13

Functional and Structural Genomics of Amino Acid Metabolism in *Streptomyces coelicolor*

by

Francisco Barona Gómez

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Biological Sciences

University of Warwick, Department of Biological Sciences

September 2003

PAGE NUMBERING AS ORIGINAL

To the memory of Salvador Ibarra, to whom the promises of scientific progress meant only a vague hope

A la memoria de Salvador Ibarra, para quien las promesas del desarrollo científico, fueron tan solo eso

Declaration

The author, Francisco Barona-Gómez, hereby states that the work described in this thesis was conducted by myself under the supervision of Dr. David A. Hodgson with the exception of those instances where the contributions of others have been specifically acknowledged. This thesis does not form any part of any other work submitted for a higher degree to another university.

Some of the work presented in this thesis has been published elsewhere in:

Barona-Gómez, F. and Hodgson, D. A. (2003). "Occurrence of a putative ancient-like isomerase involved in histidine and tryptophan biosynthesis." *EMBO Rep* **4**(3): 296-300.

Table of Contents

Introduction

Results and Discussion

Chapter II: Discovery of the enzyme phosphoribosyl isomerase (PriA) involved in both tryptophan and histidine biosynthesis

Chapter III: Structural and functional characterisation of PriA

Acknowledgements

First of all I should thank Dr David A. Hodgson whom, through a unique blend of scientific freedom, commitment and wisdom, proved to be an excellent supervisor and an endless source of knowledge and intellectual inspiration. I would also like to thank all my colleagues and friends that were always keen in helping and sharing their thoughts and ideas with me; Dr David Whitworth for introducing me into the molecular biology laboratory when I barely could distinguish between a gene and a protein; Dr Bertolt Gust for always sharing with me the latest developments of the PCR-targeting mutagenesis system: full recognition to his technical and fundamental knowledge of *Streptomyces* genetics; Dr Helena Wright for a very fruitful, and hopefully lasting, collaboration in the field of protein structure, always ready to try whatever is required to solve a structure; Drs Jonathan Blackburn and Wayne Patrick for their help in the biochemistry of tryptophan biosynthesis and interesting discussions on the evolution of the $(\beta\alpha)_8$-barrels; Dr Christine Orengo for inviting me to visit her group where I discovered the powers of bioinformatics and met Dr Stuart Rison and Dr David Lee, always ready to reply my emails with useful hints for making the computer a biological laboratory (my thanks also to them); and Dr Greg Challis whose help and support went beyond the science of natural products, particularly during the writing of this thesis. Finally, I would also like to thank all my colleagues from the Department of Biological Sciences and from the *Streptomyces* community, whom without being aware of, had a massive impact on this thesis and made this scientific experience a most enjoyable one. Citing all of them would be an endeavour as hard as writing this thesis!

List of Figures

Chapter I

Chapter II

Chapter III

Chapter IV

Chapter V

Chapter VI

List of Tables

Chapter II

Chapter IV

Chapter V

Chapter VII

Summary

An investigation of amino acid metabolism in *Streptomyces coelicolor*, including the anabolism of tryptophan, histidine, the branched-chain amino acids and proline, as well as the catabolism of the latter, is reported. The experiments reported herein were conceptually conceived within a functional genomics framework. For this purpose the complete genome sequence of *S. coelicolor* was systematically exploited. Moreover, the current knowledge on the physiology of *Streptomyces* was taken onboard, as well as the prevailing and emerging notions on the evolution of proteins and metabolic pathways. Some of the results obtained using *S. coelicolor* as a model organism were expanded to other actinomycetes, such as *Mycobacterium tuberculosis*. This was aided by a comparative genomics analysis of the actinomycetes whose genomes have been sequenced. The theoretical principles that give support to this thesis are introduced in Chapter I. This study was greatly facilitated by the development of a novel PCR-targeting mutagenesis method of which details can be found in Chapter VII.

The discovery of a common isomerase for tryptophan and histidine biosynthesis is reported in Chapter II. This discovery arose from efforts aimed at reconstructing the tryptophan biosynthetic pathway of *S. coelicolor*, since the genome sequence project of this organism failed to identifiy a *trpF* gene coding for the enzyme phosphoribosyl anthranilate isomerase. The solution of this functional genomics discrepancy led to the discovery of a putative $(\beta\alpha)_8$-barrel enzyme, termed PriA, whose preliminary functional and structural characterisation is reported in Chapter III. The evolutionary implications of the discovery of PriA are discussed within Chapters III and IV. A comparative genomics analysis of actinomycetes centred on the *priA* gene is presented in the latter Chapter, supporting the notion that this novel protein is spread across the high (G + C) content Gram-positive organisms. Indeed, it was predicted that a *priA* orthologue accounts for the lack of a *trpF* gene from the genome of *M. tuberculosis*, a hypothesis that proved to be correct. Finally, evidence to support the notion that the histidine and tryptophan biosynthetic pathways co-evolved is presented.

In contrast to the isomerisation catalysed by PriA, in which an enzyme is shared by two amino acid biosynthetic pathways, several paralogous enzymes with the potential to account for the first step of tryptophan biosynthesis from chorismate were found on the genome of *S. coelicolor*. These chorismate-utilising enzymes are investigated in Chapter V. Mutational analysis of some of this paralogues is reported and it is anticipated that the analysis and results reported therein will serve to direct future experiments aimed at identifying the *trpE* paralogue encoding the enzyme anthranilate synthase.

Chapter VI reports on the identification of the *proC* gene involved in the last step of proline biosynthesis in *S. coelicolor*. The pyrroline-5-carboxylate reductase activity of the enzyme encoded by the putative *proC* gene was extensively characterised, with particular emphasis on the interaction between primary and secondary metabolism. Furthermore, mutational analysis of *proC* suggested that paralogues of this gene are present on the genome of this organism, since its deletion did not lead to an auxotrophic phenotype. Investigation of this observation showed that two paralogous enzymes encoded by *ilvC*-like genes, involved in biosynthesis of the branched-chain amino acids, are capable of compensating for the lack of *proC*. This is the first example of a physiological link between the biosynthesis of proline and the branched-chain amino acids.

To sum up, the results reported in this thesis represent an advancement towards understanding the physiology of *S. coelicolor* as a model actinomycete, within a functional and structural genomics framework. They also offer evidence on the evolutionary principles that lead to the appearance of novel proteins and metabolic pathways in bacteria.

Abbreviations

AA	Anthranilic acid
ACT	actinorhodin
ADC	4-amino-4-deoxychorismate
ADIC	2-amino-2-deoxyisochorismate
AHB	2-aceto-2-hydroxybutyrate
AHIR	acetohydroxy acid isomeroreductase
AICAR	5′-(5-aminoimidazole-4-carboxamide) ribonucleotide
AL	2-acetolactate
AS	anthranilate synthase
ATP	Adenosine triphosphate
CDA	calcium-dependent antibiotic
CDART	conserved domain architecture retrieval tool
CDRP	1-[(2-carboxyphenyl)amino]-1-deoxyribulose 5-phosphate
CM	chorismate mutase
COG	cluster of orthologous genes
CPL	chorismate pyruvate-lyase
DHAP	2-keto-3-deoxy-D-arabinoheptusonate-7-phosphate
FMOP	flavin mononucleotide-dependent oxidoreductase and phosphate-binding
GK	glutamyl kinase
GS-DH	glutamate-γ-semialdehyde dehydrogenase
HMM	hidden Markov models
HT	his and trp (genes)
ICS	isochorismate synthase
IGP	indole-3-glycerol phosphate
ImGP	imidazole-3-glycerol phosphate
MA	minimal agar
MM	minimal medium
OSG	orthologous strings of genes
P5C	Δ1-pyrroline-5-carboxylate
PABA	*p*-aminobenzoic acid
PDB	Protein Data Base
P-DH	proline dehydrogenase
PHB	4-hydroxy benzoic acid
PRA	*N′*-(5′-phosphoribosyl)anthranilate
PRFAR	*N′*-[(5′-phosphoribulosyl)formimino]-5-aminoimidazole-4-carboxamide
ProFAR	*N′*-[(5′-phosphoribosyl)formimino]-5-aminoimidazole-4-carboxamide ribonucleotide
PRPP	phosphoribosyl pyrophosphate
P5C-R	P5C reductase
RED	tripyrrole prodiginines
SFM	soya-flour-mannitol
SPB	standard phosphate-binding
TIM	triose phosphate isomerase
VB	Vogel-Bonner

Chapter I

Chapter I

Introduction

1.1 Evolving Genes and Proteins

1.1.1 Historical background

The title *Evolving Genes and Proteins* of this Section was burrowed from a symposium held at the Institute of Microbiology of the University of Rutgers (Bryson and Vogel, 1965). Nowadays it is accepted that genes and proteins are intimately related in evolutionary terms. However, although considered to be secondary, at the time of this symposium, the first evolutionary implications of the study of molecular biology had started to be appreciated (Zuckerkandl and Pauling, 1965; cited also in Woese, 1987). Although some hypotheses for the evolution of genes, proteins and biochemical pathways were put forward before this period, they were more often intuitive (e.g. Horowitz, 1945) or derived from circumstantial evidence (e.g. Beadle and Tatum, 1941). It was not until the 1960s that new theories based on nucleotide and amino acid sequence data started to emerge within the framework of the *one gene – one enzyme* paradigm (Beadle and Tatum, 1941). Indeed, the feeling that one has after a quick glance of the program of this symposium is that the motivation behind this scientific gathering was to integrate the separate notions about the evolution of genes, evolution of proteins and evolution of metabolic pathways, into a broader theory of molecular evolution.

In the words of Tatum (1965) during the opening address of this symposium, "the evolution of metabolic pathways, the evolution of proteins, and the evolution of genes are of course intimately and causally interrelated [....] questions (that) we are conceptually ready to ask". In fact, the groundbreaking lecture by Zuckerkandl and Pauling (1965) about the evolutionary history hidden in the sequence of haemoglobins addressed this very point. The notion of a molecular monogenetic classification of living beings rather than an organismal taxonomy was implicit in the works of Zuckerkandl and Pauling (1965), preparing the field for the birth of molecular evolution as a discipline. This work settled the basis for modern phylogenetics, i.e. the classification of living beings accordingly to the similarities of their rRNA genes into the three domains of life, namely, Archeabacteria, Eubacteria and Eukaryote (Woese, 1987).

1.1.2 The role of duplication in gene evolution

The general idea of gene evolution by tandem duplication dates back to the 1950s (Lewis, 1951). Even before then, Haldane and Muller appreciated in the early 1930s the evolutionary significance of gene duplication (cited in Grauer and Li, 2000, pp. 249). Obviously these hypotheses could not have offered a molecular mechanistic model for the process of gene duplication since the structure of DNA was not discovered until the mid 1950s (Watson and Crick, 1953). It was not until the 1970s, when the evidence for polyploidy became available, that Ohno (1970) hypothesised that gene duplication provides a temporary escape from the relentless pressure of natural selection on a duplicate gene allowing the new gene copy to follow one of two alternative evolutionary fates. The classical model predicts that the duplicated gene will accumulate

deleterious mutations through selection and therefore be lost. This process is known as *non-functionalisation* and often gives rise to pseudogenes (Grauer and Li, 2000). Alternatively, although considerably less likely to occur, random mutations could accumulate under neutral selection until a new beneficial function would arise by chance that could then be selected for by natural selection (Ohno, 1970). This process is referred to as *neo-functionalisation* and it has been traditionally considered to be the main origin of protein novelties (Grauer and Li, 2000).

Despite the fact that the view of gene duplication as originally proposed by Ohno (1970) has prevailed as the main conceptual framework by which the evolution of genes is explained, Ohno's hypothesis has been challenged from different scientific frontlines since its postulation (Hughes, 1994; Graur and Li, 2000, pp. 281-283; Lynch and Conery, 2000; Kondrashov *et al.*, 2002; Gu *et al.*, 2003). For instance, the lack of function of the duplicate gene before duplication has been argued against (Hughes, 1994). The eye lens crystallins in which acquisition of new functions preceded gene duplication exemplifies the argument (Piatigorsky and Wistow, 1991; Piatigorsky, 2003). In the case of crystallins two unrelated protein functions, such as enzymatic activities and a structural role responsible for the optical properties of the lens, co-exist within the same polypeptide chain. This phenomenon is known as gene sharing (Graur and Li, 2000, pp. 302) and it represents the underlining basis of the so-called *moonlighting proteins*. These polypeptides adopt different functions depending upon cellular factors such as expression pattern or concentration of key metabolites (Jeffery, 1999; Copley, 2003). In principle, these proteins differ from multidomain mosaic proteins in that their functions are attributable to a single domain. An example of a

moonlighting protein whose function depends upon cellular localisation and concentration of its substrate is PutA of *Escherichia coli* involved in proline catabolism (Jeffery, 1999). The function of PutA switches between its pyrroline-5-carboxylate (P5C) dehydrogenase activity, which is associated with the plasma membrane, and its transcriptional repressor function by which it binds DNA in the cytoplasm, as a response to the presence of proline (Vinod *et al.*, 2002 and references therein; see Chapter VI).

Recent results obtained from genome-wide analyses have also highlighted the limitations of Ohno's hypothesis in understanding the high frequency of occurrence and organisation of gene duplication (Lynch and Conery, 2000; Kondrashov *et al.*, 2002; Gu *et al.*, 2003; for a comment see Meyer, 2003). It has been shown that gene duplications occur more often than predicted by the theory of population genetics and might be important in speciation in addition to alteration of protein function (Lynch and Conery, 2000). Furthermore, it has been suggested that in yeast, duplicate genes confer significant genetic robustness to the cell under both fermentation and respiration conditions (Gu *et al.*, 2003). Gu and co-workers observed that duplicate genes accounted for at least one quarter of the genome of *Saccharomyces cerevisiae* with one or more paralogous partner(s) somewhere else in the chromosome. This observation motivated the comparison of fitness values of single-gene-deletion mutants with and without paralogous partners, i.e. mutants of duplicate genes *versus* singleton gene mutants. The results showed that null mutants of duplicate genes have consistently a significantly higher fitness value than singleton gene mutants. Furthermore, it was discovered that less diverged homologues compensate better for the function of the

corresponding paralogue and, even when the paralogues are highly diverged, there is some degree of functional compensation (Gu *et al.*, 2003; see also Chapters V and VI).

The results discussed above, together with the observation that many genes and gene networks are similar in evolutionary diverse species, suggests that gene duplication more commonly provides a means of preserving function than being a source of gene innovation as suggested by Ohno (Meyer, 2003). A third evolutionary fate for a gene duplication event has been recently described. Although this model is based on the same assumptions as Ohno's model, i.e. that duplicated genes are redundant and diverged under neutral selection, mechanistically it is somewhat different. This process relies on complementary degenerative changes in a pair of duplicate genes, such that the duplicates together retain the original functions of their single ancestor (Force *et al.*, 1999). The nature of this process implies that the ancestral gene used to have multiple functions and hence the *sub-functionalisation* term has been coined (reviewed be Prince and Pickett, 2002; see also Chapter VI).

It appears therefore that at least in the short-term duplicated genes evolved under selection, as opposed to neutral evolution, both in eukaryotes and prokaryotes (Lynch and Conery, 2000; Kondrashov *et al.*, 2002; Gu *et al.*, 2003). One implication of this scenario is that gene duplication may confer an advantage to the cell immediately after it occurs, probably as an effect of protein dosage (Kondrashov *et al.*, 2002) or due to the presence of a function predating the duplication event (Hughes, 1994). Interestingly, the recently fully sequenced genome of the Gram-positive high (G + C) content organism *Streptomyces avermitilis* ATCC 31267 was shown it to be composed of approximately one third of duplicate genes (Ikeda *et al.*, 2003). This situation is likely to be shared by

its close relative the model actinomycete *Streptomyces coelicolor* A3(2) (Bentley *et al.*, 2002), which is used as the model organism for experimentation within this thesis (see Section 1.3 for an introduction to *S. coelicolor*). The level of presumed gene duplication detected in *Streptomyces* raises the possibility that similar evolutionary laws to those discussed above may be in place in these bacteria. In fact, it was acknowledged since very early that genetic duplication extensively occurs within the genomes of *Streptomyces* species, e.g. during the isolation of heteroclones. For an explanation of the genetic nature of *Streptomyces* heteroclones see Anderson and Roth (1977).

1.1.3 Bacterial genome evolution

Evolutionary events leading to gene duplication in bacteria have been inferred from experimental observations in the laboratory and several mechanisms had been envisaged. For instance, gene amplification in bacteria as a response to environmental factors is well acknowledged, such as the duplication or multiplication of genes conferring low-level of resistance to antibiotics (reviewed by Kondrashov *et al.*, 2002; see also Kessler *et al.*, 1989; Birch *et al.*, 1990; Dittrich *et al.*, 1991 for specific examples in *Streptomyces* species). Moreover, recent analyses of complete bacterial genome-sequences have been invaluable in understanding genome evolution on a more realistic evolutionary time-scale (Mira *et al.*, 2002). It was often believed that the main mechanism by which an organism gained duplicated DNA and therefore expanded its functional diversity was through recombination. The size of the genetic material that can be duplicated by recombination can range from sequence motifs, genes, chromosomal segments and even entire genomes (Hopwood, 1999; Betran and Long, 2002; Mira *et*

al., 2002). These genome rearrangements usually take place via homologous recombination, although illegitimate recombination has also been reported (Doolittle, 1995 and references therein). In such a context, as discussed below, mechanisms involving mobile genetic elements are crucial for the evolution of genomes.

When genetic recombination occurs during cell division the transfer of the outcome is limited to vertically clonally inheritance. However, a contrasting alternative to this process exists in the phenomenon of lateral gene transfer (LGT) by which an organism enriches its genome horizontally with DNA from either a close relative (e.g. conjugation, transduction or transformation in prokaryotes) or distantly related species. Although the role of LGT in the evolution of species and their genomes is an ongoing debate it appears that it had a larger impact than originally appreciated (for arguments in the debate of the role of LGT in genome evolution see Lawrence and Ochman, 1998; Doolittle, 1999; Martin, 1999; Glansdorff *et al.*, 2000; Ochman *et al.*, 2000; Woese, 2000; Jain *et al.*, 2002; Gogarten *et al.*, 2002; Lawrence and Ochman, 2002). This debate started when the first comparisons of complete bacterial genomes reported the unexpected observation that genomes of even closely related organisms were not conserved (Mushegian and Koonin, 1996; Tamames *et al.*, 1997; Koonin and Galperin, 1997; Lawrence and Ochman, 1998). This lack of bacterial genome conservation was at two levels: gene content and gene order (see also Section 1.1.3.1). The differences in gene content were often explained by means of LTG, although not always compelling evidence was offered (for a critical review see Glansdorff *et al.*, 2000).

The occurrence of LGT (Ochman *et al.*, 2000) and gene duplication (Mira *et al.*, 2002) has been previously positively correlated with the abundance of mobile sequences

within the recipient genome. The role of plasmids in the evolution of pathogenic bacterial genomes (Ochman *et al.*, 2000) and phages in the evolution of the DNA replication machinery (Forterre, 2002) are only some examples of how mobile genetic elements shape genome evolution, in some cases involving LGT. Thus, as judged from the lack of conservation, it may appear that bacterial genomes are better understood as highly dynamic genetic entities showing a great degree of fluidity and, to a certain degree, a lack of evolutionary memory, at least in terms of gene organisation (see below next Section). Interestingly, evidence suggesting that LGT had occurred in members of the genus *Streptomyces* has been published (Huddleston *et al.*, 1997; Egan *et al.*, 1998; Ueda *et al.*, 1999) strongly suggesting that the genome of *S. coelicolor*, the subject of study of this thesis, may have been shaped by LGT during the course of evolution. Furthermore, the phenomenon of genetic instability and genome rearrangement in members of the *Streptomyces* genus has been extensively characterised (see e.g. Leblond *et al.*, 1990; Birch *et al.*, 1990 and 1991). For instance, it has been shown that more than a million base pairs of DNA at either end of the linear chromosome of *Streptomyces* can undergo extensive deletions and amplifications without compromising viability, at least under laboratory conditions (Volff and Altenbuchner, 1998; see also Section 1.3.2).

1.1.3.1 Lack of conserved gene organisation in bacterial genomes

With the advent of the availability of a large collection of bacterial genomes it has been possible to compare bacteria at a whole-genome level. These efforts stand within the field of *comparative genomics*, which is defined as the study of the

differences and similarities in genome structure and organisation in different organisms (Primrose and Twyman, 2003). One of the first fruits of this emerging discipline is the demonstration that bacterial genome organisation is poorly conserved, as originally recognised by Koonin and Galperin (1997). Within this context, the highest level of organisation is represented by operons, which are groups of genes transcribed in a single polycistronic mRNA co-ordinately expressed from a single promoter situated upstream (Jacob *et al.*, 1960; see also de Daruvar *et al.*, 2002 for the modern view on operons). Despite operons being frequently found in bacterial genomes, it has been previously shown that operons tend not to be conserved (Mushegian and Koonin, 1996; Tamames *et al.*, 1997; Itoh *et al.*, 1999). Only in the cases where operons encode for proteins that physically interact has some level of conservation been detected (Dandekar *et al.*, 1998). The lack of conservation in gene organisation observed in bacterial genomes has been exploited to develop methods for the prediction of gene function through genomic context analysis. These methods are used in Chapter IV of this thesis.

1.1.4 Mechanisms for the evolution of protein function

Irrespective of the evolutionary trajectory leading to the appearance of a coding region, once a new gene has been fixed in a recipient genome, the corresponding protein function would evolve through the process of random mutation and natural selection. With the advent of structural genomics studies aimed at understanding the biology of an organism from the 3D representation of its proteome, "the prevailing trends in emergence and diversification of protein domains" have just started to be understood (Aravind *et al.*, 2002; see also Section 1.1.5.2 and references therein). These studies

have confirmed the fundamental prediction that only a few thousands different structures encompass the immensity of the protein universe (Murzin *et al.*, 1995; Murzin, 1998). This implies extensive reuse of the same topological solutions during the course of protein evolution. Thus natural selection has exerted its effect upon existing protein folds leading to generation of new functions, and not the invention of new protein folds to carryout new functions. It appears therefore that the extant proteins diverged from a relatively small number of ancestral polypeptides implying a common origin of each fold (Copley and Bork, 2000; Todd *et al.*, 2001). However, a common evolutionary trajectory was not necessarily followed in order to reach each function (Doolittle, 1994; Galperin *et al.*, 1998; Aravind *et al.*, 2002).

The evolution of proteins is often accomplished through different mechanisms independently or in combination. The genetic mechanisms include point mutations, insertions, deletions, domains shuffling and gene fusions (Doolittle, 1995; Gupta, 2000; Todd *et al.*, 2001). Proteins consequently can evolve by forming different oligomerisation associations and post-translational modifications (Todd *et al.*, 2001). A schematic diagram that illustrates the mechanisms by which new protein functions could arise is shown in Figure 1.1. Furthermore, it has become clear that protein evolution has been centred in domain architecture rather than in domain innovation, i.e. the appearance of new protein functions rely more often in linear rearrangements of domains in a protein than in the appearance of new protein domains (Todd *et al.*, 2001; Betran and Long, 2002).

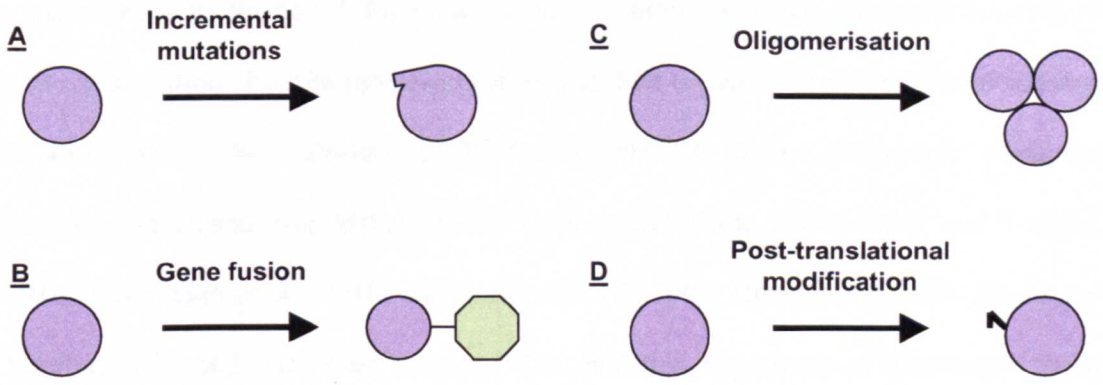


Figure 1.1 Possible mechanisms for protein evolution. Once a gene has been duplicated, it undergoes modifications leading to new protein functions. The modifications include **A** incremental mutations (accounting for point mutations, insertion/deletions and domain shuffling); **B** gene fusion; **C** oligomerisation and **D** post-translational modification. New functions often evolve through a combination of these mechanisms. Adapted from Todd *et al.* (2001).

1.1.5 Mechanisms for the evolution of enzyme function

In general terms, the same principles of protein evolution apply to the evolution of enzyme function. However, detailed structural mechanistic studies on the catalysis of enzymes have provided very sophisticated insights into how a new enzymatic activity might appear (see Gerlt and Babbitt (2001b) for a very comprehensive review on enzyme evolution). The current knowledge on enzyme evolution from a structural perspective is discussed in Section 1.1.5.2, after introducing the early notions that prevailed before the structural genomics era in Section 1.1.5.1.

1.1.5.1 Early notions on enzyme evolution

From its infancy the field of enzyme evolution was conceptually shaped by studies on the origin of life. The first hypothesis with a bearing on enzyme evolution took into account the environmental constrains imposed by the prebiotic soup in decay (Horowitz, 1945). This is the retrograde evolution hypothesis, which will be discussed

in more detail in Section 1.1.6 in terms of its contribution for the understanding of pathway evolution. For the time being it will suffice to realize that the idea of enzyme evolution driven by substrate-specificity is implicit in the retrograde evolution hypothesis (Lazcano and Miller, 1999; Copley and Bork, 2000; Gerlt and Babbitt, 2001b; Teichmann *et al.*, 2001a and b; Alves *et al.*, 2002; Rison *et al.*, 2002; Rison and Thornton 2002 and references therein). The modified hypothesis of Horowitz (1965) implies that enzymes can evolve after gene duplication by *retention of their substrate-binding pocket and modification of their catalytic capabilities.* This possibility seems plausible particularly within the context of the evolution of enzymes involved in the same pathway since such a collection of enzymes would share substrates and products. It is assumed therefore that a certain degree of overlapping specificities should exist among these proteins (Horowitz, 1965). Furthermore, this type of enzyme evolution is not limited to enzymes from the same pathway since similar and even identical substrates and products can be common to different metabolic pathways.

Since the evolution of enzymes driven by substrate-specificity ultimately modifies the chemistry, the mechanism of reaction catalysed by the ancestral and new enzymes is not related. This, as a consequence, gives birth to *functionally distinct suprafamilies*, i.e. groups of homologous enzymes that catalyse different overall reactions without sharing any common mechanistic attribute (Gerlt and Babbitt, 2001b; see also Section 1.1.5.2). Examples of enzymes that have been historically considered to evolve in this fashion are enzymes involved in tryptophan biosynthesis as originally proposed by Horowitz (1965) (see also Figure 1.2). This proposal has been recently supported by the observation that these tryptophan biosynthetic enzymes bind similar

phosphate-substrates through a common motif present in their $(\beta\alpha)_8$-barrel structure suggesting a common origin (Wilmanns *et al.*, 1991 and 1992). For an examination of this hypothesis in the evolution of the tryptophan biosynthetic pathway in *S. coelicolor* see Chapter II.

With the advent of more functional (biochemical) studies dealing with the activity and specificity of enzymes a different view on enzyme evolution started to emerge. It became a common experience to discover enzymes with broad-substrate specificities and overlapping activities as recognised by Jensen (1976), and recently reviewed by O'Brien and Herschlag (1999) and Copley (2003). This extensive body of evidence, together with results obtained from global analyses of the protein universe from a structural perspective (see below Section 1.1.5.2), suggested that enzymes were mostly recruited on the basis of similar catalytic capabilities in a process by which the *substrate-binding domain evolved to fit new substrates whereas the chemistry remained unchanged* (Jensen, 1976; Gerlt and Babbit, 1998; O'Brien and Herschlag, 1999; Gerlt and Babbitt, 2001b). One of the consequences of such large-scale evolutionary recruitment is the large number of protein families composed of homologous enzymes leading to superfamilies which can be of two types: (a) superfamilies whose members catalyse the same reaction with differing substrate specificities, termed *specificity diverse superfamilies*; and (b) superfamilies whose members share a common mechanistic attribute but perform different overall reactions, termed *mechanistically diverse superfamilies* (Gerlt and Babbitt, 2001b). Indeed, the existence of these superfamilies is evidence of the catalytic promiscuity of enzymes, a feature that appears to be intrinsic to the nature of enzymes.

1.1.5.2 Enzyme evolution from a structural perspective

Recent detailed structural analyses of the catalytic mechanisms of enzymes have added another level of sophistication to the picture on enzyme evolution. Thanks to the increasing number of enzyme's structures that have been elucidated it has been possible to discover evolutionary relationships (i.e. the basis of protein classification into families) that remained uncovered until before the structural genomics era. On the one hand, 3D structure elucidation has been applied to the study of enzymes whose substrates can range from metabolites containing a few atoms to macromolecules. These enzymes are typical examples of specificity diverse superfamilies. One of the most drastic examples where the substrate varies the most is that of the glycosyl hydrolase superfamily, which has the $(\beta\alpha)_8$-barrel domain. Members of this superfamily act on monosaccharides as well as polysaccharides, including cases in which proteins and organic groups are linked to the polymeric sugar (cited in Todd *et al.*, 2001). On the other hand, the paradigm of the mechanistically diverse superfamilies is represented by the enolase superfamily, which also has the $(\beta\alpha)_8$-barrel domain (Neidhart *et al.*, 1991; Hasson *et al.*, 1998; Gerlt and Babbitt, 1998 and 2001b). Each of the reactions accomplished by members of this superfamily involves formation of an enolic intermediate by abstraction of the α-proton in the carboxylate of the substrate, but depending upon the exact reaction the intermediate is diverted to the product by different active site functional groups. Furthermore, the position of the catalytic atoms within the scaffold of members of the enolase superfamily remains constant despite the fact that the atoms are not always in the same residues (Hasson *et al.*, 1998; Bartlett *et*

al., 2002; Todd *et al.*, 2002; Gerlt and Raushel, 2003), a remarkable conservation pattern only detectable at the structural level.

Recently a third mechanism by which an enzyme function can evolve has been described with the aid of structural studies, showing the opportunistic nature of evolution (Gerlt and Babbitt, 2001b; Wiles *et al.*, 2002). This involves the use of a structural scaffold and functional residues from a progenitor enzyme in completely different mechanistic and metabolic contexts. Wise *et al.* (2002) showed that the enzymes orotidine 5'-monophosphate decarboxylase from *Bacillus subtilis* and 3-keto-L-gulonate 6-phosphate decarboxylase, encoded by the *ulaD* gene of *E. coli*, are homologues despite the fact that they perform completely mechanistically different reactions making use of a common $(\beta\alpha)_8$-barrel domain and structurally equivalent residues. Thus, these enzymes are members of the same *functionally distinct suprafamilies* although they are involved in different metabolic pathways (see also Section 1.1.5.1). This mechanism of enzyme evolution has been recently expanded to enzymes with different folds (Bartlett *et al.*, 2002; Todd *et al.*, 2002; Anantharaman *et al.*, 2003). It is interesting to note that the cases presented above involve protein families that adopt the $(\beta\alpha)_8$-barrel structure, a reflection of the structural plasticity of this fold (Todd *et al.*, 2001; Gerlt and Babbitt, 2001a and b; Nagano *et al.*, 2002; Gerlt and Raushel, 2003).

1.1.5.3 Divergent evolution of enzyme function

The existence of superfamilies and suprafamilies as inferred from the structural evidence discussed above is strong evidence that enzyme recruitment is a major force in

the *divergent evolution* of enzyme function. However, independent inventions during the course of enzyme evolution have been reported (Doolittle, 1994; Perona and Craik, 1997; Galperin *et al.*, 1998; Aravind *et al.*, 2002), i.e. *convergent evolution* of analogous enzymes with the same activity but different origin (as judged from sequence and structure comparisons). Nevertheless, the same principles of enzyme recruitment should apply in both evolutionary processes. Moreover, the existence of protein domains with the same enzyme activity but different fold is a strong argument for divergent, as opposed to convergent evolution (Galperin *et al.*, 1998; Aravind *et al.*, 2002). This observation supports the notion discussed in Section 1.1.4 that the extant proteins diverged from a relatively small number of ancestral polypeptides.

As discussed in Section 1.1.2 gene duplication is not a *per se* condition for the evolution of protein function (Hughes, 1994). In other words, and within the context of the evolution of enzyme function, two different enzyme activities could evolve independently and simultaneously within the same polypeptide. This view is strongly supported by the catalytic promiscuity observed in a large number of protein families with enzymatic activity (Jensen, 1976; O'Brien and Herschlag, 1999; Copley, 2003). It appears therefore that enzymes with overlapping specificities may have been selected for during the course of evolution. These enzymes would be encoded by the same gene until environmental conditions changed and new pressures for independent and specific enzymes arose. Then a new enzyme would appear after gene duplication and specialisation. In fact, it has been previously proposed that the catalytic promiscuity observed in enzymes is an ancestral feature that has been retained during the course of evolution (Jensen, 1976; Lazcano and Miller, 1999; O'Brien and Herschlag, 1999;

Anantharaman *et al.*, 2003). As discussed in the following Section this feature of enzyme evolution has a bearing on the evolution of metabolic pathways.

1.1.6 Evolution of metabolic pathways

As discussed in Section 1.1.5.1 the retrograde evolution hypothesis of Horowitz (1945) was originally proposed as a model for the evolution of metabolic pathways. This hypothesis states that pathways evolved in a *backward* stepwise fashion in heterotrophic life forms driven by the scarceness of organic compounds in the prebiotic soup (for reviews in these concepts see Lazcano and Miller, 1999; Rison and Thornton, 2002). In this context the first organisms would thrive heterotrophically from organic compounds such as amino acids and nucleotides (represented by compound [A] in Figure 1.2) produced abiotically. The Horowitz hypothesis assumes that as life appeared compound A started to run out and its availability declined drastically imposing a pressure for the appearance of another means of synthesising compound A, which presumably by then was already essential for sustaining life. The retrograde hypothesis proposes that this pressure would drive the appearance of an enzyme (i.e. Enz 1 in Figure 1.2) that could make use of the environmentally available compounds B and C as substrates for the synthesis of compound A. On an iterative process the new enzyme Enz 1 would make use of compound B (and C) to the point in which its presence would inevitably drop leading to a new pressure for the appearance of a new enzyme function (i.e. Enz 2 in Figure 1.2) that could use compounds D and E for the synthesis of compound B. This process would account for the evolution of metabolic pathways (Horowitz, 1945; Lazcano and Miller, 1999).

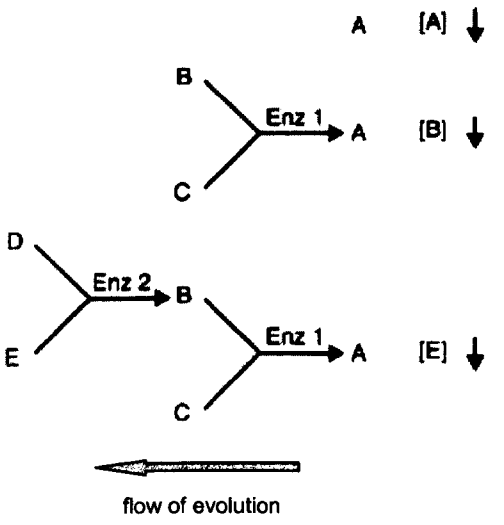


Figure 1.2 The retrograde (Horowitz) model of pathway evolution. An organism using key compound A uses up all of the environmental supply of the compound. The fortuitous recruitment of an enzyme, Enz 1, capable of synthesising A from B and C confers a survival advantage to the organism. In turn, environmental concentrations of B (and C) drop, compensated by the recruitment of a new enzyme, Enz 2, capable of producing compound B from D and E. By this process the pathway evolves backwards starting from the final-product of the pathway and ending at the initial pathway metabolite(s). Adapted from Rison and Thornton (2002) with permission from the first author.

Although the retrograde hypothesis as originally proposed (Horowitz, 1945) remains the most plausible explanation of what may had been the driving force in the evolution of metabolic pathways there is one obvious problem. This problem is that this model fails to explain the evolution of metabolic pathways that use as intermediaries labile compounds that could not have accumulated in the environment (for alternative proposals to overcome this problem see Lazcano and Miller, 1999; Roy, 1999; Anantharaman *et al.*, 2003). Moreover, in a paper presented at the symposium *Evolving Genes and Proteins* held at the University of Rutgers (see Section 1.1.1) Horowitz (1965) complicated his retrograde hypothesis when attempting to accommodate into his theory the recently developed operon model of Jacob, Monod and co-workers (Jacob *et al.*, 1960). In this paper, inspired by the observation that in *E. coli* and relatives certain amino acid biosynthetic genes (e.g. histidine, tryptophan and leucine) were shown to cluster together and form an operon, Horowitz proposed that metabolic pathways evolved by duplication of genes encoding enzymes within the same metabolic pathway (Horowitz, 1965). One implication of this modified retrograde model is that pathways

and therefore enzymes retained substrate-specificity during the course of evolution. As extensively discussed in Section 1.1.5 it is now clear that enzymes far more often evolved by a process in which the catalytic capabilities remained unchanged whereas the substrate specificity was modified. A second implication of the Horowitz proposal of 1965 is that enzymes from within a single pathway are homologues of each other, i.e. they share a common origin. Global metabolic analyses have shown that homologous enzymes tend not to be conserved within a given metabolic pathway but instead are scattered randomly across the network of the small-molecule metabolic (SMM) pathways (Teichmann *et al.*, 2001a and b; Rison *et al.*, 2002; Rison and Thornton, 2002; Alves *et al.*, 2002). From this observation that the architecture of the SMM pathways has been compared to a 'mosaic' in which enzymes, represented by pebbles in the comparison, have been recruited in the basis of catalytic capabilities (Teichmann *et al.*, 2001a and b).

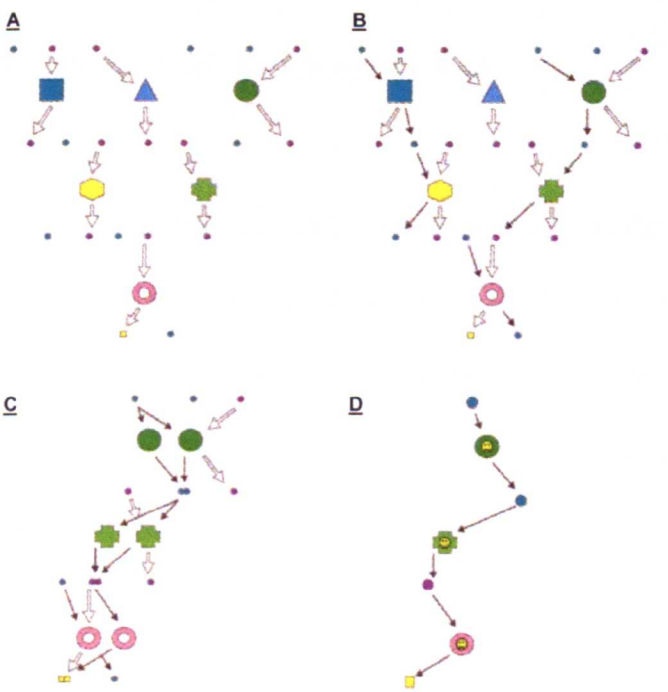


Figure 1.3 The patchwork model of pathway evolution. In the patchwork model, enzymes may have a favoured substrate and catalytic mechanism **A**, but exhibit broad substrate specificities and are capable of catalysing other reactions **B**. Therefore, many metabolic chains synthesising key metabolites (e.g. yellow square) may have existed, such as the one catalysed by the olive circle, the green cross and the pink doughnut. Duplication of any gene in such a pathway **C** would be advantageous, as more of the key metabolite would be synthesised. This duplication, followed by enzyme specialisation **D**, would account for extant pathways. The Figure was taken from Rison and Thornton (2002) with permission from the first author.

The data discussed above strongly supports Jensen's patchwork hypothesis (Jensen, 1976), which states that pathways evolved through a general process of enzyme recruitment driven by the catalytic capabilities of the enzymes. A corollary of this hypothesis is that ancestral enzymes used to have broad-substrate specificity and therefore were able to participate in more than one enzymatic reaction and metabolic pathway (Figure 1.3; see also Figure 3.2). During the course of evolution some of these ancestral enzymes may have been duplicated and specialised in order to perform one of the ancestral functions more efficiently. However, as suggested by the catalytic promiscuity of enzymes discussed in Section 1.1.5.1 this feature may have been retained in some cases during the course of evolution. These evolutionary possibilities are shown in Figure 1.3. For an examination of this hypothesis in the evolution of the tryptophan biosynthetic pathway in *S. coelicolor* see Chapter II.

1.2 History of the $(\beta\alpha)_8$-barrels

The first protein that was shown to adopt the $(\beta\alpha)_8$-barrel structural fold was the enzyme triose phosphate isomerase (TIM; Banner *et al.*, 1975). From this circumstance the $(\beta\alpha)_8$-barrels are also known as TIM-barrels. Since the report by Banner *et al.* (1975) the number of proteins discovered showing the $(\beta\alpha)_8$-fold has increased at a staggering rate. By the late 1980s a total of seventeen proteins shared the $(\beta\alpha)_8$-fold and were subject of a review as a structural protein family for the first time (Farber and Petsko, 1990). By the mid 1990s, there were thirty known proteins that adopt the $(\beta\alpha)_8$-

fold (Reardon and Farber, 1995) reaching 147 non-redundant $(\beta\alpha)_8$-barrel structures elucidated by 2001 (Nagano *et al.*, 2002). This situation is evident after inspection of the classification database CATH (version 1.7) which includes 889 different $(\beta\alpha)_8$-barrels comprising 503 Protein Data Base (PDB) entries, 76 sequence families and 18 different superfamilies (Nagano *et al.*, 2002). For this thesis the CATH database is preferred over other protein databases (e.g. SCOP) since the work by Nagano *et al.* (2002), which is based on data derived from CATH, represents the latest comprehensive review of the $(\beta\alpha)_8$-barrel structural fold.

The increasing degree of attention that the $(\beta\alpha)_8$-barrels have gained is shown by the number of reviews published to date (Farber and Petsko, 1990; Farber, 1993; Reardon and Farber, 1995; Nagano *et al.*, 1999; Pujadas and Palau, 1999; Copley and Bork, 2000; Henn-Sax *et al.*, 2001; Höcker *et al.*, 2001a; Wierenga, 2001; Nagano *et al.*, 2002; Gerlt and Raushel, 2003). Although all these works have a bearing in the evolution of the $(\beta\alpha)_8$-barrels, and in some aspects they overlap, different emphases can be discerned. These include structure classification (Farber and Petsko, 1990; Farber, 1993; Reardon and Farber, 1995; Nagano *et al.*, 1999; Nagano *et al.*, 2002) structure organisation and stability (Pujadas and Palau, 1999; Höcker *et al.*, 2001a; Wierenga, 2001) divergent evolution (Henn-Sax *et al.*, 2001; Gerlt and Raushel, 2003) and implications for the evolution of metabolic pathways (Copley and Bork, 2000). In addition, the $(\beta\alpha)_8$-barrels are very often found as central in more general reviews (e.g., O'Brien and Herschlag, 1999; Gerlt and Babbitt, 2001b; Todd *et al.*, 2001). Therefore, it can be conclude from these works that the evolutionary history of the $(\beta\alpha)_8$-barrels is of particular interest: it remains to be clarified whether the extant $(\beta\alpha)_8$-barrels arose by

convergent evolution following different evolutionary trajectories to a stable fold or by divergent evolution from a common ancestor (see also Section 3.1.2).

1.2.1 Structural features of the $(\beta\alpha)_8$-fold

In general terms, the $(\beta\alpha)_8$-fold typically consists of eight twisted parallel β strands positioned in the centre of the domain, forming a cylinder by hydrogen bonding between each adjacent β strand (i.e. β1 strand bonds β8 and β2 strands, etc). The β strands are connected by eight intercalated α helices, which are situated externally, wrapping the β-barrel in the centre. The eight (βα) modules, as well as the β strands and α helices forming the modules, are linked through loops of different lengths. Interestingly, all known $(\beta\alpha)_8$-barrel enzymes conserve the location of their active site residues at the C-terminal end of the β strands and the loops that connect the β strands with the subsequent α helices (Farber and Petsko, 1990; Henn-Sax *et al.*, 2001; Nagano *et al.*, 2002; see Figure 1.4). These residues form the catalytic face of the barrel whereas the residues of the loops connecting the α helices with the β strands have been proposed to be important in the overall stability of the structure (Höcker *et al.*, 2001a; Wierenga, 2001). An schematic representation of the topology of a canonical $(\beta\alpha)_8$-barrel can be seen in Figure 1.4. However, it should be noted that deviations from the canonical $(\beta\alpha)_8$ topology are not rare. These variations include antiparallel β-strands or lack of β-strands, irregular number of α-helices or lack of certain α-helices, presences of other domains and natural circular permutations (Pujadas and Palau, 1999).

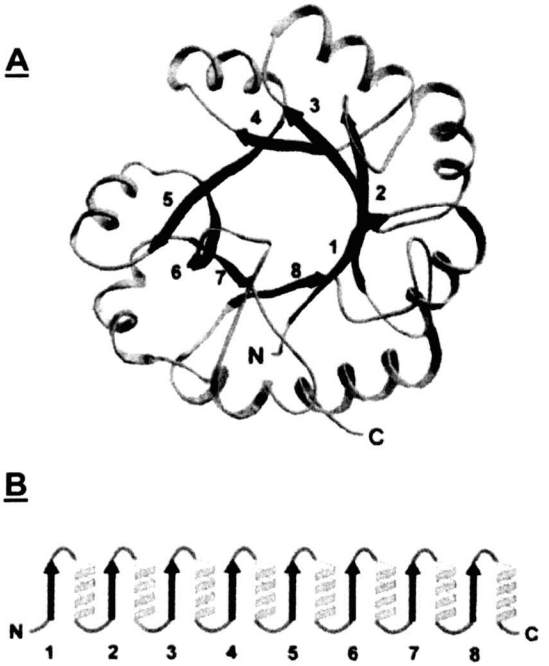


Figure 1.4. Schematic representation of a canonical $(\beta\alpha)_8$-barrel. A View onto the C-termini of the β-strands (black) that form the cylindrical parallel β-sheet. The β-barrel is surrounded by eight α-helices (grey). B Topological diagram of the eight (βα) modules. The active site residues of all known $(\beta\alpha)_8$-barrel enzymes are located at the C-terminal ends of the β-strands and in the loops that connect the β-strands with the subsequent α-helices. Taken from Henn-Sax *et al.* (2001).

1.2.2 The centrality of the $(\beta\alpha)_8$-barrels to life

The degree of interest that the $(\beta\alpha)_8$-barrels arouse in different scientific communities comes as no surprise. The vast majority of the proteins adopting the $(\beta\alpha)_8$-fold have some sort of enzymatic activity (Nagano *et al.*, 2002). Moreover, 85% of the $(\beta\alpha)_8$-barrels are directly involved in fundamental processes that sustain life (Pujadas and Palau, 1999; Nagano *et al.*, 2002) i.e. energy metabolism, macromolecule metabolism and small molecule metabolism, as defined by Rison *et al.* (2002). The ubiquitous nature of the $(\beta\alpha)_8$-barrels is also evident from the estimation that this fold may account for 10% of all enzymes (Farber and Petsko, 1990; Gerlt, 2000) covering an enormous range of biochemical reactions (Pujadas and Palau, 1999; Nagano *et al.*, 2002). For instance, the $(\beta\alpha)_8$-fold is the structure with the highest number of functions including sixteen different enzyme activities covering the EC primary numbers one to

five (Gerstein, 1998; Hegyi and Gerstein, 1999; Nagano *et al.*, 2002). The functional diversity shown by the $(\beta\alpha)_8$-barrel enzymes is reflected in the eighteen superfamilies of $(\beta\alpha)_8$-barrels, which comprise 76 different sequence families (Nagano *et al.*, 2002). Hence the $(\beta\alpha)_8$-fold has been dubbed *superfold*, as defined by Orengo *et al.* (1994), i.e. a stable structural arrangement that unite many distinct sequence families.

The centrality of the $(\beta\alpha)_8$-barrels to life can be further inferred from its broad occurrence in different forms of life including the three cellular domains (Gerstein, 1998; Hegyi and Gerstein, 1999). In addition, the $(\beta\alpha)_8$-barrel enzymes typically participate in metabolic pathways that are considered to be ancestral, such as the glycolytic pathway (Copley and Bork, 2000) and the biosynthesis of the amino acid histidine, an ancestral metabolite that is believed to be present in the prebiotic soup (Alifano *et al.*, 1996; Lazcano and Miller, 1999). Therefore, it can be concluded that the $(\beta\alpha)_8$-fold was one of the earliest structural solutions to appear during the course of evolution. Hence, understanding the evolution of this ancient protein fold has a bearing on the mystery of life sustained by a metabolic network catalysed by enzymes.

1.2.3 Applied aspects of the $(\beta\alpha)_8$-barrels

The $(\beta\alpha)_8$-fold faultlessly exemplifies the *one fold – many functions* paradigm (Nagano *et al.*, 2002) a fact that has been considered as a reflection of the plasticity of this protein structure (Todd *et al.*, 2001; Gerlt and Babbitt, 2001a and b; Nagano *et al.*, 2002; Gerlt and Raushel, 2003). In other words, the $(\beta\alpha)_8$-barrel scaffold is well-suited to enzyme evolution, which in addition to being a key asset for evolution of life on Earth, has its applied side. Recently, the scaffold of the $(\beta\alpha)_8$-barrels has gained

increasing interest from the protein engineering community since it is anticipated that it will offer a rational framework for the design of novel proteins with application as biocatalysts (Petsko, 2000; Gerlt, 2000; Gerlt and Babbitt, 2001a and b; Höcker *et al.*, 2001a; Gerlt and Raushel, 2003). The increasing interest comes after the apparent demonstration that it is possible to rationally direct the evolution of the activity of a $(\beta\alpha)_8$-barrel enzyme involved in tryptophan biosynthesis from the scaffold of a related barrel involved in the same pathway (Altamirano *et al.*, 2000; see also Section 2.2.2.1). The possibility of a rational approach of directing the evolution of the $(\beta\alpha)_8$-barrels adds a new alternative to the more conventional methods based on random mutagenesis (e.g. Jürgens *et al.*, 2000).

1.3 *Streptomyces coelicolor:* a 'multitalented' organism

1.3.1 Taxonomy and characteristics of the *Streptomyces* genus

Streptomyces is the largest eubacterial genus including approximately 450 different species (Miyadoh *et al.*, 1997). It is probably safe to state that this genus is the most notorious subgroup of organisms of the Gram-positives with high (G + C) content or actinomycetes (Miyadoh *et al.*, 1997; Kieser *et al.*, 2000, pp. 2-6). However, the genus *Streptomyces* is taxonomically hitherto the sole member of the family *Streptomycetaceae* (Miyadoh *et al.*, 1997; Anderson and Wellington, 2001). The streptomycetes are ubiquitous and can be isolated from a widespread of natural environments, soil being the main ecological niche colonised by these organisms where

they act as general saprophytes (see reviews by Hodgson, 1992 and 2000). Although there are some pathogenic streptomycetes (e.g. *S. scabies* is a plant pathogen; cited in Kieser *et al.*, 2000, pp. 6) the main threat of actinomycetes species to human health are members of other genera, such as *Corynebacterium*, *Nocardia* and *Mycobacterium*.

Members of the *Streptomyces* genus exhibit a striking variety of morphological and physiological characteristics. Morphologically, the genus is characterised by a mycelial growth involving different cellular forms: after germination of a spore, mycelium substrate develops and lyses in order to support growth of aerial mycelium, from which chains of spores are produced after septation of aerial hyphae (see Figure 1.5). The developmental biology of *Streptomyces*, using *S. coelicolor* as a model system, has been extensively studied for the last three decades or so (reviewed by Chater, 1993 and 2001). These studies have lead to the identification of a large number of genes, namely, the *whi* and *bld* genes, whose products interact within a complex regulatory network controlling the multicellular development of *Streptomyces*. Other factors, which act as extracellular signals in the initiation of aerial mycelium formation, have also been identified in different species (reviewed by Chater and Horinouchi, 2003). These factors notoriously include the γ-butyrolactones, a family of structurally related but diverse metabolites first isolated from *S. griseus* (Khokhlov *et al.*, 1967), which have been demonstrated to regulate mycelium development in addition to secondary metabolism (reviewed by Horinouchi, 2002).

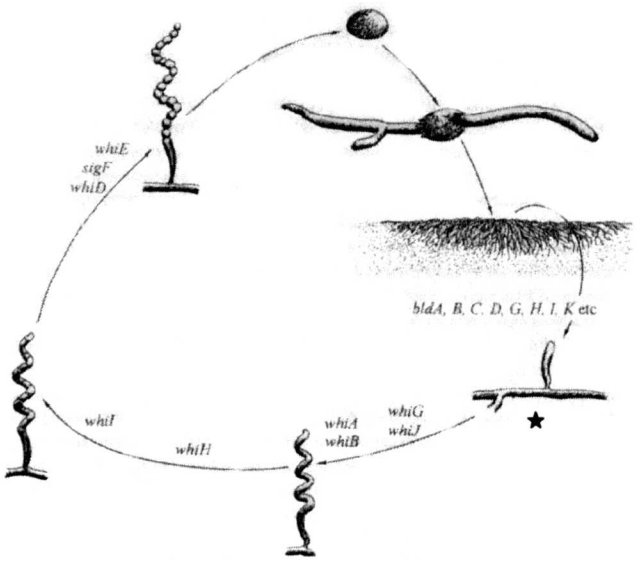


Figure 1.5. The cell cycle of *Streptomyces*. The prototypical life cycle of *Streptomyces*, based on data from *S. coelicolor*, is shown. Germ tubes emerge from a spore and grow by tip extension and branching to give rise to a substrate mycelium. Then, aerial hyphae grow up in a process that involves the action of a large number of *bld* genes. Aerial mycelium forms a spiral syncytium and when it stops growing multiple septa subdivide into pre-spore compartments. These spores subsequently mature and the grey spore pigment appears. Sporulation depends on the action of the products of the *whi* genes and the sigma factor *sigF*. The onset of secondary metabolism is highlighted with a star. Adapted from Kieser *et al.* (2000, pp. 19).

Another interesting characteristic of the *Streptomyces* genus is the observation that its morphological development is intrinsically correlated with the physiological state of the cell. In addition to a primary level of metabolism, which shows significant differences to other eubacteria (see below Section 1.3.1.1), the genus *Streptomyces* is characterised by a secondary level of metabolism. This latter physiological state only starts until formation of aerial mycelium in solid medium (Figure 1.5) and in the transition from exponential growth to stationary phase in liquid culture. Despite the fact that little is known about secondary metabolism's *raison d'être* in *Streptomyces*, i.e. the biological role of secondary metabolites in real environmental conditions, this characteristic has made this genus very popular amongst industrial microbiologists. This popularity rests on the fact that the *Streptomyces* genus has served for more than fifty years as the main source of natural products with relevant biological activities for the pharmaceutical and agrochemical industries (Demain, 1999; Behal, 2000). These

biologically active metabolites originally were limited to antibiotics, such as streptomycin produced by *S. griseus*, which was clinically used to treat tuberculosis. However, several other biological activities have been later on discovered and exploited by the biotechnological industry as well (reviewed by Demain, 1999; Behal, 2000; Kieser *et al.*, 2000, pp. 10-16). Examples of these secondary metabolites include the immunosuppressant agent FK506 produced by *S. tsukubaensis*; the anti-tumour agent daunorubicin produced by *S. peucetius*; the antihelminthic agent avermectin produced by *S. avermitilis*; and the herbicide bialaphos produced by *S. hygroscopicus*. Moreover, it has been estimated that the potential of this genus for the discovery of novel natural products has not been as yet exhausted (Watve *et al.*, 2001; Demain, 2002). The reader is referred to the *Streptomyces* genetics manual published by the John Innes Centre for a very comprehensive introduction to this genus (Kieser *et al.*, 2000, pp. 1-41).

1.3.1.1 Amino acid metabolism in *Streptomyces*

Despite the fact that there is a decent understanding of how the biosynthesis of secondary metabolites is directed and regulated at the genetic and biochemical levels (see reviews by Pfeifer and Khosla, 2001; Lautru and Challis, 2003), relatively little is known about primary metabolism in streptomycetes, which serves as source of precursors for synthesis of secondary metabolites. Hodgson (2000) reviewed primary metabolism in this group of organisms for the first time. This review brought to light the observation that regulation of amino acid metabolism in streptomycetes differs to the typical *E. coli* paradigm of feedback regulation. It can be generalised that amino acid biosynthesis in *Streptomyces* is constitutively expressed and therefore very few

examples in which there is clear evidence of feedback regulation have been reported (see Figure 1.6). Hodgson (2000) also concluded that around half of the amino acid catabolic pathways are constitutive while the other half is induced by the cognate amino acid. It seems therefore that while *E. coli* and other copiotrophs have developed tight regulatory systems to react efficiently to the constant 'feast' time seen in the gut, and 'famine' time seen outside of the gut, for *Streptomyces* and other oligotrophs it has not been worthwhile to invest in amino acid biosynthesis genetic regulation. This has been explained from the fact that soil is poor in nitrogen so a constant 'famine' time for amino acids would be expected (Hu *et al.*, 1999; Hodgson, 2000). Overall, it seems that streptomycetes have opted for a global regulatory system where amino acid metabolism is regulated by growth phase and growth rate (Hood *et al.*, 1992; Hu *et al.*, 1999). The general aim of this thesis is to contribute to the understanding of amino acid metabolism in streptomycetes, using *S. coelicolor* A3(2) as a model system, within a genomics framework.

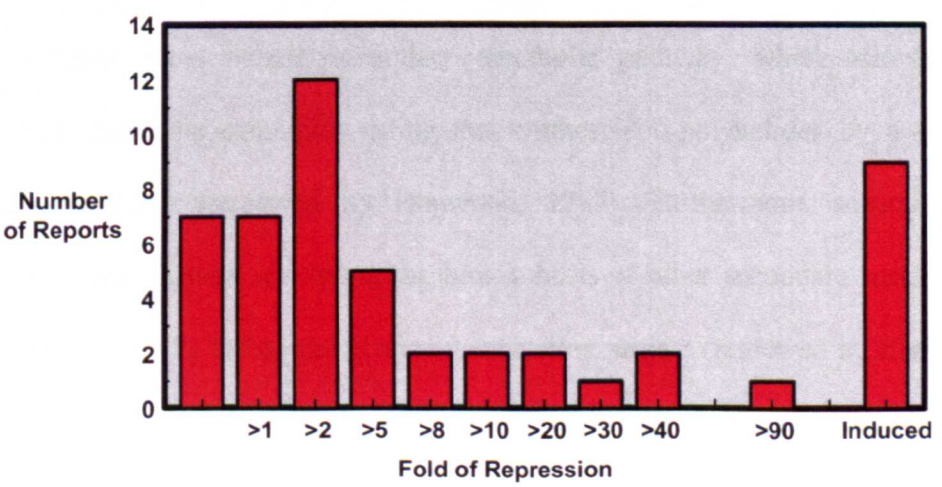


Figure 1.6 Repression of amino acid biosynthesis in streptomycetes. The graph shows the number of reports dealing with amino acid biosynthesis in streptomycetes against the fold of repression reported (best repressor chosen if more than one). Please note that in a significant number of reports induction was detected. Figure kindly provided by D.A. Hodgson.

1.3.2 *Streptomyces coelicolor* A3(2): a model system

Among the large number of streptomycetes that have been discovered hitherto *S. coelicolor* A3(2) has received the largest amount of attention. Without doubt it can be safely stated that this strain is the actinomycete genetically best characterised (Hopwood, 1999). Since the pioneering studies by Hopwood more than forty years ago *Streptomyces* have gone through different genetics eras: from *in vivo* through *in vitro* to *in silico* (see Hopwood (1999) for an excellent historical account of the genetics of *Streptomyces* with emphasis on A3(2)). The consolidation of A3(2) as a model system is a paradoxical situation since this strain does not produce any commercially important secondary metabolite, the main motivation for the study of streptomycetes (see Section 1.3.1). However, it is interesting to note that one of the reasons why Hopwood decided to work on A3(2) is that this strain produced a blue-pigmented secondary metabolite, which could be used as a genetic marker (Hopwood, 1999). The blue pigment, which gave name to this isolate (*coelicolor* = sky blue colour), turned out to be the aromatic polyketide actinorhodin (called ACT; *act* genes). Actinorhodin was the first example of cloning of genes for a whole secondary metabolic pathway, which allowed the elucidation of the basic principles ruling the synthesis of polyketides by a type II polyketide synthetase (reviewed by Hopwood, 1997). Furthermore, several other chromosomal gene clusters specifying the biosynthesis of other secondary metabolites were identified in A3(2) before the genome-sequencing project (reviewed by Bentley *et al.*, 2002). These metabolites include the complex of the red-pigmented tripyrrole prodiginines (called RED; *red* genes, see also Chapter VI); the non-ribosomal peptide calcium-dependent antibiotic (called CDA; *cda* genes, see also Chapter V); the plasmid-

borne methylenomycin (*mmy* genes); and the grey spore pigment (*whi* genes). Figure 1.7 shows colonies of *S. coelicolor* producing ACT and RED.

In contrast to other actinomycetes *S. coelicolor* is amenable to genetic manipulation, a situation that has been exploited for the development of a large collection of genetics tools (Kieser *et al.*, 2000). As discussed in Section 1.1.3 the *Streptomyces* chromosome is fairly unstable and can undergo huge deletions and amplifications (reviewed by Chen *et al.*, 2002), an observation that appears to be related to the fact that *Streptomyces* contains a huge linear chromosome (of around 8.7 Mb for *S. coelicolor*). This feature escaped preliminary genetic analysis and was not discovered until the use of pulsed-field gel electrophoresis in *S. lividans* 66 (Lin *et al.*, 1993), a close relative of *S. coelicolor* A3(2). By analogy with the results obtained from studies on huge linear plasmids, also present in *Streptomyces* species (e.g. SCP1 in *S. coelicolor* is 365 kb), a chromosomal replication model was proposed (Chang and Cohen, 1994). In this model, conventional bidirectional replication takes place from the centrally located origin of replication (*oriC*). Replication is followed by the so-called *end-patching* replication primed by proteins attached covalently to the terminal inverted repeats (TIR) at the free 5' ends of the chromosome to fill the gap left by removal of the RNA primer for the last Okazaki fragment on each discontinuous strand. This model proved to be correct when the experiments were done on the chromosome (Bao and Cohen, 2001 and 2003).

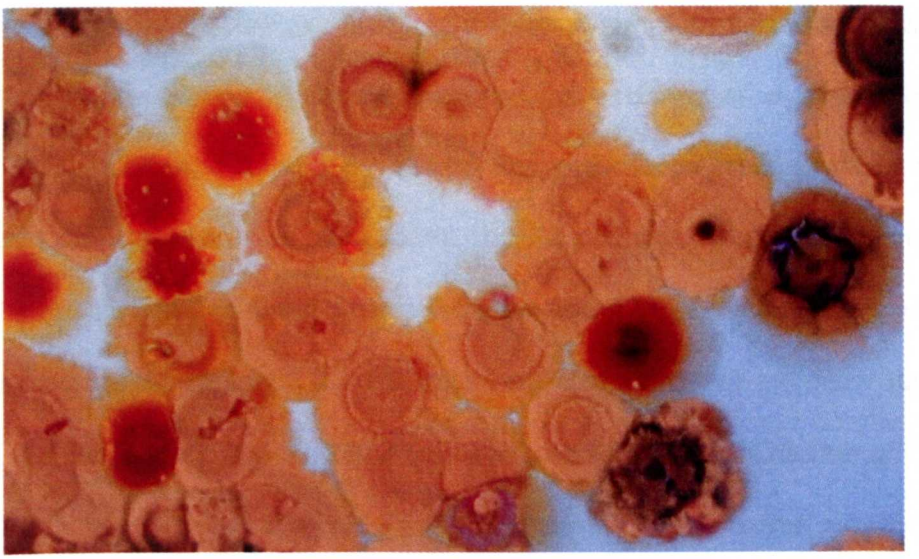


Figure 1.7 Colonies of *S. coelicolor* producing secondary metabolites. The picture shows three different strains derived from A3(2) growing together (see Chapter VI). The production of the blue-pigmented ACT can be seen in the background as it migrates into the medium and as dark blue regions in the colonies towards the right-hand side of the picture. RED can be seen in the bright-red colonies lacking sporulation.

1.3.2.1 Genomics of *S. coelicolor*

Sequencing of the entire genome of *S. coelicolor* by the Sanger Centre (http://www.sanger.ac.uk/Projects/S_coelicolor/) started in 1997 and it was not completed until 2001 (Bentley *et al.*, 2002). The genome sequence was obtained by sequencing an ordered cosmid library that covered the entirety of the chromosome of a plasmidless prototrophic derivative of A3(2), termed M145 (Redenbach *et al.*, 1996). Due to this reason M145 was the strain of choice for this thesis and the cosmids prepared by Redenbach *et al.* (1996) were extensively used. The great impact of the genome sequence to the field was immediately anticipated (for a comment see Thompson *et al.*, 2002; Donadio *et al.*, 2002). This impact was enhanced by the fact that the annotated sequences generated by the Sanger Centre were immediately made public. Examples of reports that exploited the sequence data as it was generated are innumerable and the first

publications dealing with the so-called *omics* technologies, i.e. proteomics (Hesketh *et al.*, 2002; Novotna *et al.*, 2003) and transcriptomics (Huang *et al.*, 2001; Bao and Cohen, 2003), have also appeared. Indeed, the genome sequence proved to be an invaluable resource for this thesis. All the experiments reported herein were conceptually conceived within a genomic framework. The genomics of *S. coelicolor*, with emphasis on the technologies that were generated as part of a broader functional genomics initiative in the United Kingdom, have been reviewed by Chater *et al.* (2002).

There are several discoveries from the analysis of the genome sequence alone that are revealing about the biology of *S. coelicolor* (Bentley *et al.*, 2002; see also the review by Paradkar *et al.*, 2003). Remarkably, the number of coding sequences found in the chromosome of this bacterium, at 7825 predicted proteins, is roughly twice as many as in *E. coli* (4289) and *M. tuberculosis* (3959) and almost one quarter more than in the eukaryote *Saccharomyces cerevisiae* (6203). In Addition, the number of regulatory proteins identified (965), representing 12.3% of the total, is truly remarkable, a fact that is likely to be related to the complex morphology and physiology of this organism (see Section 1.3.1). These proteins include ECF sigma factors, transcriptional regulators, two-component systems, serine/threonine protein kinases and DNA-binding proteins. It was also surprising to discover that, in addition to genes directing the synthesis of the five secondary metabolites that were known to be produced by A3(2) (see Section 1.3.2), the genome sequence contained at least a further eighteen gene clusters that potentially encode sets of enzymes characteristic of secondary metabolism (Bentley *et al.*, 2002). This characteristic, together with the discovery of a myriad of secreted enzymes, such as general hydrolases, chitinases and celullases, are likely to be related to

the life-style of this organism and to its extraordinary capacity to colonise the soil (see Sections 1.3.1 and 1.3.1.1).

The genome sequence of *S. coelicolor* has also proved to be useful within the context of comparative genomics. For example, the genome sequence of *S. coelicolor* has been used to study the evolution of the genome of actinomycetes at the cladistic level. Bentley *et al.* (2002) recognised that on the basis of the essentiality of the genes encoded the linear chromosome of *S. coelicolor* can be divided into two different classes of genomic regions: a *core* region that appears to extend from around 1.5 Mb to 6.4 Mb and two uneven *arms* with lengths of approximately 1.5Mb (left arm) and 2.3 Mb (right arm). The core region was predicted to encode genes for essential functions, such as DNA processing and amino acid metabolism. In contrast, the arms were found to encode non-essential functions such as secondary metabolism and exoenzymes. Moreover, it was found that the core region of *S. coelicolor* showed conserved synteny with the entire chromosomes of the pathogenic actinomycetes *Corynebacterium diphtheriae* and *Mycobacterium tuberculosis*, which are circular (Bentley *et al.*, 2002). This observation suggests that the core region of the genome of *S. coelicolor* may reflect the ancestral chromosome of the actinomycetes and that the DNA present in the arms has been recently acquired, probably by LGT (Bentley *et al.*, 2002) *via* integration of a linear plasmid that linearised the actinomycetes ancestral circular chromosome (Chen *et al.*, 2002). Interestingly, some of the regions that showed the greatest synteny between the different actinomycetes that were compared encode amino acid metabolism functions (Bentley *et al.*, 2002). Recently, a second genome of a streptomycete was fully sequenced, namely, the avermectin-producer *S. avermitilis* (Ōmura *et al.*, 2001; Ikeda *et*

al., 2003). It is anticipated that this advancement will give a further boost to the field of actinomycetes' comparative genomics (for a comment see Hopwood, 2003).

Chapter II

Chapter II

"Discovery of the enzyme phosphoribosyl isomerase (PriA) involved in both tryptophan and histidine biosynthesis"

2.1 Introduction

Back in the 1970s it was demonstrated that the enzyme activity *N'*-(5'-phosphoribosyl)anthranilate (PRA) isomerase (TrpF, EC 5.3.1.24) encoded by *trpF* was detected in cell-free extracts of *S. coelicolor* (Smithers and Engel, 1974). This situation was also found in *S. venezuelae* (Paradkar *et al.*, 1991). TrpF is responsible for the third step of the tryptophan biosynthetic pathway and represents the fourth enzymic function in the synthesis of tryptophan from chorismate (Yanofsky, 2001 and 2003; see also Figures 4.6 and 5.1). TrpF isomerises the aminoaldose PRA into the aminoketose 1-[(2-carboxyphenyl)amino]-1-deoxyribulose 5-phosphate (CDRP) via an Amadori rearrangement (Figure 2.1; Kirschner *et al.*, 1987). In eubacteria the *trpF* gene is often associated with *trpC*. For example, in the γ subdivison of purple eubacteria (enteric bacteria) these genes are fused in a single polypeptide, TrpC-F, while in the low (G+C) content subdivison of Gram-positive eubacteria (*Bacillus* and relatives) they are organised next to each other, being *trpC* promoter proximal (Crawford, 1989; Doolittle *et al.*, 1992).

In the early genetic map of the *trp* genes in *S. coelicolor* reported by Smithers and Engels (1974) the *trpC-F* fusion observed in enteric bacteria and the *trpCF* gene

order characteristic of *Bacillus* were not implied. The *trpF* gene was positioned about 10 o'clock in the chromosome together with *trpD* and far apart from *trpC* (Smithers and Engel, 1974). Unfortunately, the *S. coelicolor* Trp7 strain deficient in PRA isomerase activity (Smithers and Engel, 1974) used for the crosses is not available (D. Hodgson, personal communication) removing the possibility of cloning the *trpF* gene by complementation.

During the course of the investigations on the regulation of tryptophan biosynthesis in *S. coelicolor* in Hodgson's laboratory (Hood *et al.*, 1992; Hu, 1995; Hu *et al.*, 1999) a suggestion that the *trpF* gene was missing started to take root. No hint of a *trpF*-like sequence was found in the flanking regions of *trpC* as expected by analogy to other bacteria (Hood *et al.*, 1992). Moreover, sequencing of the ORFs flanking *trpD* (Hu, 1995; Hu *et al.*, 1999) did not lead to the discovery of a *trpF* gene as suggested by the genetic map of Smithers and Engel (1974). This paradox of a function without a gene was confirmed after completion of the sequencing of the complete genome of *S. coelicolor* M145 (Bentley *et al.*, 2002). A similar situation was encountered in the genomes of the actinomycetes *Mycobacterium tuberculosis* H37Rv (Cole *et al.*, 1998) and *Mycobacterium leprae* (Cole *et al.*, 2001) where no *trpF* gene was identified (see also Chapter IV). These observations taken together indicate that the *trpF* genes of actinomycetes are of a different class to those found in other bacteria. A caveat to this generalisation is the discovery of a *trpF* gene in *Corynebacterium glutamicum* (Matsui *et al.*, 1986), although this sequence proved to be strikingly similar to those found in enteric bacteria and point to the possibility that lateral gene transfer had occurred (Crawford, 1989; see also Section 4.2.2.3).

This Chapter reports the basis of PRA isomerase or TrpF activity in *S. coelicolor*. The extensive genetic and biochemical understanding of the biosynthetic pathway of tryptophan (reviewed by Yanofsky, 2001 and 2003) was exploited in order to propose hypotheses aimed to assign a gene to the enzyme function PRA isomerase. As a result, compelling genetic evidence *in vivo* is provided to support the view that a *hisA* homologue (SCO2050) putatively encoding *N'*-[(5'-phosphoribosyl)formimino]-5-aminoimidazole-4-carboxamide ribonucleotide (ProFAR) isomerase (HisA, EC 5.3.1.16) participates not only in histidine biosynthesis, but also in tryptophan biosynthesis (Figure 2.1). The evolutionary implications of this discovery and the results presented in this Chapter have been published elsewhere (see Appendix 1; Barona-Gómez and Hodgson, 2003).

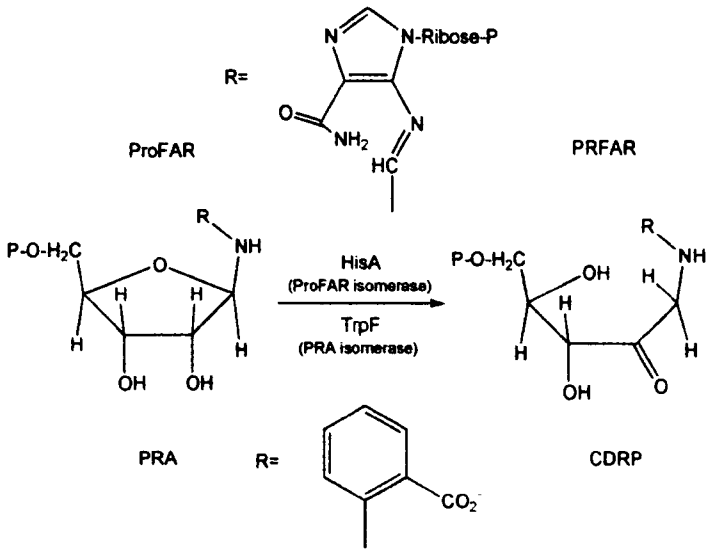


Figure 2.1 Biochemical reactions performed by TrpF and HisA. TrpF and HisA catalyse analogous Amadori rearrangements (Jürgens *et al.*, 2000; Henn-Sax *et al.*, 2002) in the biosynthesis of the aromatic amino acids tryptophan and histidine, respectively. The aminoaldoses *N'*-(5'-phosphoribosyl)anthranilate (PRA) and *N'*-[(5'-phosphoribosyl)formimino]-5-aminoimidazole-4-carboxamide ribonucleotide (ProFAR) are isomerised into their corresponding aminoketoses 1-[(2-carboxyphenyl)amino]-1-deoxyribulose 5-phosphate (CDRP) and *N'*-[(5'-phosphoribulosyl)formimino]-5-aminoimidazole-4-carboxamide (PRFAR) by TrpF (Kirschner *et al.*, 1987) and HisA (Margolies and Goldberger, 1967), respectively.

2.2 Results and Discussion

2.2.1 Functional genomics *in silico* of *trpF*

The standard method used for assigning function to a gene that has just been identified (e.g. by a genome-sequencing project) involves comparison of its sequence against the totality of the sequences deposited in the database using mathematical algorithms. Putative functions of the new genes are assigned on the basis of similarity with genes of known function (Parkhill, 2002; Primrose and Twyman, 2003). The case of the lack of *trpF* represents the opposite situation: a known function remains to be assigned to a gene. Therefore, a different approach for doing functional genomics *in silico* was required. It was hoped that a *trpF* sequence from the actinomycete *C. glutamicum* (Matsui *et al.*, 1986) could be used to identify a *trpF*-like sequence in the genome sequence of *S. coelicolor*. Despite the proposal that this gene of *C. glutamicum* has been acquired by lateral gene transfer, this sequence represents a suitable probe since it shows the typical high (G + C) content of the actinomycetes (Crawford, 1989). Two different algorithms were used: the standard BLAST (Altschul *et al.*, 1997) and the more powerful hidden Markov models (HMM; Karplus *et al.*, 1998). Both of them failed to identify a *trpF* homologue in the genome of *S. coelicolor*. This result supports the original prediction that the *trpF* gene of *S. coelicolor* is of a different class. Thus, novel approaches were required if this gene was to be identified *in silico*. Different scenarios for the evolution of TrpF were proposed and strategies for computational searches were designed based upon these scenarios.

2.2.2 Evolution driven hypotheses for the identification of *trpF*

The debate on the evolution of metabolic pathways and enzymes is seen in the proposed paths for the evolution of tryptophan biosynthesis (Copley and Bork, 2000; Henn-Sax *et al.*, 2001 and references therein). There are two main possible evolutionary paths that could have accounted for the appearance of this pathway. If the *retrograde* hypothesis is assumed, i.e. the last enzyme of the pathway arose first and the rest of the pathway evolved on a backward stepwise fashion (Horowitz, 1965), TrpF may have evolved from TrpC (Figure 2.2). On the other hand, if tryptophan biosynthesis evolution followed the *patchwork* hypothesis proposed by Jensen (1976) it would imply that TrpF arose by recruitment of an enzyme function more likely to be involved in another metabolic pathway performing an analogous reaction, i.e. an Amadori rearrangement. See also Sections 1.1.5.1 and 1.1.6 in Chapter I.

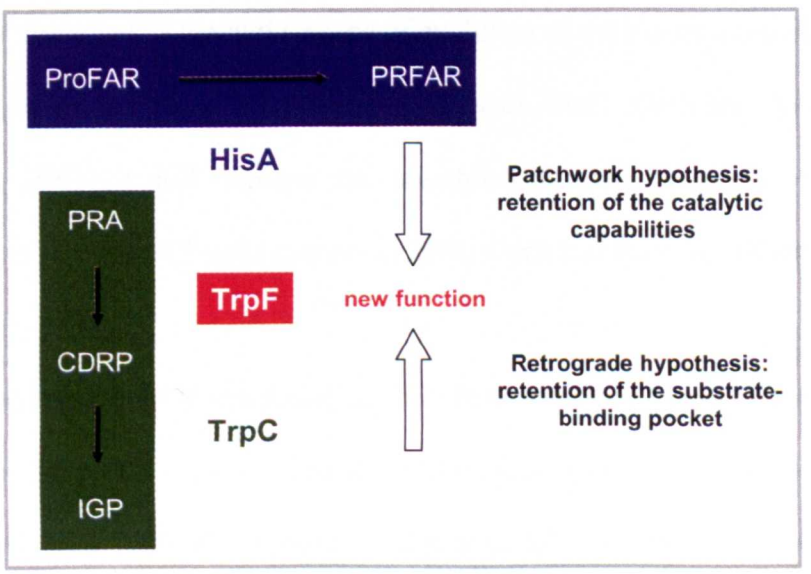


Figure 2.2 Evolutionary scenarios of the appearance of TrpF. Two alternative paths that may have accounted for evolution of TrpF are represented schematically. Steps of tryptophan biosynthesis are depicted in green: PRA → CDRP → IGP (indole-3-glycerol phosphate). Following a retroevolution flow (from green to red - arrow) TrpF could have arisen from TrpC. A single step of histidine biosynthesis is depicted in blue: ProFAR → PRFAR. If the evolution flow is dictated by recruitment of an enzyme function from a different pathway (from blue to red - arrow) TrpF could have arisen from HisA.

2.2.2.1 Examination of the retrograde evolution hypothesis

Based on sequence and structural homology TrpF and TrpC have been previously proposed to share a common ancestry (Wilmanns *et al.*, 1991; Eder and Kirschner, 1992; Gerlt and Babbitt, 2001b). Although TrpF and TrpC share low levels of sequence similarity (22% identity) they are the closest homologues of each other (Gerlt, 2001). Moreover, both TrpF and TrpC are known to be members of the $(\beta\alpha)_8$-barrel structural family (Priestle *et al.*, 1987; Wilmanns *et al.*, 1992) and to show the standard phosphate-binding (SPB) motif characteristic of phosphate-binding enzymes (see Section 3.1.4; Wilmanns *et al.*, 1991; Bork *et al.*, 1995). Therefore, a retroevolution of TrpF from TrpC, implying retention of the substrate-binding domain and modification of the catalytic capabilities of TrpC to perform the new TrpF function, appears to be plausible (Figure 2.2). Although this type of evolution has been categorically discarded as the *main* means of evolution of the extant metabolic pathways in general (Copley and Bork, 2000; Teichmann *et al.*, 2001; Gerlt and Babbitt, 2001b; Todd *et al.*, 2001) it still remains one possibility for the evolution of tryptophan biosynthesis in particular (Teichmann *et al.*, 2001; Gerlt and Babbitt, 2001b; Todd *et al.*, 2001).

The hypothesis of the evolution of TrpF from TrpC has been previously tested in the laboratory. It appears that, at least tentatively, it is feasible to evolve TrpF activity from the scaffold of TrpC after random mutagenesis followed by DNA shuffling and *in vivo* selection (Altamirano *et al.*, 2000). Unfortunately, all the authors of this report other than M.M. Altamirano retracted their original conclusions when they discovered that one of their recently evolved TrpF clones was contaminated with a *trpF* wild type

sequence (Altamirano *et al.*, 2002). Nevertheless, one of the authors of the work has stated, "there is no reason to believe that it will not work again" (J. Blackburn, personal communication).

We tested therefore whether or not the sequence of TrpC (SCO2039) of *S. coelicolor* could reveal a *trpF*-like sequence of an actinomycete in the database by performing iterative searches using the algorithm PSI-BLAST (Altschul *et al.*, 1997). Again, no *trpF*-like sequence from any actinomycete was identified before convergence was reached, despite the fact that other TrpF sequences from other organisms were identified. This result supports the view that in some actinomycetes the *trpF* gene is of a different class and hinted to the possibility that a *non-orthologous gene displacement*, as defined by Koonin *et al.* (1996) had occurred.

2.2.2.2 Exploitation of the patchwork evolution hypothesis

The patchwork hypothesis assumes the ability of enzymes to perform analogous reactions in different biochemical pathways (Jensen, 1976). A corollary of this hypothesis is that ancestral enzymes with broad-substrate specificity existed at some point during the course of evolution. These enzymes would serve as starting material for the evolution of new enzyme functions following a process of gene duplication and subsequent specialisation (Jensen, 1976; see also Section 1.1.6). When applying this model to the functional genomics discrepancy of *trpF* in *S. coelicolor*, a scenario on which TrpF arose from an enzyme performing an Amadori rearrangement can be envisaged. This implies that TrpF in *S. coelicolor* might appear homologous to an

isomerase rather than to a synthase (i.e. TrpC), as suggested by the retrograde hypothesis.

In general terms TrpF isomerises a sugar, a reaction that is not rare in primary metabolism, but limited to a comprehensible number (Michal, 1992). Furthermore, the phosphoribosyl nature of the substrate of TrpF can be taken into account to pair TrpF with analogous enzymes. One reasonable candidate to be analogous with TrpF is ProFAR isomerase of histidine biosynthesis, which is encoded by *hisA*. HisA isomerises the aminoaldose ProFAR, a phosphoribosyl ligand similar to PRA, into the aminoketose *N'*-[(5'-phosphoribulosyl)formimino]-5-aminoimidazole-4-carboxamide (PRFAR) (Figure 2.1; Margolies and Goldberger, 1967). Moreover, both TrpF (Priestle *et al.*, 1987; Wilmanns *et al.*, 1992) and HisA (Lang *et al.*, 2000) are known to be $(\beta\alpha)_8$-barrels containing the SPB motif characteristic of phosphate-binding enzymes (see also Section 3.1.4).

The similarity between the barrels encoded by *hisA* and *trpF* was exploited by Jürgens *et al.* (2000) to direct the evolution of TrpF activity from the scaffold of HisA by random mutagenesis and selection *in vivo*. It appears that TrpF activity was obtained by exchange of a single residue in the scaffold of HisA, albeit at the expense of its original activity. Moreover, it seems that one of the mutant enzymes showed that both TrpF and HisA activities could co-exist within the same protein (Jürgens *et al.*, 2000). These observations suggest that HisA and TrpF share a common ancestry and, therefore, that they are evolutionary related. Along these lines, the conservation of the mechanism of the reactions catalysed by HisA and TrpF has been recently confirmed (Henn-Sax *et al.*, 2002). These data are in accordance with the patchwork hypothesis, and it was the

basis in this study for the proposal of the hypothesis that in *S. coelicolor* a HisA paralog has PRA isomerase activity, which would account for the lack of *trpF* in this organism.

More than a decade ago a *hisA* homologue was cloned from *S. coelicolor* as part of a histidine biosynthetic cluster (Limauro *et al.*, 1990). Unfortunately, no functional characterisation was reported. Please note that the *hisA* gene encoding for the enzyme ProFAR isomerase is not related to the genetic marker *hisA* (alleles *his*-1, 120 and 132) historically used in the mapping experiments by Hopwood and co-workers (Hopwood *et al.*, 1973; Hopwood *et al.*, 1985). Following the annotation of the *his* genes in *E. coli* (Alifano *et al.*, 1996) the *S. coelicolor hisA* marker corresponds to *hisD*, which encodes for the bifunctional enzyme histidinol dehydrogenase (HisD, EC 1.1.1.23). Therefore, no real *hisA* mutant has been isolated until this study.

After completion of the genome-sequencing project of *S. coelicolor* (Bentley *et al.*, 2002) it became apparent that the histidine biosynthetic cluster cloned by Limauro *et al.* (1990) containing the *hisA* homologue (SCO2050), lay in the chromosome just upstream of the *trp* operon cloned in our laboratory (Hu, 1995; Hu *et al.*, 1999). The genomic association of genes in prokaryotes more often reflects functional association between their proteins. This situation is particularly recurrent in metabolic genes and has been exploited to predict function of such genes (for in-depth discussions of these concepts see Chapter IV and references therein). Therefore, when considering the genetic organisation of the *his* and *trp* genomic region of *S. colicolor* shown in Figure 2.3, the ORF SCO2050 becomes a good candidate for having PRA and ProFAR isomerase activities and to be involved in both histidine and tryptophan biosynthesis.

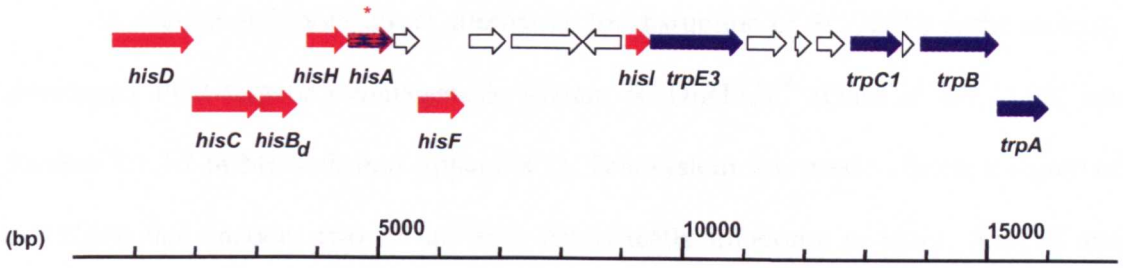


Figure 2.3 Genomic organisation of the *S. coelicolor his* and *trp* region. The majority of the genes required for synthesis of histidine (*his*; red) and tryptophan (*trp*; blue) cluster together in a genetic locus of around 16 kb. Genes of unknown function are uncoloured. The *hisA* homologue proposed to be involved in the biosynthesis of both amino acids is marked with an asterisk and coloured in blue and red. The Figure was constructed using the inverse complement of the sequence and taking into consideration the coding frame. The genes that are potentially translationally coupled are shown (*hisD*/*hisC*, *hisC*/*hisB*$_d$, SCO2051/*hisF* and *trpB*/*trpA*). For a discussion on the evolution of this cluster see Chapter IV.

2.2.3 Mutational analysis of SCO2050

In order to test the hypothesis of the bifunctionality of the *hisA* homologue in *S. coelicolor*, the gene SCO2050 was knocked out. Two different genetic approaches were tested. First, a *S. coelicolor* mutant library generated by massive transposon mutagenesis with Tn*4560* was screened by PCR amplification (Fowler, 2002; Gust *et al.*, 2003). In this system, the desired mutant is identified by using primers that anneal to the flanking regions of the gene of interest and Tn*4560* (see Section 7.1.3.1 in Methods). The specificity of the PCR primers was increased by targeting 20meres of nucleotides that are not repeated on the chromosome (T. Kieser, personal communication; see also Section 7.1.3.1 in Methods). Two putative positive PCR products were obtained and their sequences were determined in order to find out the exact position of the insertion of the Tn*4560* transposon. The genes disrupted were identified by BLAST searches as a putative transcriptional accessory protein (SCO6743) and one of the five *trpE*

paralogues (*trpE1* or SCO2117) present on the genome of *S. coelicolor*. Due to this negative result this strategy was abandoned.

A second and more direct alternative for disruption of SCO2050 is the recently developed PCR-targeting mutagenesis system REDIRECT® (Gust *et al.*, 2003; see Section 7.1.3.2 in Methods and Appendix 2). This system was used to delete a region of SCO2050 that encodes two of the three catalytically important residues, Asp130 and Thr166, identified by aligning the sequences of SCO2050 and HisA from *Thermotoga maritima* (Figure 2.4; Jürgens *et al.*, 2000; Henn-Sax *et al.*, 2002). The region deleted lies between Ala126 and Gly210. The overall strategy followed is shown in Figure 2.5. The ORF SCO2050 was first mutagenised in cosmid Sc4G6 (Redenbach *et al.*, 1996) by homologous recombination (double crossover) replacing the corresponding 255 base pairs from its 3' end. The *aac(3)IV* plus *oriT* disruption cassette from pIJ773 was removed by the FLP-recombinase system. The FLP-recombinase excision leaves behind an in-frame 'scar' of 81 nucleotides with no stop codons, which avoids possible polar effects on the *trp* genes.

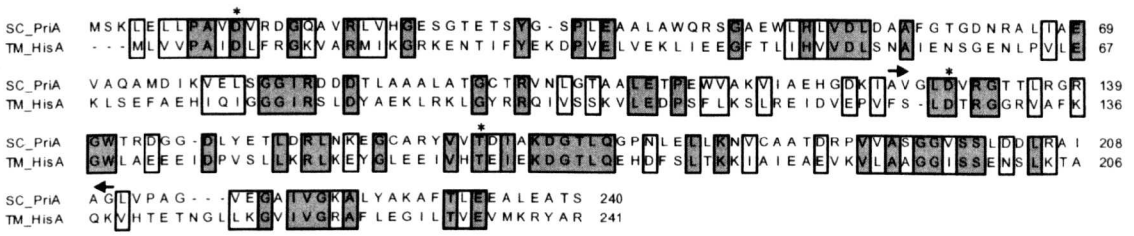


Figure 2.4 Sequence alignment of HisA from *T. maritima* and SCO2050. The alignment of the HisA homologues of *S. coelicolor* (SC_PriA) and *T. maritima* (TM_HisA) shows that these proteins are 25% identical. Identical (shadow) and similar (boxed) residues are shown. Data from *T. maritima* (Jürgens *et al.*, 2000; Henn-Sax *et al.*, 2002) was used to identify the catalytically important residues in SCO2050 (Asp11, Asp130 and Thr166 – asterisks). The replaced region in *S. coelicolor* WH101 (this study) spans from Ala126 to Gly210 (between arrows).

The mutagenised cosmid was re-engineered by replacing the kanamycin *neo* marker from the backbone of SuperCos1 with the selectable marker *aac(3)IV* and an RP4 *oriT*. The resulting construct was introduced to *S. coelicolor* M145 by RP4-based conjugation and selected with apramycin (Figure 2.5). Authentic double crossovers were not obtained until two rounds of growth on soya-flour-mannitol (SFM) plates supplemented with tryptophan and histidine, implying that loss of SCO2050 was selected against. The replacement of the targeted region of SCO2050 was implied in colonies that were apramycin sensitive as screened by replica plating. Interestingly, these colonies appeared to be consistently affected in growth and sporulation levels even when both amino acids were supplemented. The presence of the wild type or mutated form of SCO2050 was detected by PCR and the correct in-frame replacement was confirmed by sequencing.The apramycin sensitive colonies were picked and transferred onto a SFM-containing plate supplemented with histidine and tryptophan and allowed to grow until sporulation. The patches were screened for tryptophan and histidine auxotrophy by replica plating onto minimal medium (MM) plates. As hypothesised, the resulting mutant (named WH101) could not grow on minimal medium unless supplemented with both amino acids. Figure 3.6A shows growth of *S. coelicolor* WH101 in MM supplemented with or without histidine and tryptophan. These results are summarised in Table 2.1. At this point, SCO2050 was renamed *priA* (phosphoribosyl isomerase A) to reflect the common isomerase activities of TrpF and HisA upon phosphoribosyl ligands. The construction of the SCO2050 mutant was repeated using different source of cosmid and new PCR product obtaining the same result, making it unlikely that the phenotype observed was due to a second site mutation.

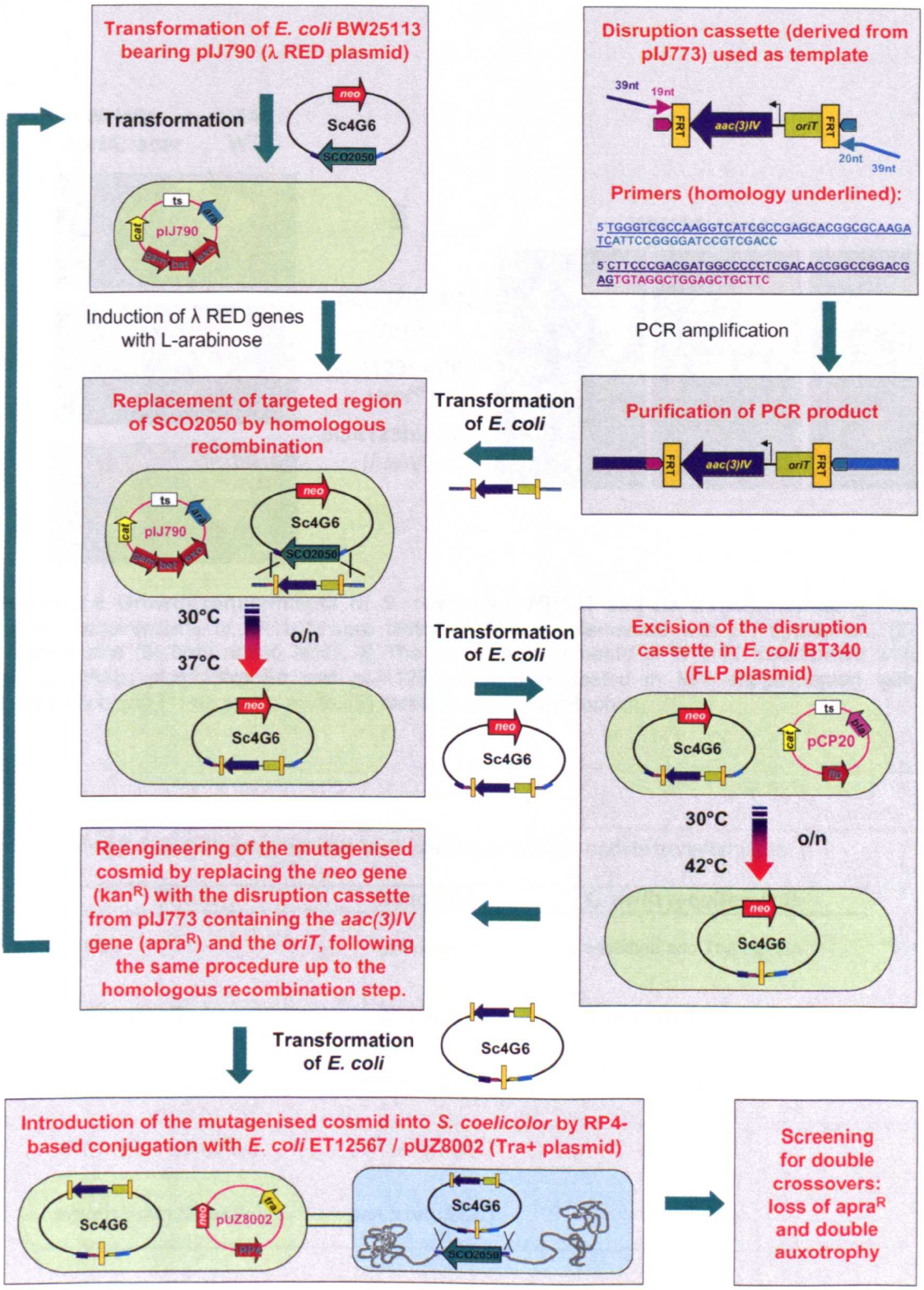


Figure 2.5 Strategy for the generation of a SCO2050 knockout mutation. For details and explanation refer to text and Section 7.1.3.2 in Methods and Appendix 2.

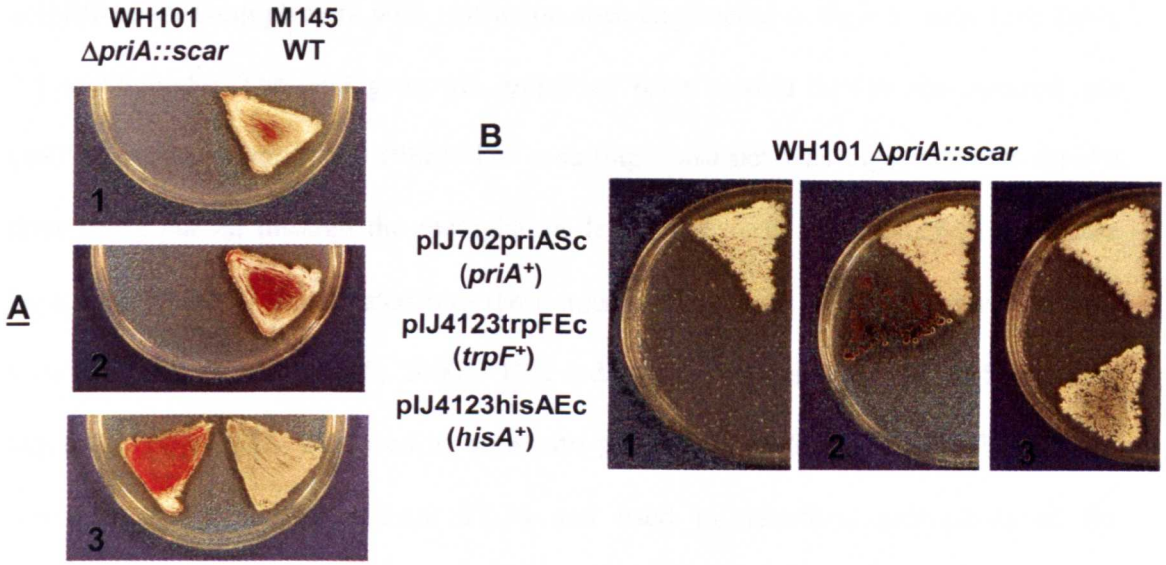


Figure 2.6 Growth requirements of *S. coelicolor* WH101 and its transformants. **A** The growth requirements of WH101 were tested in MM supplemented with (1) tryptophan, (2) histidine and (3) both amino acids. **B** The growth requirements of WH101 transformed with pIJ702priASc, pIJ4123trpFEc and pIJ4123hisAEc were tested in MM supplemented with thiostrepton and (1) no amino acids, (2) histidine and (3) tryptophan.

Table 2.1 Growth requirements of *S. coelicolor* WH101 and its transformants.

Strain	Genotype	Growth requirements
WH101	*priA*::scar	Histidine and Tryptophan
WH101(pIJ702)	*priA*::scar	Histidine and Tryptophan
WH101(pIJ702priASc)	*priA*::scar (*priA*$^+$)	Prototrophic
WH101(pIJ4123trpFEc)	*priA*::scar (*trpF*$^+$)	Histidine
WH101(pIJ4123hisAEc)	*priA*::scar (*hisA*$^+$)	Tryptophan

2.2.4 Complementation of *priA*

All the DNA fragments intended for cloning were prepared by PCR amplification using primers with restriction sites engineered at their 5' ends (see Table 7.1 in Methods). The *priA* gene was amplified from cosmid Sc4G6 and inserted into pMTL22 (Chambers *et al.*, 1988). The resulting construct was digested with *Bam*HI after being passed through the methylation-deficient *E. coli* host GM2929, to release *priA*, which was in turn ligated into the unique *Bgl*II site of the *Streptomyces* cloning vector pIJ702 (Kieser *et al.*, 2000). This led to the construct pIJ702priASc, which, together with pIJ702, was used to transform protoplasts of *S. lividans* TK24. Correct constructs were recovered from TK24 and used to transform protoplasts of the auxotrophic strain WH101. Prototrophic transformants of WH101 were obtained at a high efficiency only when pIJ702PriASc was used, demonstrating that the tryptophan and histidine auxotrophy of WH101 was due exclusively to the lack of *priA* (Figure 2.6B). The WH101 derivatives transformed with pIJ702 turned out to be auxotrophic for both amino acids.

It was tested whether the *trpF* and *hisA* genes of *E. coli* could complement the WH101 tryptophan and histidine auxotrophies. The cloning strategy was similar. The *hisA* gene was amplified from chromosomal DNA of the *E. coli* strain MC1061 and cloned into the expression vector pET-3a using the restriction sites *Nde*I and *Bam*HI engineered in the 5' ends of the primers used (see Table 7.1 in Methods). The *hisA* gene was cleaved from this construct using the same restriction enzymes and inserted into the *Streptomyces* expression vector pIJ4123 (Takano *et al.*, 1995) under the control of the thiostrepton-inducible promoter (Holmes *et al.*, 1993). An artificial gene that only

encodes the PRA isomerase function of the *trpC-F* fusion of *E. coli* had been constructed (Altamirano *et al.*, 2000). A derivative of this *trpF* gene was cleaved from pMS401 (see also Table 7.2 in Methods) with *Nco*I and *Bam*HI and sub-cloned into pET-22a, from where *trpF* was subsequently cleaved with *Nde*I and *Bam*HI and sub-cloned into the expression vector pIJ4123 (Takano *et al.*, 1995). Thiostrepton-dependent expression of *trpF* from *E. coli* (pIJ4123trpFEc) in WH101 restored tryptophan-independence alone, whilst expression of *hisA* from *E. coli* (pIJ4123hisAEc) restored histidine-independence alone. However, it seems that *trpF* from *E. coli* does not fully complement the PRA isomerase deficiency because sporulation and morphological development seems to be affected (Figure 2.6B). These results are summarised in Table 2.1.

2.2.4.1 Intergenic complementation of *hisA* and *trpF* mutants of *E. coli* by *priA*

The ability of *priA* to complement independent *hisA* and *trpF* mutations in *E. coli* was also investigated. For this purpose, *priA* was cloned into the expression vector pGEX-4T-1. This vector is suitable for complementation experiments in *E. coli* since expression of the insert from the *lacI*-dependent promoter does not require a particular genetic background. The *priA* gene of *S. coelicolor*, and the *hisA* and *trpF* genes of *E. coli* were cloned following a similar strategy. The restriction sites *Eco*RI and *Xho*I were introduced in the 5' ends of the cloning primers (see Table 7.1 in Methods). The construct bearing *priA*, pGEXpriASc, was used to transform *E. coli* auxotrophs with mutations in *trpF*, W3110 *trpC*(*F*del) (Darimont *et al.*, 1998), and *hisA*, Hfr G6 (Matney *et al.*, 1964) and the transformants were tested for tryptophan and histidine auxotrophy

in modified Vogel-Bonner (VB) and minimal agar (MA) media, respectively (see Section 7.2 in Methods). Complementation of both defects was achieved after incubation overnight by background levels of expression of *priA*, i.e. without induction by IPTG. For some unknown reason, induction with as little as 1 mM IPTG inhibited growth of the transformants. In accordance with this observation, Henn-Sax *et al.* (2002) also reported that when purifying HisA from *E. coli* they noticed that IPTG inhibited expression of *hisA*. The constructs bearing the *hisA* and *trpF* genes of *E. coli*, pGEXhisAEc and pGEXtrpFEc, served as controls. The results of these complementation studies are summarised in Table 2.2.

Table 2.2. Complementation of *E. coli* strains W3110 *trpC*(Fdel) and Hfr G6 by *priA*

Plasmid	**Strain**	
	W3110 *trpC*(Fdel)	**Hfr G6**
pGEX-4T-1	-	-
pGEXpriASc (*priA*$^+$)	+	+
pGEXhisAEc (*hisA*$^+$)	-	+
pGEXtrpFEc (*trpF*$^+$)	+	-

Complementation detected (+) or not detected (-) by expression of the insert (between parenthesis) from the *lacI*-dependent promoter without addition of IPTG.

2.2.5 An explanation of the bifunctionality of PriA

An obvious mechanistic explanation that might account for the bifunctionality of PriA is that this enzyme exhibits a broad specificity for the substrates PRA and ProFAR. This is in agreement with the broad-substrate specificity observed in other enzymes involved in primary metabolism (O'Brien and Herschlag, 1999). The broad occurrence of this mechanistic feature of enzymes is strong evidence supporting the patchwork

hypothesis and, therefore, stands at the core of the evolution of metabolic pathways. Examples of enzymes with broad substrate specificity involved in amino acid metabolism are some enzymes taking part in the biosynthesis of the branched-chain amino acids (Umbarger, 1996; see also Chapter VI) and glutamate aminotransferases (reviewed by Mehta *et al.*, 1993; Jensen and Gu, 1996).

The latter family of enzymes includes another example of an enzyme involved in the biosynthesis of both histidine and aromatic amino acids, namely, the product of *hisC* in *Bacillus subtilis* and *Zymomonas mobilis* (called *hisH* in *E. coli*, EC 2.6.1.9; Alifano *et al.*, 1996). The main role of HisC in these organisms is in histidine biosynthesis, as revealed by the phenotype of the corresponding mutants (Nester and Montoya, 1976; Gu *et al.*, 1995). However, the phenotype of these mutants also implies collaborative involvement of HisC in the final step of the biosynthesis of some aromatic amino acids, leading the authors of these reports to propose that the dual function of HisC can be explained by a promiscuous activity of HisC overlapping with the activity of other aromatic amino acids aminotransferase. Interestingly, the physical position of the *hisC* gene in the chromosomes of *B. subtilis* (Nester and Montoya, 1976; Henner *et al.*, 1986; Berka *et al.*, 2003) and *Z. mobilis* (Gu *et al.*, 1995) is within a cluster of genes involved in the biosynthesis of aromatic amino acids and not as part of the histidine-controlled *his* operon. This genomic organisation has been explained in the basis of the double function of these enzymes (Jensen and Gu, 1996).

The case of PriA is different from that of HisC because there is no collaboration of enzyme functions through overlapping activities, but a single enzyme function 100% responsible for steps in both histidine and tryptophan biosynthesis. If the broad-substrate

specificity of a given enzyme is in fact an ancient feature (Jensen, 1976; O'Brien and Herschlag, 1999; see also Section 2.2.2.2) it is implicit that this feature has been retained in PriA through the course of evolution. Therefore, it is pertinent to address the question of why did *S. coelicolor* fail to develop independent *trpF* and *hisA* genes (see also Chapter IV). On the whole streptomycetes fail to regulate amino acid biosynthesis by feedback repression of gene expression (see Section 1.3.1.1 and references therein). This physiology has been proposed to be related to the oligotrophic status of soil (Hodgson, 2000). It seems that tryptophan biosynthesis in *S. coelicolor* is regulated by growth rate and/or growth phase (Hu, 1995; Hu *et al.*, 1999). Therefore, as amino acid biosynthesis is not co-ordinately regulated on a pathway specific basis, enzymes with roles in multiple pathways are possible. The central position of the *priA* gene within the *his* and *trp* genomic region of *S. coelicolor* (Figure 2.3) fits well with the bifunctionality of its product.

2.3 Conclusions

The functional genomics discrepancy of the lack of *trpF* in *S. coelicolor* was successfully addressed following the development of evolutionary driven hypotheses. It is hereby demonstrated that the genetic mapping of *trpF* in *S. coelicolor* by Smithers and Engel (1974) was wrong. Genetic evidence is offered that demonstrates that tryptophan and histidine biosynthesis in *S. coelicolor* converges to a common phosphoribosyl isomerase (PriA) in the analogous steps of the isomerisation of PRA (TrpF) and ProFAR (HisA), extending the common physiology for these two amino acids observed in other organisms. Broad-substrate specificity is proposed for PriA by which this enzyme participates in both biosynthetic pathways, an idiosyncrasy that is in accordance with the patchwork hypothesis.

Chapter III

Chapter III

"Structural and functional characterisation of PriA"

3.1 Introduction

3.1.1 General introduction

As discussed in Chapter I there is little doubt about the enormous importance of the $(\beta\alpha)_8$-barrels for both scientific and applied reasons (Section 1.2). Despite the large amount of information that has been generated over the last three decades, leading to a reasonable understanding of the structural logistic of the $(\beta\alpha)_8$-barrels, the evolutionary history of this protein family remains to be clarified (Copley and Bork, 2000; Henn-Sax *et al.*, 2001; Gerlt and Raushel, 2003). Arguments that favour either convergent evolution to a stable form (Lesk *et al.*, 1989; Hennig *et al.*, 1992) or divergent evolution from a common ancestor (Farber and Petsko, 1990; Reardon and Farber, 1995; Copley and Bork, 2000; Henn-Sax *et al.*, 2001) have been put forward. These arguments are discussed in the following Chapter in the light of the discovery of the enzyme PriA.

It has been previously acknowledged that only with the advent of more sequences and structures that may serve as *stepping-stones*, more evolutionary linkages between the members of the $(\beta\alpha)_8$-barrels can be established (Nagano *et al.*, 2002). The discovery of PriA, a putative ancient-like $(\beta\alpha)_8$-barrel enzyme as inferred from its dual-substrate specificity, may prove to be important for understanding the evolution of this structural fold. Thus, it is anticipated that the structure of PriA might act as a stepping-

stone. This Chapter presents experiments done in order to characterise PriA at the molecular level including its purification, enzyme activity determination and crystallisation. The experimental approach followed was complemented by an extensive computational analysis of the sequence of PriA. Such analyses were performed with the hope of identifying differences in the sequence and model structure of PriA, which might help to understand the unusual substrate specificity associated with this $(\beta\alpha)_8$-barrel. Furthermore, the results derived from these analyses might prove to be useful to direct future experiments.

3.1.2 Evolution of the $(\beta\alpha)_8$-barrels: the lack of consensus

The evolutionary history of the $(\beta\alpha)_8$-barrels remains a mystery that has been the subject of a vigorous debate. The *a priori* observation that all $(\beta\alpha)_8$-barrels are enzymes with a broad range of catalytic functions, involving substrates and cofactors of different structures, has the intuitive implication of convergent evolution (Farber and Petsko, 1990), i.e. parallel evolutionary trajectories leading to the same solution. This conclusion was supported by the fact that sequences of related $(\beta\alpha)_8$-barrel enzymes, as inferred from other observations such as function and conserved sequence motifs, share very little identity (Chothia, 1988; Farber and Petsko, 1990), i.e. below the standard threshold of 20% – 30% that is used to infer relatedness (Parkhill, 2002; Primrose and Twyman, 2003). This view was predominant until more $(\beta\alpha)_8$-barrel structures started to become available, showing the high degree of structural similarity that exist between enzymes with unrelated sequences and functions (e.g. Neidhart *et al.*, 1990). With the advent of more sophisticated algorithms and 3D superimposition methods used in the

field of structural biology, more evolutionary links between the structures and sequences of extant $(\beta\alpha)_8$-barrels have been established (Farber, 1993; Reardon and Farber, 1995; Nagano *et* al., 1999; Nagano *et al.*, 2002), expanding the initial relationships proposed by Farber and Petsko (1990).

The identification of sequence motifs and structural features common to the $(\beta\alpha)_8$-fold suggests divergent evolution, in other words, evolution from a common ancestor. For instance, the catalytic residues are always located at the C-terminal face of the β-barrel and loops that link the β strands with α helices (see Figure 1.4). Farber and Petsko (1990) highlighted this pattern, and so far no exception has been reported (Nagano *et al.*, 2002). Moreover, such functional residues more often occupy 3D structural positions that are equivalent regardless of their function (for a review see Nagano *et al.*, 2002). Another feature that is frequently conserved amongst members of a family is the number and location of extra structural elements, e.g. insertion of an α-helix 8 prime (α8') in the structures of the $(\beta\alpha)_8$-barrels of tryptophan biosynthesis (Eder and Kirschner, 1992), which is shared with the $(\beta\alpha)_8$-barrels of enzymes involved in histidine biosynthesis (Lang *et al.*, 2000). It is also common for the substrate-binding domain to be located towards the C-terminal end of the chain, e.g. the C-terminus location of the so-called *standard phosphate-binding* (SPB) motif (see Section 3.1.4; Wilmanns *et al.*, 1991; Bork *et al.,* 1995; Nagano *et al.*, 2002).

These structural and sequence data is further supported by functional analyses. For example, three possible strategies that $(\beta\alpha)_8$-barrel enzymes may have followed for divergent evolution of function have been proposed and experimentally demonstrated (see also Section 1.1.5.2): (i) substrate specificity is dominant, e.g. evolution of TrpF

from TrpC (Section 2.2.2.1 and Figure 2.2); (ii) chemical mechanism is dominant, e.g. evolution of TrpF from HisA (see Section 2.2.2.2 and Figure 2.2); and active site architecture is dominant, which refers to the opportunistic adoption of a structural framework and catalytic residues by the $(\beta\alpha)_8$-barrel evolutionary product that can be used in different mechanistic and metabolic contexts (Gerlt and Babbitt, 2001b; Wise *et al.*, 2002). Based on this wealth of data it can be concluded that a significant number of the extant $(\beta\alpha)_8$-barrels are related. However, although these observations clearly favour divergent evolution, the possibility that the $(\beta\alpha)_8$-fold appeared in evolution more than once cannot be ruled out (Gerlt and Raushel, 2003). It remains to be seen if enough evidence will be gathered to provide support for the evolutionary relatedness of the *entirety* of the family of the $(\beta\alpha)_8$-barrels.

Another approach that is used in the scientific method is to test a hypothesis by challenging it through falsification. Lesk *et al.* (1989) concluded that the $(\beta\alpha)_8$-barrels can be placed in two broad classes based on side-chain packing considerations. Through their eyes the $(\beta\alpha)_8$-barrels analysed had substantially different packing arrangements that could not be interchanged through simple changes in the genetic code that are likely to occur in nature. However, it has been proposed that such packing patterns in some $(\beta\alpha)_8$-barrels can be modified *via* single-residue mutations (Raine *et al.*, 1994). Moreover, it has been proved that genetic permutations leading to a change in the side-chain packing pattern are feasible (Luger *et al.*, 1989; Luger *et al.*, 1990; Urfer and Kirschner, 1992) and that they do occur in nature (Copley and Bork, 2000). Another piece of evidence that suggests convergent evolution is the discovery of a $(\beta\alpha)_8$-barrel

that presumably lacks any enzyme activity, the implication being that there are enzymatic and non-enzymatic $(\beta\alpha)_8$-barrels (Hennig *et al.*, 1992).

What is the contribution of the discovery of PriA to this debate? We argue that PriA did not appear as a consequence of loss of *trpF* and subsequent broadening of the specificity of HisA (Barona-Gómez and Hodgson, 2003). On the basis of the broad-substrate specificity of PriA, one hypothetical scenario is that TrpF and HisA could arise from PriA after gene duplication and specialisation, as suggested by the patchwork hypothesis (Jensen, 1976). This scenario is an example of divergent evolution (Henn-Sax *et al.*, 2001) and would support the view that the ancestor of PriA was one of the older of the $(\beta\alpha)_8$-barrels (Fani *et al.*, 1994; Copley and Bork, 2000; Lang *et al.*, 2000), although it is also possible that its ancestor arose relatively recently (Nagano *et al.*, 2002). An alternative scenario is that a TrpF function could arise from another enzyme such as TrpC (Wilmanns *et al.*, 1991; Eder and Kirschner, 1992; Gerlt and Babbitt, 2001b) and then pre-existing PriA could lose its TrpF activity to become HisA exclusively, an example of convergent evolution. Therefore, the possibility of either scenario suggest that divergent or convergent evolution may have accounted for the evolution of the extant $(\beta\alpha)_8$-barrel enzymes as a whole, and in particular for the $(\beta\alpha)_8$-barrels involved in tryptophan and histidine biosynthesis (Barona-Gómez and Hodgson, 2003). However, these scenarios are based upon the assumption that PriA is a $(\beta\alpha)_8$-barrel and therefore the structure of this enzyme is required.

3.1.3 The 'tortuous' folding pathway of the $(\beta\alpha)_8$-barrels

Studies on the folding pathway of $(\beta\alpha)_8$-barrel proteins aimed to understand how the soluble barrel fold have added another level of ambiguity to the evolutionary problem discussed in the previous Section (Thoma *et al.*, 1998). Although these studies were undertaken to learn more about the relationship between amino acid sequence and structure, these results might also provide circumstantial evidence as to how proteins evolved from smaller progenitors (for a review see Gunasekaran *et al.*, 2001). It has been speculated that conservation of the folding pathway of the $(\beta\alpha)_8$-barrels can be used to infer evolutionary relatedness (Eder and Kirschner; 1992; Forsyth and Matthews, 2002). Very early partially unfolded forms of the α subunit of tryptophan synthase (TS) were detected, suggesting that the single domain of the $(\beta\alpha)_8$-barrels could be composed by further secondary structural elements that can fold autonomously (Miles *et al.*, 1982). Over the years evidence has accumulated that suggest that different $(\beta\alpha)_8$-barrel enzymes fold following the so-called *6 + 2* mechanism in which the folding of the first six $(\beta\alpha)$ modules is followed by the folding and association of the last two $(\beta\alpha)$ modules. Indeed, this seems to be the case for the three $(\beta\alpha)_8$-barrels involved in tryptophan biosynthesis as revealed by unfolding and fragment complementation studies: the α subunit of TS or TrpA (Miles *et al.*, 1982; Zitzewitz *et al.*, 1999); IGP synthase or TrpC (Sánchez del Pino and Fersht, 1997; Forsyth and Matthews, 2002); and PRA isomerase or TrpF (Eder and Kirschner, 1992; Jasanoff *et al.*, 1994; Patrick and Blackburn, personal communication). This situation is also encountered in other well studied $(\beta\alpha)_8$-barrels such as TIM (Bertolaet and Knowles, 1995; Silverman *et al.*, 2001) and the α subunit of the bacterial luciferase (Noland *et al.*, 1999).

Recently Silverman and Harbury (2002) reported evidence for the existence of three cooperatively unfolding subdomains within the structure of TIM, as inferred from misincorporation proton-alkyl exchange data. Based on their results they suggested a *3 + 3 + 2* model. Along these lines, the results of Zitzewitz *et al.* (1999) suggest, "that individual βαβ modules have a strong preference for structure". As discussed in Section 1.2.1 the loops between β strands and α helices accommodate the catalytically important residues while the loops connecting α helices and β strands are more important for stability. Thus, this minimal βαβ module simplifies the complexity of the $(\beta\alpha)_8$-barrels. How do these pieces of information fit in the 6 + 2 model? At a first glance the two models may appear to be at odds. However, further inspection leads to an obvious geometrical solution, i.e. that the two $(\beta\alpha)_3$-subdomains correspond to the $(\beta\alpha)_6$ unfolded form previously identified (Silverman and Harbury, 2002).

It has been proposed from sequence (Fani *et al.*, 1994; Thoma *et al.*, 1998) and structural (Lang *et al.*, 2000; Höcker *et al.*, 2001b) studies that the homologous $(\beta\alpha)_8$-barrel enzymes ProFAR isomerase (HisA) and the cyclase moiety of imidazole-3-glycerol phosphate (ImGP) synthase (HisF) of histidine biosynthesis are composed of two duplicated superimposable $(\beta\alpha)_4$-barrel subdomains. Reconstitution of the enzyme HisF into its soluble and active form from its N and C-terminal halves expressed independently strongly supports this view (Höcker *et al.*, 2001b). These results led to the postulation of a *4 + 4* folding mechanism. However, in spite of the sound evidence offered, the dearth of other $(\beta\alpha)_8$-barrels showing this internal duplication raises some doubts about the universality of the model. Lang *et al.* (2000) have argued that internal duplications in other $(\beta\alpha)_8$-barrels are not detected because they have been erased

during the course of evolution at the sequence level as the $(\beta\alpha)_8$-fold is an ancestral domain. There is only one other example of a $(\beta\alpha)_8$-barrel where duplication of two half-barrels has been suggested, namely, the structure of one of the domains of the enzyme phosphoinositide-specific phospholipase C (Heinz *et al.*, 1998). However, reconstitution of the activity of this enzyme from its putative $(\beta\alpha)_4$-barrel subdomains has not been reported.

Whether these two folding mechanisms, i.e. the 3 + 3 + 2 and 4 + 4 models, are correct is virtually impossible to answer without a lot of speculation. Therefore, one can only rationalise the implications of the veracity of each model. This acquires relevance because if the $(\beta\alpha)_8$-barrels evolved from smaller subdomains, "phylogenetic analyses must include the consequences of such modular complexity" (Gerlt and Babbitt, 2001a). Furthermore, as it can be noted, the majority of the evidence comes from experiments using the $(\beta\alpha)_8$-barrel enzymes involved in tryptophan and histidine biosynthesis. The current view is that these proteins are homologous to each other (Wilmanns and Eisenberg, 1993; Bork *et al.*, 1995; Copley and Bork, 2000) and form a discrete evolutionary group (Nagano *et al.*, 2002; see also Section 3.1.4). Both models cannot be correct leaving us with an irreconcilable situation even for the small $(\beta\alpha)_8$-barrels subfamily of TrpF, TrpC, TrpA, HisA and HisF, as originally recognised by Thoma *et al.* (1998). It is hoped that structural and functional characterisation of PriA may help to solve this conundrum.

3.1.4 Structural features of the FMOP superfamily

The general structural features of the $(\beta\alpha)_8$-fold were introduced in Section 1.2.1 and shown in Figure 1.4. Such features are expanded in this Section with emphasis on the $(\beta\alpha)_8$-barrel superfamily of the flavin mononucleotide-dependent oxidoreductase and phosphate-binding enzymes (abbreviated FMOP; Nagano *et al.*, 2002). The membership of the FMOP superfamily includes a total of twelve sequence families representing seventeen non-redundant $(\beta\alpha)_8$-barrel structures (CATH version 1.7; Nagano *et al.*, 2002). The sequence families corresponding to the $(\beta\alpha)_8$-barrel enzymes involved in tryptophan and histidine biosynthesis are members of this superfamily.

The main characteristic of the FMOP superfamily is the presence of the *standard phosphate-binding* (SPB) motif located towards the C-terminal end of the sequence. The SPB motif was first identified by Wilmanns *et al.* (1991) as a glycine-rich motif laying between the C-terminal end of β strand 7 and α helix 8' – including loop 7, α helix 7 and β strand 8 – in the sequences of at least seven $(\beta\alpha)_8$-barrel enzymes including those involved in tryptophan biosynthesis (see Section 3.1.4.1). Bork *et al.* (1995) expanded the SPB motif family to include other $(\beta\alpha)_8$-barrels such as those involved in histidine biosynthesis (see Section 3.1.4.2). The binding of two phosphate ions, as shown in Figure 3.1, evidences the two SPB motifs in the 3D structure of the cyclase subunit of ImGP synthase or HisF (Lang *et al.*, 2000).

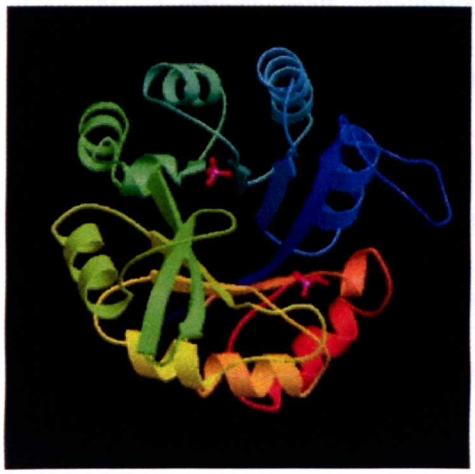


Figure 3.1 Three-dimensional structure of HisF represented by ribbons. The structure of HisF from *T. maritima* was resolved to a resolution of 1.45 Å (PDB 1thf). The binding of two phosphate-ions [violet (phosphorous atoms) and red (oxygen atoms)] reveals the location of the two SPB sites in HisF. It has been proposed by analogy to HisF that the phosphate moieties of ProFAR bind to HisA in a similar fashion (Lang *et al.*, 2000).

3.1.4.1 Structural features of the TrpF sequence family

There are two structures available for the enzyme PRA isomerase: the C-terminal domain of TrpC from *E. coli* (eTrpF; PDB code 1pii; Priestle *et al.*, 1987; Wilmanns *et al.*, 1992) and the monofunctional TrpF from the hyperthermophile *T. maritima* (tTrpF; PDB code 1nsj; Henn-Sax *et al.*, 2002). Remarkably, inspection of the structures of these TrpF orthologs reveals some differences. The must obvious one is the absence of the α-helix 5 from the structure of eTrpF, which is replaced by a peptide segment without a defined secondary structure (Priestle *et al.*, 1987; Wilmanns *et al.*, 1992). In contrast, if the α helix 8' of the SPB motif is disregarded, the tTrpF structure is a perfect $(\beta\alpha)_8$-barrel (Henn-Sax *et al.*, 2002). There are some other structural differences between eTrpF and tTrpF, such as the length of the loops connecting α helices with β strands, although these are likely to reflect the constraints of the hyperthermophilic tTrpF in comparison to the mesophilic eTrpF (Henning *et al.*, 1997).

The catalytic residues in tTrpF have been identified and characterised through directed mutagenesis (Jürgens *et al.*, 2000) and elucidation of the structure of a complex of tTrpF with an analogue of its product (Henn-Sax *et al.*, 2002). These are Asp126

acting as the general acid, which protonates the furanose ring oxygen, and Cys7 acting as the general base, which abstracts a proton from the C2′ atom of the ribose in the irreversible Amadori rearrangement (Jürgens *et al.*, 2000; Henn-Sax *et al.*, 2002). The residues Cys7 and Asp126 are located towards the C-terminal end of their cognate β strands, i.e. β1 and β6 strands, respectively. Additionally, residue Arg36 of tTrpF was shown to form a salt bridge between its side chain and the carboxylate group of anthranilate (Henn-Sax *et al.*, 2002).

3.1.4.2 Structural features of the HisA sequence family

Based on structure prediction (Wilmanns and Eisenberg, 1993) and sequence alignments of the regions containing the phosphate-binding sites (Bork *et al.*, 1995) the ProFAR isomerase or HisA sequence family was proposed to be related to the sequence families of the $(\beta\alpha)_8$-barrels taking part in tryptophan biosynthesis, implying that these enzymes may share the $(\beta\alpha)_8$-barrel fold. This prediction was confirmed by elucidation of the structure of HisA (tHisA; PDB code 1qo2) and HisF (tHisF; PDB code 1thf; see also Figure 3.1) from *T. maritima*, which were shown to adopt the $(\beta\alpha)_8$-barrel fold (Lang *et al.*, 2000). The main feature of the HisA and HisF sequence families that places them within the FMOP superfamily is the presence of the SPB motif located at the C-terminal end of the chain (Nagano *et al.*, 2002). In addition, some of the catalytic residues of HisA and HisF are in equivalent positions to the catalytic residues of other members of the superfamily such as TrpF (Henn-Sax *et al.*, 2002). For tHisA the essential residues are Asp8 in β strand 1 and Asp127 in β strand 5, presumably acting as the general base and general acid, respectively, as inferred by analogy to TrpF (Jurgens

et al., 2000; Henn-Sax *et al.*, 2002). Additionally, residue Thr164 of tHisA was shown to be important for catalysis although not essential (Henn-Sax *et al.*, 2002).

The HisA and HisF sequence families are different to the rest of the members of the FMOP superfamily in that they show the internal duplication discussed in Section 3.1.3. Moreover, although this duplication implies that the structures of HisA and HisF are symmetrical (probably as a consequence of the symmetry of their substrates) extensive structural elements have been inserted particularly towards the C-terminal end of the β strands and subsequent loops. The inserted structural elements for HisA are β strands 1' and 1'', and α helices 2', 4' 6' and 8', the latter being the α-helix involved in the SPB motif (Lang *et al.*, 2000).

3.2 Results and Discussion

3.2.1 Computational analysis of the sequence of PriA

One implication of the bifunctionality of PriA is that TrpF and HisA share a common ancestor that might resemble PriA. Therefore, a comparative analysis of the sequence of PriA with respect to the sequences of HisA and TrpF is pertinent. Moreover, such analysis might bring some clues about the molecular basis of the dual function of PriA. The question addressed in the following Sections is whether or not differences can be detected in the sequence of PriA that might help us to understand its substrate specificity.

3.2.1.1 Considerations in the quality and handling of the data

The sequences of PriA and tHisA share about 25% identity at the amino acid level whereas the sequences of PriA and tTrpF share less than 10% identity. Furthermore, the catalytic important residues Asp8, Asp127 and Thr164 of tHisA are conserved in PriA of *S. coelicolor* (Asp11, Asp130 and Thr166) in addition to the N and C-terminus SPB motifs, as well as other conserved residues of unknown function (see Figure 3.4). Thus, it is clear that PriA resembles more closely HisA than TrpF. Based on this observation a multisequence alignment of sixty six selected bacterial homologs of PriA, including seven from actinomycetes, was constructed (Figure 3.2). The sequences used were identified by BLAST (Altschul *et al.*, 1997) and HMM (Karplus *et al.*, 1998) searches in public databases up until January 2003. The alignment was constructed automatically with the program ClustalX 1.81 and corrected by eye using data from the structure of tHisA as *anchor residues* (Lang *et al.*, 2000). This approach enhances the quality of the alignment and allows analysis to be done by sections of residues associated with structural elements (Graur and Li, 2000). Furthermore, the alignment shown in Figure 3.2 was analysed in the context of the results reported by Jürgens *et al.* (2000) and Henn-Sax *et al.* (2002). Of particular relevance is the structure-based amino acid sequence alignment of tHisA and tTrpF done by these workers after superimposition of these structures with a root mean square (rms) deviation of 2.2Å. The other body of information considered from the work of Jürgens *et al.* (2000) were the replaced residues that led to the establishment of TrpF catalytic activity on the scaffold of tHisA.

3.2.1.2 General analysis of the sequence alignment

The first observation that can be made of the sequence alignment is that the actinomycetes sequences are highly conserved forming a discrete group. Based on this observation, albeit only for nomenclature purposes until their biological role is confirmed, all the sequences from actinomycetes are called PriA. The HisA protein from *Deinococcus radiodurans* appears to be similar to the conserved actinomycetes PriA proteins, although it's clustering with the PriA sequences is not unambiguous (see also Section 4.2.4). The putative conservation of PriA across and beyond the actinomycetes is also discussed in Chapter IV. A phylogenetic tree derived from this multisequence alignment is shown in Figure 4.5 and a discussion is presented in Section 4.2.4.

Figure 3.2 Multisequence alignment of PriA homolgues *(see the alignment in the following pages).* The homologues are named after the prefix HisA (HisA_seq) other than the actinomycetes homolgues for which PriA is used (PriA_seq, **red**). The sequence of *T. maritima* (HisA_Tm, **blue**) was used to assign the structural elements. The β strands are identified with **light blue** arrows and the α helices with pink cylinders. The N-terminal end insertions in HisA_Mht, HisA_Na and PriA_Tf were deleted. The invariant residues are shown in HisA_Tm and PriA_seq in bold font. The residues relevant to the novel function of PriA (discussed in the text) are highlighted in yellow. The suffix identifies the origin of the sequence (right-hand side column). The abbreviations used in the order in which the sequences appear in the alignment, with the accession number between parentheses, are as follows: Sf *Shigella flexneri* (24113409), Ec *Escherichia coli* (16129965), Se *Salmonella enterica* (16761007), Yp *Yersinia pestis* (16121817), Bch *Buchnera* sp. (15616725), Pm *Pasteurella multocida* (15603068), Hi *Haemophilus influenza* (16272421), Vc *Vibrio cholerae* (15641150), So *Shewanella oneidensis* (24373629), Cj *Campylobacter jejuni* (15792906), Xc *Xanthomonas campestris* (21231261), Xf *Xylella fastidiosa* (15838806), Ml *Mycobacterium leprae* (15827647), Mt *Mycobacterium tuberculosis* (15608741), Tf *Thermobifida fusca* (23018086), Sc *Streptomyces coelicolor* (21220531), Bl *Bifidobacterium longum* (23465861), Ce *Corynebacterium efficiens* (25028552), Cg *Corynebacterium glutamicum* (19553293), Dr *Deinococcus radiodurans* (15807480), Pya *Pyrobaculum aerophilum* (18312333), Ss *Sulfolobus solfataricus* (15897511), Tm *Thermotoga maritima* (15643795), Pf *Pyrococcus furiosus* (18978034), Aa *Aquifex aeolicus* (15606512), Mht *Methanothermobacter thermoautotrophicus* (15678863), Mb *Methanosarcina barkeri* (23051200), Mj *Methanococcus jannaschii* (15669727), Af *Archaeoglobus fulgidus* (11498591), Hal *Halobacterium* sp. NRC-1 (15791103), Nm *Neisseria meningitidis* (15793809), Ne *Nitrosomonas europaea* (22954624), Bf *Burkholderia fungorum* (22985724), Rs *Ralstonia solanacearum* (17547666), Pa *Pseudomonas aeruginosa* (15600334), Av *Azotobacter vinelandii* (23103281), Md *Microbulbifer degradans* (23027167), Gm *Geobacter metalireducens* (23055863), Mag *Magnetococcus* sp. (22998453), At *Agrobacterium tumefaciens* (15887398), Srm *Sinorhizobium melitoti* (15963805), Brm *Brucella melitensis* (17988325), Rr *Rhodospirillum rubrum* (22965732), Bj *Bradyrhizobium japonicum* (27375764), Rpp *Rhodopseudomonas palustris* (22962307), Na *Novosphingobium aromaticivorans* (23109767), Cc *Caulobacter crescentus* (16127966), The *Thermoanaerobacter ethanolicus* (20455020), Ct *Clostridium thermocellum* (23020793), Dh *Desulfitobacterium hafniense* (23112135), Tre *Trichodesmium erythraeum* (23043144), Nos *Nostoc* sp. (17231998), Syn *Synechococcus* sp. (24251254), Te *Thermosynechococcus elongatus* (22297760), Prm *Prochlorococcus marinus* (23131183), Sm *Streptococcus mutans* (24379686), Lm *Listeria monocytogenes* (16802607), Lnm *Leuconostoc mesenteroides* (23023419), Ll *Lactococcus lactis* (15673195), Li *Leptospira interrogans* (24212826), Oi *Oceanobacillus iheyensis* (23098003), Dd *Desulfovibrio desulfuricans* (23475575), Sa *Staphylococcus aureus* (21284322), Bs *Bacillus subtilis* (16080541), Bh *Bacillus halodurans* (15616141), Cht *Chlorobium tepidum* (21673312).


```
  1         11        21        31        41        51        61        71        81        91       100
  |         |         |         |         |         |         |         |         |         |         |
1 ---MMIIPALDLIDGTVVRLHQGDYGKQRDYGNDPLP-RLQDYAAQGAEVLHLVDLTGAKDPAKRQIPLIKTLVAGVN---VPVQVGGGVRSEEDVAALL   93    HisA_Sf
1 ---MMIIPALDLIDGTVVRLHQGDYGKQRDYGNDPLP-RLQDYAAQGAEVLHLVDLTGAKDPAKRQIPLIKTLVAGVN---VPVQVGGGVRSEEDVAALL   93    HisA_Ec
1 ----MIIPALDLIDGTVVRLHQGDYARQRDYGNDPLP-RLQDYAAQGAGVLHLVDLTGAKDPAKRQIPLIKTLVAGVN---VPVQVGGGVRTEEDVAALL   92    HisA_Se
1 ----MIIPALDLIEGKVVRLHQGDYGQQRDYGNHPLP-RLQDYQQQGAQVLHLVDLTGAKDPAARQIPLLRELLAGVD---VPVQVGGGIRNEQDVVALL   92    HisA_Yp
1 ----MIIPAFDLINGRTVRLYQGDYSNQKNYNVNLFN-SLEVYKSKGIEIVHLVDLDGAKNSANRQIELFKKIVSHTT---VPVQVGGGIRTTKDISTLL   92    HisA_Bch
1 MKTSQIIPALDLIDGQVVRLYQGDYGQKTLYSDNPIA-QFQDYVAQGAKYLHLVDLTGAKDPTKRQTQLIGEIINAVN---CSIQVGGGIRTEQDVADLL   96    HisA_Pm
1 MKQSIIIPALDLINGQVVRLHQGDYAKQTTYSDNPIK-QFDNYVRQGAKQLHLVDLTGAKNPQSRQTALIGKIVEATQ---CKVQVGGGIRTEQDVADLL   96    HisA_Hi
1 ----MIIPALDLIEGQVVRLYQGDYGQVTEYKVDPAE-QFNLYHQAGAGWLHLVDLTGAKNTAARQLNLIARLLASTP---ANIQIGGGVRSEQDVVDLL   92    HisA_Vc
1 ----MIIPAIDLIDGKVVRLYQGDYGQQTTFDLSPLA-QLQSYQAQDASWLHIVDLTGAKDPAKRQTALIAALTAGLS---ANIQVGGGIRTEEQVAELL   92    HisA_So
1 --MTQIIPALDLIDGEVVRLVKGDYEQKKVYKYNPLK-KFKEYEKAGAKELHLVDLTGAKDPSKRQFALIEKLAKEVS---VNLQVGGGIRSKEEVKALL   94    HisA_Cj
1 -MSFTVYPALDIRDGRVVRLLQGDYARETRYGDDVLP-RAQAFADAGAQWMHLVDLDAAKAGGYTLAALLGQISRQTG---LQVQTGGGVRSREDVARIL   95    HisA_Xc
1 -MNFIVYPALDIRNGAVVRLQQGDYARQTRYDDQVLP-RALAFADSGATWMHLVDLDAAKAGGYTLAPLLRQMTRATG---LQVQTGGGVRSRDDVARIL   95    HisA_Xf
1 ---LILLPAVDVVEGRAVRLVQGKAGSENDYG-SALD-AALCWQRDGADWIHLVDLDAAFGRGSN-RELLSEMVGKLD---VQVELSGGIRDDDSLNAAL   91    PriA_Ml
1 ---LILLPAVDVVEGRAVRLVQGKAGSQTEYG-SAVD-AALGWQRDGAEWIHLVDLDAAFGRGSN-HELLAEVVGKLD---VQVELSGGIRDDESLAAAL   91    PriA_Mt
1 P-ILELLPAVDVAGGQAVQLVQGKDGTGGRYG-DPLN-AALTWQNAGAEWIHLVDLDAAFGRGHN-RELLASIVGKLD---IKVELSGGIRDDESLNAAL   93    PriA_Tf
1 --KLELLPAVDVRDGQAVRLVHGESGTETSYG-SPLE-AALAWQRSGAEWLHLVDLDAAFGTGDN-RALIAEVAQAMD---IKVELSGGIRDDDTLAAAL   92    PriA_Sc
1 ---LTLLPAVDVRDGKAVRLRQGESGSETDYG-SPFE-AARTWVEAGAEWIHLVDLDAAFGTGNN-RDQLREIVHELGD-RVNIELSGGVRDDASLDAAL   93    PriA_Bl
1 ---FTILPAVDVVNGQAVRLDQGEAGTEKSYG-TPLE-SALRWQEQGAEWLHFVDLDAAFNRGSN-HELMAEITRQLD---IKVELTGGIRDDASLERAL   91    PriA_Ce
1 ---FTILPAVDVVNGQAVRLDQGEAGTEKSYG-TPLE-SALKWQEQGAKWLHFVDLDAAFNRGSN-HEMMAEIVGKLD---VDVELTGGIRDDESLERAL   91    PriA_Cg
1 LPLPQVIPCVDIQSGRAVRLYEGDPDRETVYFDSPRA-AARHWVTLGAGFLHLVDLDAATGRGEN-RAVIAEIVRELG---VPVEVGGGVRDRAAAQSLL   95    HisA_Dr
1 ----MIIPSIDIEGGRAVKRIRGQRG-NYIFQGDPLE-LAARFRKA--PLVHVVDLDGAEAGGLVNISVISKIAEILEG---RCELGGGLRSLEAVERAL   89    HisA_Pya
1 --MNEIIPSIDISFGKAVKRIRGVRG-TGLVLGNPIE-LANKLYNEGYSRIHVVDLDAAEGVG-NNEMYIKEICKIGFD---WIQVGGGIRDVEKAKRLI   92    HisA_Ss
1 ---MLVVPAIDLFRGKVARMIKGRKENTIFYEKDPVE-LVEKLIEEGFTLIHVVDLSNAIENSGENLPVLEKLSEFAEH----IQIGGGIRSLDYAEKLR   92    HisA_Tm
1 -MKFRIYPAIDIMDGKVVRLYKGKKDEVKVYGD-PLK-FAEKFSEY-VDKIHIVDLDGAFSGKPQNLDIVEKIIRDVG---VRVQIGGGFRDYESIKKAY   93    HisA_Pf
1 NLKEFIIPAIDLMEGKAVRLYKGDFNKVKVYSERPWE-LAKNFSDLGFKRLHVVDLEGAEGGKPKNLEVIRKIRENFE---GEVEVGGGIRSYEVAKALF   96    HisA_Aa
1 KDRMLIIPAVDIKDGKCVQLVQGKPGTEQVVLDDPAG-VAGRWESLGAETIHVVDLDGAL-GLEKNTGILKEITERVS---VPLQIGGGIRSREYAGKLL   95    HisA_Mht
1 ---FEVIPAVDMRGGKCVQLVQGVPGSEIVSIDDPVE-VALDWVRKGAKTLHLVDLDGAIEGERKNAPIIEKIVSACQKKGVQIQVGGGIRSFEDAASLL   96    HisA_Mb
1 ---MIIIPAVDLKDKKCVQLIQGDPNKKHLELNNPVE-VAKKFVDEGAEYLHIIDLDAAF-GTGNNRDVIKNIIKEVN---VPVEVGGGIRNLEIAKELI   92    HisA_Mj
1 --MFRVIPAIDLKDGKVVRLRQGKEDEVTFEATNPIE-VAKQWVEIGAKVLHVIDLSGAFQGRLRHEEIIAEIAKMAE-----VQVGGGIRDFRVAERLF   92    HisA_Af
1 S--FTVLPAVDMQDGQVVQLVQGERGTERTYGDPVD--AATDWVAAGAEALHLVDLDGAFEGARANATAVEDILDATD---VSVQVGGGIRSAEDATALL   93    HisA_Hal
1 ---MLLIPAIDLKEGRCVRLKQGLMEEATVFSDSPAE-TALHWFEQGARRLHLVDLNGAFAGVPQNLPAIKDILAAVAK-DIPVQLGGGIRDLKTIGQYL   95    HisA_Nm
1 ---MLIIPAIDLKDGHCVRLKQGIMENATVFSENPET-VALHWLDNGARQLHLVDLNGAFAGKPKNGEAIRAIVEAVDG-RIPIQLGGGIRDLETIEYYL   95    HisA_Ne
1 ---MLLIPAIDLKDGQCVRLKQGDMDQATIFSEEPAA-MARHWVDRGARRLHLVDLNGAFAGKPKNEDAIRAIIAEVGG-EIPVQLGGGIRDLNTIERYL   95    HisA_Bf
1 ---MLLIPAIDLKDGQCVRLKQGDMDQATVFSEDPAA-MARHWVDQGARRLHLVDLNGAFVGKPRNEAAIKAIIAEVGS-EIPVQLGGGIRDLNTIERWL   95    HisA_Rs
```


```
  1         11        21        31        41        51        61        71        81        91       100
  |         |         |         |         |         |         |         |         |         |         |
1 ---MLIIPAIDLKDGACVRLRQGLMEDATVFSDDPVA-MAAKWVDGGCRRLHLVDLNGAFEGKPVNGEVVTAIARRYP--DLPIQIGGGIRSLETIEHYV   94    HisA_Pa
1 ---MLIIPAIDLKDGACVRLRQGRMDDSTVFSDDPVA-MAARWVEAGCRRLHLVDLNGAFEGQPINGEVVTAIARRYP--DLPIQIGGGIRSLETIEHYV   94    HisA_Av
1 ---MIVIPAIDLKDGACVRLRQGLMEDSTVFSDDPAA-MARKWVEQGAKRLHLVDLNGAFAGEPVNGEVVKAIAKAYP--NLPIQIGGGIRSLETIKHYL   94    HisA_Md
1 ---MIVIPAIDLKEGKCVRLEQGLMEKDTVFCDNPAD-QARAWERQGAELLHIVDLDGAFAGEPKNRASIEAIVKAIA---IPTQLGGGIRDIPTIEAYL   93    HisA_Gm
1 ---MRIIPAIDLKDGQCVRLFKGDMDQNTVYSDNPGE-TAKQWAEQGAERMHVVDLNGAFAGEPVNADAIAAIRKAIT---IPMQLGGGIRTLETLQKLF   93    HisA_Mag
1 ---MILFPAIDLKDGECVRLKLGDMEQATVYNEDPGA-QAKAFEDQGFEWLHVVDLNGAFAGETVNGAAVDAILKSTK---NPVQLGGGIRTLDHIEAWL   93    HisA_At
1 ---MILFPAIDLKDGQCVRLKLGDMEQATVYNPDPAA-QARAFEEQGFEWLHVVDLNGAFAGETVNGAAVDAILKATK---NPVQLGGGIRTLEHIENWL   93    HisA_Srm
1 ---MILFPAIDLKDGQCVRLKLGDMDQATIYNEDPAA-QAKAFEDQGFEWLHVVDLNGAFAGESVNGTAVEAILKATK---NPVQLGGGIRTLAHIENWL   93    HisA_Brm
1 ---MILFPAIDLKDGQCVRLRKGEMDQATVFNDSPAD-QAAAFALAGCRWIHVVDLNGAFAGRPVNAPAVEAILAAVD---LPVQLGGGIRDLAAIDLWL   93    HisA_Rr
1 MEAVILFPAIDLKNGQCVRLEQGDMARATVFNLNPAA-QAQSFAEQGFEYLHVVDLDGAFAGKPVNAQAVEAMLKTIK---IPVQLGGGIRDLATVEAWL   96    HisA_Bj
1 MEAVILFPAIDLKNGQCVRLEQGDMARATVFNLDPAA-QAQSFAAQGFQYLHVVDLDGAFAGKPMNAQAVEAMLKVVS---MPVQLGGGIRDLKTIEAWL   96    HisA_Rpp
1 ARAMIVFPAIDLKAGQVVRLAEGDMDRATVYGDNPAH-QASLFAQAGSQYLHVVDLDGSFAGRAENREAVEGILKSFP---GHVQLGGGIRTREAVSGWF   96    HisA_Na
1 ---MILYPAIDLKDGQCVRLLHGDMDKATVFNTSPAD-QAQRFVQDGFSWLHVVDLNGAIEGKSVNTAAVEQILESIS---IPVQLGGGIRTLEGVEAWI   93    HisA_Cc
1 ---MIIYPAVDIKDGRCVRLVQGEFDKVTVYSDNPVE-MGLKWERMGAQYLHVVDLDGARTGQIQNTPIISEMAVKLG---IPVQLGGGIRTVETIETLL   93    HisA_The
1 ---MIIYPAVDIKDGRCVRLVQGEFDKVTVYSDNPVE-MGLKWERMGAQYLHVVDLDGARTGQIQNTPIISEMAVKLG---IPVQLGGGIRTVETIETLL   93    HisA_Ct
1 -MRMRLFPAIDLKEGKAVRLLQGRMEDATVYGEQPVE-VARKFKEQGADSLHVVDLDGAFAGKPVNDAVILKLIQSSG---LRVQVGGGIRTLERIEELL   95    HisA_Dh
1 ---MNIIPAIDILGGRCVRLYQGDYKQTQIFNDNPVD-VAKRWADEGATKLHLVDLDAAKVGQPVNQKTIEAIIKAVD---IPIQVGGGLRNYSTVASLL   93    HisA_Tre
1 ---MDVIPAIDLLAGRCVRLYQGDYDRSQVFSENPVD-VAKQWVDEGATRLHIVDLDGAKAGKVVNLGAIEAIAQAIS---IPIEIGGGLRDRSSVEQVF   93    HisA_Nos
1 ---MDVIPAIDLLGGQCVRLFQGDYDQAEVYGKDPVG-MALRWAEAGAQRLHLVDLDGAKEGSPVNAEAIATIAQRLS---IPVQVGGGLRDRDTVARLL   93    HisA_Syn
1 ---MDVIPAIDLLDGKCVRLVQGNYDKVNVFHDDPLK-VALYWQRLGAPRLHLVDLDAAKTGEPRNYDLIGKIVASLE---IPVQVGGGLRSRQAVADLF   93    HisA_Te
1 ---MEIIPAIDLLDSVCVRLHQGDYEKVTRFSEDPVD-QALSWQKQGATRLHLVDLDGAKSGEPVNDSCVRAITSALN---IPVQLGGGVRTLERAEELL   93    HisA_Prm
1 ---MQILPAIDIKNGHAVRLVKGDFEQETVVNDNVLD-QASIFMEAGIQYIHVVDLDGALEGRAANRDLIAKIKEMTG---LAIEVGGGIRTIEQIEDYL   93    HisA_Sm
1 ---MQIFPAIDLKNGQCVRLFQGDFSKKTVVNEDPIA-QAKAFATDGATYLHIVDLDGALEGRPINLEVIQKMKITAK---IPVQVGGGIRSMAQVDYYL   93    HisA_Lm
1 ----MIFPAIDLLNGQSVRLYQGDYEKETTINPDPLK-QAKQIESAGLKHLHLVDLDGAKEGKPVNLNVIQSLREQTN---LFIELGGGIRTLEQVNQYL   92    HisA_Lnm
1 ---MKIIPAIDLQNGEAVRLYKGDYDKKTVYSKTPLE-IAQKFERMGATDLHLVDLDDANIGQTRNLEIVRKIKEETR---LKIEIGGGIRDFDTVRMYL   93    HisA_Ll
1 ---MIVIPAIDLFDNCAVRLFKGNYEEKKIYSSEPWK-LAESFAKNGATLLHLVDLNGARNQLGVNEDSILKIRETTS---LKVQLGGGIRDKEKL-AYY   92    HisA_Li
1 ---MIIFPAIDIRNGKCVRLRQGDYNQETIYSNSPVE-MAKEWEASGAEYLHTVDLDGAKSGESNNINIIQDVAQALS---IPIQVGGGIRSLEVIDKYI   93    HisA_Oi
1 AWSVIIFPAVDIKGGQAVRLRQGKADQETVFSGDPVA-MARHWEEQGGRWLHVIDLDGAFSGTPVNRELIRTICASVS---IPVQLGGGIRDTVTAQAYL   96    HisA_Dd
1 --MIELWPAIDLIGSTSVRLTEGKYDSEEKMSRSAEESIAYYSQFECVNRIHIVDLIGAKAQYAREFDYIKSLRRLTT---KDIEVGGGIRTKSQIMDYF   95    HisA_Sa
1 MSAFTLYPAIDMRNGKCVRLVQGDYNKETIYGDSPYD-MAELFEKEGAEWIHLVDLDGAKEGKRVNDRHVIEIAQKLN---LKVEIGGGIRSENDVYEYL   96    HisA_Bs
1 MAKFIIYPAIDMRDGKCVRLVQGDYNQETVYGDSPFT-MAKSFAEQGAEWIHMVDLDGAKAKERMNHEHVVKVATELD---VNVQVGGGIRTAEDIAYYL   96    HisA_Bh
1 ---MLIIPAIDIKEGKCVRLTRGDFAQKKIYLDNPCD-MAVIWRKQNAKMIHVVDLDAALTGETVNFERIREIVNVLD---IPIQVGGGIRSVEAVEKYL   93    HisA_Cht
```


```
    101       111       121       131       141       151       161       171       181       191       200
    |         |         |         |         |         |         |         |         |         |         |
94  EAGVARVVVGST-AVKSPDVVKGWFERFGADALVLALDVRIDEQGNKQVAVSGWQENSGVSLEQLVETYLPVGLKHVLCTDISRDGTLAGSNVSLYEEVC  192   HisA_Sf
94  EAGVARVVVGST-AVKSQDMVKGWFERFGADALVLALDVRIDEQGNKQVAVSGWQENSGVSLEQLVETYLPVGLKHVLCTDISRDGTLAGSNVSLYEEVC  192   HisA_Ec
93  KAGVARVVIGST-AVKSPDVVKGWFERFGAQALVLALDVRIDEHGNKQVAVSGWQENSGVSLEQLVETYLPVGLKHVLCTDISRDGTLAGSNVSLYEEIC  191   HisA_Se
93  EAGAARVVVGST-AVKQPEMVQQWFERYGAEAIVLALDVRINEAGCKHVAISGWQENSDATLEQIVEQYLPYGLKHVLCTDISRDGTLSGSNVELYQEVC  191   HisA_Yp
93  DLGVKRVVIGSS-IVSNKKQVKQWLNFYGPDAIVLALDVHVDGSNKKEISIDGWQKKTNFILEEIIEYFLSSGLKHVLCTDISRDGTLLGPNFKLYKEIC  191   HisA_Bch
97  AVGANRVVIGST-AVKQPEMVKQWFKKYGAEKFVLALDVNINAQGEKQIAVSGWQENSGVSLEALIADFREVGLQHVLCTDISKDGTLQGSNVALYQEIC  195   HisA_Pm
97  AVGANRVVIGST-AVTHRSMVKNWFIKYGAEKFVLALDVNINASGQKIVAISGWQEESGVLLETLIEDFQTVGLQQVLCTDISRDGTLTGSNIGLYQEIC  195   HisA_Hi
93  EAGAQRVVVGST-AVKQPQLVKSWIEKYGAEKIVLALDINIDEQGTRKVAISGWQEDSGVTIEALIDDYLTVGLQHVLCTDISRDGTLAGSNVELYRDLC  191   HisA_Vc
93  SLGVKRVVIGSL-AVKEPELVKGWFNKFGNEAICLALDINITPSGEKIVAVSGWQNGGGKNLESIVEDFSQVGLKHALVTDISRDGTLTGANTELYCELS  191   HisA_So
95  DCGVKRVVIGSM-AIKDATLCLEILKEFGSEAIVLALDTILKED--YVVAVNAWQEASDKKLMEVLDFYSNKGLKHILCTDISKDGTMQGVNVRLYKLIH  191   HisA_Cj
96  DAGAARVVVGSL-AVRDSATVIGWLQEFGTDRLTIALDTRQDAAGVWQLPVHGWTETAEATLDQLATQYAQAGLQHLLCTDIARDGMLSGPNMGLYAHLR  194   HisA_Xc
96  DAGAARVVIGSL-AVRESARVIEWLQAFGPERITVALDTRQDAGGVWRLPVHGWTEVAEATLDVLAQQYAAAGLRHLLCTDIARDGMLSGPNMDVYTYLR  194   HisA_Xf
92  A-TGCARVNLGTAACENPHWCAQVIAEHG-DKIAVGLDVQIVDG-QHRLRGRGWET-DGGDLWDVLENLDRQGCSRFIVTDVTKDGTLDGPNLDLLASVS  187   PriA_Ml
92  A-TGCARVNVGTAALENPQWCARVIGEHG-DQVAVGLDVQIIDG-EHRLRGRGWET-DGGDLWDVLERLDSEGCSRFVVTDITKDGTLGGPNLDLLAGVA  187   PriA_Mt
94  A-TGCHRVNIGTAALENPDWCAKVIAEHG-DRVAIGLDVRGTT-----LAARGWTR-DGGDLYETLERLEAAGCARYIVTDVNKDGTLKGPNLDLLRDVC  185   PriA_Tf
93  A-TGCTRVNLGTAALETPEWVAKVIAEHG-DKIAVGLDVRGTT-----LRGRGWTR-DGGDLYETLDRLNKEGCARYVVTDIAKDGTLQGPNLELLKNVC  184   PriA_Sc
94  E-AGAARVNIGTAALENPDWTASVIKKYG-DRVAVGLDVRGHT-----LAARGWTR-EGGDLFETMKFLDSVGCSRYVVTDVAKDGMMSGPNIQLLSEVA  185   PriA_Bl
92  A-TGATRVNIGTAALEKPEWIADVIRRHG-EKIAVDIAVRLENG-EWRTKGNGWVS-DGGDLWEVLERLDSQGCSRFVVTDVSKDGTLTGPNVDLLRDVA  187   PriA_Ce
92  A-TGARRVNIGTAALEKPEWIASAIQRYG-EKIAVDIAVRLEDG-EWRTRGNGWVS-DGGDLWEVLERLDSQGCARFVVTDVSKDGTLSGPNVELLREVA  187   PriA_Cg
96  T-AGVARVVIGTAAVRNPELVAELIAEYGPERVVVSLDARGLE-----VATHGWAAGSGVSVADLTPRLAEAGLRTLIFTDVTRDGTLRGLDRDLMRQVR  189   HisA_Dr
90  S--LCQYAVVGSLPFKNWPLFR-RAAELYRDRLAVSLDYRGGE-----VLVGGWREKAASIAEAVRRLEEAGPYAAVIVTSVEVEGTGGGVKADIDVKSL  181   HisA_Pya
93  S-LDVNALIFSTIVFTNFNLFYNIVREIGNNRVIVSIDYDDTKR----VLISGWKERSMEIIDGIKKVNEL-ELLGIILTYVTNEGTVKGI--DYSVKDY  184   HisA_Ss
93  K-LGYRRQIVSSKVLEDPSFLKSLREIDV--EPVFSLDTRGGR-----VAFKGWLAEEEIDPVSLLKRLKEYGLEEIVHTEIEKDGTLQEHDFSLTKKIA  184   HisA_Tm
94  D-IGVENVIIGTKALDIS--FLERITSEF-EGITVSLDSKGGK-----IFTKGWIEEESLLVEDAYKMLRKFVNR-FVYTVIDRDGTLLG--IEKIKRFW  181   HisA_Pf
97  D-EGIDFVVIGTLAYKNKEEFLKILENFP-NRVILAIDSKQGK-----VAIGGWKEETAVSPEEFAKEYENYPIWGYLYTVIERDGSLEGVDVEPYKEIK  189   HisA_Aa
96  D-MGFERVILGTMAIENPAIVEELAGEYGSERIMVSLDSRDSR-----VVIRGWTEKVPFTAEEMARKFQERGAGSILFTNVDFEGLLSGFDLKPVSELV  189   HisA_Mht
97  E-LGVSRVILGTAALKNPELVQQLSSTFGNEHVTVALDAKNGK-----ISIKGWTEECAQTPVEMGRKFEELGAGSLLFTNIDTEGLMQGVNPVPTKNLV  190   HisA_Mb
93  S-LGVDRVIVGTKAILEPKFIDDLNKEIGKDKIVLAVECKEGK-----VVIKGWKEKVDKTPIEVIKEFEDK-VGYILFTNVDVEGLLKGINVDIIKELI  185   HisA_Mj
93  E-LGVDRVIFGTIAVERVEELREFAKKWK-GRVMVAIDSKKGR-----VAVKGWKEVVELTPVQLAELYDDL-DVSFLYTNIDVEGLVSGIERERIEEVV  184   HisA_Af
94  D-RGVDRVILGTAAIETPDIVGEIAAAYS-DGVLVSLDAKDGE-----VVVEGWTEGTGMDPVAAAQRYADLGAAGILFTDVDVEGKQEGVRTDPVRDLV  186   HisA_Hal
96  D-LGLNDVIIGTAAVKNPDFVREACKAFP-GRIIVGLDAKDGM-----AAIDGWATVTGHHVIDLAKRFEDDGVNSIIYTDIGRDGMMSGVNIDATVKLA  188   HisA_Nm
96  D-NGITYVIIGTAAVKVPGFLHDACYAFP-GQIMVGLDAKSGK-----VAVDGWSKVTGHDVIDLAKKFQDYGVEAIIHTDIGRDGMLSGLNIEATVELA  188   HisA_Ne
96  D-DGLSYVIIGTAAVKNPGFLQDACTAFG-GHIIVGLDAKDGK-----VATDGWSKLTGHEVADLARKFEDYGCESIIYTDIGRDGMLQGINIEATVRLA  188   HisA_Bf
```


```
    101       111       121       131       141       151       161       171       181       191       200
    |         |         |         |         |         |         |         |         |         |         |
96 D-DGLSYVIIGTAAVKNPGFLQDACTAFG-GHIIVGLDARDGK-----VATDGWSKLTGHEVVDLARKYEDYGVESIIYTDIGRDGMLQGINIDATVKLA  188   HisA_Rs
95 R-AGVSYVIIGTKAVKQPEFVGEACRAFP-GKVIVGLDAKDGF-----VATDGWAEVSEVQVIDLARRFEADGVSAIVYTDISKDGMMQGCNVEATAALA  187   HisA_Pa
95 K-AGVGYVIIGTKAVKQPEFVGEACRAFP-GKVIVGLDARDGF-----VATDGWAEVSSVQVVDLARRFEADGVSAIVYTDIAKDGMMQGCNVEATAALA  187   HisA_Av
95 D-AGVTYVIIGTKAVKEPEFVKQACDAFP-GHIIVGLDAKDGY-----VATDGWAEVSTVKATDLAKQFAADGVSEIVYTDIARDGMMQGVNIEATVEMA  187   HisA_Md
94 S-LGIGRVILGTAAQRNPELVEEACRLFP-GRIVVGIDAKDGM-----VAVQGWAEVTGVTAVDLAKRFEGYGVAAIIYTDIARDGMMQGPNIAATRALA  186   HisA_Gm
94 N-LGVDFAILGSVAARDPELVFRACEQFP-GRISVGIDARDGK-----VAVEGWAETTDLNAVDLAKKFEDAGVAEIIFTDIARDGTLTGPNVAATRLMA  186   HisA_Mag
94 S-RGLARVILGTVAVRDPVLVIEACKRFP-GQVAVGIDAKGGK-----VAVEGWAEASELGIIELAKRFEGAGVAAIIYTDIDRDGILAGINWASTLELA  186   HisA_At
94 S-RGLKRVILGTVAVRDPALVIEACRKFP-GRVAVGIDAKGGK-----VAVEGWAEASELGVVELARKFEGAGVAAIIYTDIDRDGILTGINWASTLELA  186   HisA_Srm
94 S-RGLRRVILGTVAVRDPALVMEACKAFP-GQVAVGIDAKGGK-----VAVEGWAEASRLGVIELAKKFEGAGVAAIIYTDIDRDGVLAGINWDSTLALA  186   HisA_Brm
94 D-KGVRRVILGTVALRDPQLVREACKRHP-GRIAVGIDARGGM-----VAVEGWAETSAVSALDLALRFEDAGVAAIVHTDIDRDGLLAGPNLDATAALA  186   HisA_Rr
97 E-KGITRVIIGTAAVRDPDLVKAAAKKFP-GRVAVGLDARDGK-----VAVEGWAETSQVTALEIAQRFEDAGVAAIIFTDIARDGLLKGLNLDATIALA  189   HisA_Bj
97 S-KGIARVIIGTAAVRDPALVKEAAKAFP-GRVAVGLDARDGN-----VAVEGWAESSQVTALDIAQRFEDAGVAAIIFTDIARDGLLKGINWDATIALA  189   HisA_Rpp
97 D-LGVSRVVMGTAALKDPQFVKDMAKEFP-GGIVVAVDARDGM-----VATEGWADVSDVSVVDLARRFEDAGVASLLFTDIGRDGLLKGVNLDATVELA  189   HisA_Na
94 E-AGVSRVILGTVAVHDPELVKKAARLWP-EQIAVAVDVRDGK-----VAVDGWTGLSDLSAIDLSRRFEDAGVAALIITDISRDGALTGVNVEGVGELA  186   HisA_Cc
94 C-KGIHRVILGTSAVKNPELVKQALKTFE-DSVVIGIDAKDGM-----VAIEGWAKTSEFTAIRFAKKMEELGAKTIIYTDISRDGMLAGPNLKAMEEMV  186   HisA_The
94 C-KGIHRVILGTSAVKNPELVKQALKTFE-DSVVIGIDAKDGM-----VAIEGWAKTSEFTAIGFAKKMEELGAKTIIYTDISRDGMLAGPNLKAMEEMV  186   HisA_Ct
96 K-LGVERVILGTVAVRNPELVEKAVQRFE-EAVVIGIDAKDGL-----VAVQGWAEKTEIKALDLALRMKKVGVKHLVFTDISRDGMLQGPNIQSTVELA  188   HisA_Dh
94 A-LGVQQAILGTVAIEQPELVSDFCQKFP-GQIIVGIDARDGK-----VATKGWLETSQVLATDLALQMAKLKVAAIIYTDIHRDGTLKGPNIEALRELA  186   HisA_Tre
94 N-LGVQWAILGTVAVEQPQLVQELCQQYP-GQIIIGIDARNGL-----VATRGWLETSEVLATQLAIQMQEMGAAAIIYTDIHRDGTLAGPNLDALRELA  186   HisA_Nos
94 D-SGVERAILGTVAVERPALVEALAGEFP-GQIAVGIDARSGK-----VAT-GAAGRFWLTAVALHSRWQTWRLRTDLHRHWAR-WHASGPNLEELRAIA  184   HisA_Syn
94 L-QGVNRAILGTVALENPKLVASLAAEYP-GRIWVGLDARDGY-----VATRGWLETSKILATDLAQNMAAMGVAGFVYTDIQRDGTLQGPNIPALRQLL  186   HisA_Te
94 A-YGLEQVILGTVAIEQPQLVKQLAQRNP-GRIIVGIDAKNGK-----VATRGWISQSEVNATDLASDFNAAGIAAIISTDIATDGTLEGPNLESLRAMA  186   HisA_Prm
94 S-VGINRVIIGSMAVKHPQFVKEALDKFGSDKIVVGIDAKNGM-----VATEGWLETSTVDYISLALEMEKMGVRLLIYTDVDRDGTLTGPNFEHYEKLI  187   HisA_Sm
94 E-SGIDRVIIGSAALTDPDFLRAAVQKYG-AKIAAGIDAKNGF-----VATRGWLDVSQVSYLDLAKRMEKVGVETIIYTDISRDGTLTGPNLEQMANLK  186   HisA_Lm
93 N-IGINRVIIGSAALTHPELVRTAVAKYGSDKIVVGVDGRDEK-----VATQGWLENSDTSFDDIVEAMLSVGVSNFVVTDIARDGTLSGPNIELLSRLQ  186   HisA_Lnm
94 EQIGVERVILGTAAVEKPDFLKELLIKYGPSKIIVGVDIREGF-----VSTSGWLEKTSLPYLSFLKKLERIGVKTTIITDISKDGTLTGPNFKLYDEIS  188   HisA_Ll
93 DKIGINRFILGTAAVTNPDLLKYALDNYGKERVVVAVDARDGI-----VKIAGWEKDSGIHYRDLLERLVKAGIEHIVFTDIAQDGTLAGPNLEAYREIL  187   HisA_Li
94 Q-AGVSRVILGTAAITDSVFLQTAVESYA-EKIAVSIDARNGF-----IATDGWTKNSTVEAIPFIKQLESIGVKTIIYTDILKDGMMSGPNFQELDAVQ  186   HisA_Oi
97 D-AGVERLIIGTIALTDPDLFGAMCSAFP-GRIGVSLDAEGGR-----LKTKGWVEDSGLSIYDVLPRLTEQGAAFLIYTDIDRDGMQSGVNVEALSELA  189   HisA_Dd
96 A-AGINYCIVGTKGIQDTDWLKEMAHTFP-GRIYLSVDAYGED-----IKVNGWEEDTELNLFSFVRQLSDIPLGGIIYTDIAKDGKMSGPNFELTGQLV  188   HisA_Sa
97 S-AGVERVILGSSAVSNPPFVKKMLKQYG-EKIAIGLDARNGF-----VSTEGWLETSTLKATELGKELANEGAEVFIFTDIATDGMLSGPNVKSTVELA  189   HisA_Bs
97 D-RGISRVILGSVAVNNPAFTKEMLQKYGGEKIAIGLDARDGY-----VATEGWLETSSVKAETLGQELARHGAEVFIFTDISRDGMLSGPNTEAIVQLA  190   HisA_Bh
94 D-IGVSRVVIGSAAVTNPGLIADLLKKYRPSQIVVGIDAEHGV-----PKIKGWTESSNMQDYELAGEMKKLGVERIIYTDITRDGMLQGVGYETTKRFA  187   HisA_Cht
```


```
    201        211        221        231        241        251        261        271        281
    |          |          |          |          |          |          |          |          |
193 ARYPQVAFQSSGG---IGDINDVAA-----LRGTGVRGVIVGRALLEGKFT-------VKEAIACWQNA---------------- 246  HisA_Sf
193 ARYPQVAFQSSGG---IGDIDDVAA-----LRGTGVRGVIVGRALLEGKFT-------VKEAIACWQNA---------------- 246  HisA_Ec
192 ARYPQIAFQSSGG---IGDIDDIAA-----LRGTGVRGVIVGRALLEGKFT-------VKEAIQCWQNV---------------- 245  HisA_Se
192 QRYPQVAFQASGG---IGCLDDIAR-----LRGSGVQGVIVGRALLDGKFN-------VKEAIACWQNV---------------- 245  HisA_Yp
192 SNFKNINFQASGG---VASLQDIIF-----LKKTGVKSIIIGRSLLEKKFT-------IEEAVKCWQRES---------------- 246  HisA_Bch
196 TQFPTIQFQSSGG---IGSIADIDA-----LKGTGVAGVIVGRALLEGKFN-------VQEAIACWQNA---------------- 249  HisA_Pm
196 EKYPPIQFQSSGG---IGSLADIEA-----LKGTGVSGVIVGRALLEGKFT-------LSEAIKCWQNG----------------. 249  HisA_Hi
192 RQYPQVQFQSSGG---IGSLADIEA-----LKGTGVAGVIVGRALLDGKFT-------AQEAFACWQNG---------------- 245  HisA_Vc
192 SRYPNIAWQASGG---IASLEDVAA-----VRDSGAAGIIIGKALLINQFN-------VAEAIQCWPNE---------------- 245  HisA_So
192 EIFPNICIQASGG---VASLKDLEN-----LKG-ICSGVIVGKALLDGVFS-------VEEGIRCLAN---------------- 243  HisA_Cj
195 ALAPQLQVQVSGG---ARNLADVAA-----AKAAGCAGIVLGKALLEGHLD-------LKDALAC------------------- 244  HisA_Xc
195 ALVPAVQLQVSGG---ARDVADVVA-----AKMAGCAGIVLGKALLEGRLA-------LKEAVQQGSVADPSDPLPCGELTEPVC 264  HisA_Xf
188 DRTN-VPVIASGGVSSLDDLRAIAK-----FTERGIEGAIVGKALYAERFT-------LPQALAVVRM---------------- 242  PriA_Ml
188 DRTD-APVIASGGVSSLDDLRAIAT-----LTHRGVEGAIVGKALYARRFT-------LPQALAAVRD---------------- 242  PriA_Mt
186 SRTD-KPVVASGGISSLDDLRALAA-----LVPEGVEGAIMGTALYEGAFT-------LEEALATVREVAQ------------- 243  PriA_Tf
185 AATD-RPVVASGGVSSLDDLRAIAG-----LVPAGVEGAIVGKALYAKAFT-------LEEALEATS----------------- 238  PriA_Sc
186 ERTD-AKVTASGGISKLDDLRAIKE-----LAEIGVDSAILGKSLYARAFT-------LQEALEVAK----------------- 239  PriA_Bl
188 AATD-APIVASGGISTLEDVLGLAK-----YQDEGIDSVIIGKALYEHRFT-------LAEALEAVEKLG-------------. 244  PriA_Ce
188 AATD-APIVASGGISVLEDVLELAK-----YQDEGIDSVIIGKALYEHKFT-------LEEALAAVEKLG--------------- 244  PriA_Cg
190 GLWH-GELIVGGGVADTNDVRLLA--------EEGIEGAIVGRAIYEGTLP-------YPVRL------------------- 236  HisA_Dr
182 KRIARRV-FYAGGVRDCNDVERAYA--------LGFDGVIVGYALYAGDLK---------KCANWY---------------- 229  HisA_Pya
185 AKLIRGVKEYAGGVSSDSDITFLKN--------VGFDYIIVGMAFYLNKIR---------GTNVV----------------- 232  HisA_Ss
185 IEAE-VKVLAAGGISSENSLKTAQKVHT--ETNGLLKGVIVGRAFLEGILT--------VEVMKRYAR-------------- 241  HisA_Tm
182 G--DEEFIYA-GGVATPTDIIQLAEIG--------FSGVIIGKALYEGTIS-------LEEALEVAKSVGKKNYSSS-------- 240  HisA_Pf
190 KHVK-KPVIASGGVSSLEDIKKLYGI---------VEGVVVGKAIYEGRIT-------LEDLQNF----------------- 237  HisA_Aa
190 EAVE-IPVIYSGGVSSLNDLRMLQG--------TGVMGVVIGSAIYRGLID-------LTEALKYQDLK------------- 242  HisA_Mht
191 ESVS-IPVIASGGVSSLQDLRTLKK--------TGASGVVVGSALYTGRFT-------LEAAIETTQQE------------- 243  HisA_Mb
186 EKTD-IPIIYSGGITTLEDIKALKE--------LGIYGVVIGSALYKGLID-------LKKALEIVKN------------- 237  HisA_Mj
185 KSLR-NPVYVAGGISSIEDIRFAKR--------VGAAGVVIGSALYTKKLK-------FEEAIKVEYEKV------------ 238  HisA_Af
187 DSVD-IPVIASGGVATVDDVRALEA--------AGAAGAVVGTALYEGNFT-------LADAQAAVED------------- 238  HisA_Hal
189 QAVR-IPVIASGGLTGLDDIRALCA-----AEKRGVAGAITGRAIYEGSID-------FAQAQQLADSLD------------ 245  HisA_Nm
189 QALT-IPVIASGGVTNLDDIRKLCE-----VQSEGITGVITGRAIYQGSLD-------FKEAQALADQLDAATI--------- 249  HisA_Ne
189 RAVK-IPVIASGGLSNLADIESLCE-----VEDEGIEGVICGRAIYSGDLD-------FAAAQTLADRLRESDDA-------- 250  HisA_Bf
```


```
    201        211        221        231        241        251        261        271        281
    |          |          |          |          |          |          |          |          |
189 QSVS-IPVIASGGLSSLKDIDHLCA-----VESEGVEGVICGRAIYSGDLN-------FREAQDHADKLGAA------------  247  HisA_Rs
188 NATR-IPVIASGGIHNLGDIQKLLD-----ARTPGIIGAITGRAIYEGTLD-------VAEAQALCDNFKA------------  245  HisA_Pa
188 AATR-IPVIASGGIHDLGDIRKLLD-----ARAPGIVGAITGRAIYEGTLD-------VAEAQALCDGFKG------------  245  HisA_Av
188 RAGN-IPIIASGGITNMDDIKGLMA-----VAHEGITGAITGRAIYEGSLD-------LAEAQAYCDAQ-------------  243  HisA_Md
187 EAIS-IPVIASGGVSSLKDIENLMT-----IEASGIAGAITGKAVYTGAIN-------LAEAVALTKRGGA------------  244  HisA_Gm
187 ESAT-IPVIASGGVSCLADIQALLENSGPYPNGNRISGVITGKAIYDGRLD-------LAAAIALSRRWEN------------  249  HisA_Mag
187 DAVS-IPVIASGGLASIDDIKRMLQ-----PDAAKLEGAISGRALYDGRID-------PTEALDLIKAAKEVRA----------  247  HisA_At
187 DAVS-IPVIASGGLASMDDIRRMTE-----PDAQKLEGAISGRALYDGRID-------PKEALDLIREARKGKL----------  247  HisA_Srm
187 EAVS-IPVIASGGLASMEDIRRLAT-----PEMRKLEGAISGRALYDGRID-------PAEALSVLRAAA------------  243  HisA_Brm
187 ERLT-TPVIVSGGVASLDDLRAIRA-----RAAAGLEGVISGRALYDGRID-------LAEALAVLAD-------------  241  HisA_Rr
190 DSIS-IPVIASGGLASIDDVKAMLT-----PRAKKLAGAIAGRALYDGRLD-------PAAALTLIRNARAA----------...  248  HisA_Bj
190 EAIS-IPVIASGGLASIEDVRTLLS-----PRAKKLEGAIAGRALYDGRLD-------PTEALALIGAAKAA----------..  248  HisA_Rpp
190 RRVD-IPVIASGGVKGIDDIRMLKI-----HEGDGIEGVITGRALYDGRLD-------LKAAIEMAEA-------------  244  HisA_Na
187 DAVS-IPVIASGGVAAVADIERLKA-----RQGVEIAGAILGRSLYAGTIK-------PSEALIAAAA-------------  241  HisA_Cc
187 KAVN-IEVIASGGVRNIDDIRNLKN--------VGVSGVIVGKALYTGDLD-------LKEAIEVAK--------------  237  HisA_The
187 KAVN-IEVIASGGVRNIDDIRNLKN--------VGVSGVIVGKALYTGDLD-------LKEAIEVAK--------------  237  HisA_Ct
189 RLSG-LQVVASGGVSRLEDLRLLQEE---ANRGVSLEGAIVGKALYAGAFS-------LAEALRVVGQRSEGK---------  250  HisA_Dh
187 TAVS-IPVIASGGISSINDLLNLLA-----LETIGVKGAIVGRALYTGGIS-------LKEANQAVGQGRWQDVPPNLGSSTFV-  257  HisA_Tre
187 AAIS-IPVIASGGVSSVTDLLSLLA-----LEPQGVTGVIVGRALYTGDIS-------LKEALQAIGPGRIQDIPPNLDFSSFA-  257  HisA_Nos
185 AAVS-IPVIASGGVGSLTDLLSLLP-----LEAQGVSGVIVGKAFYTGAVD-------LQEALRAIGSGRWQDVAVD-DSSRLA-  254  HisA_Syn
187 AVTN-RPVIASGGVSSLNDILSLFA-----LTPHGLVGAIVGKALYTKAMD-------LREAVRAVGQGRWQDIPPDLGSTTWA-  257  HisA_Te
187 NASS-VPVIASGGVGCMADLLSLLA-----LEPYGVSGVIVGRALYDGKVD-------LKEAIRAIGDGRLQDPPTSKPLMA---  255  HisA_Prm
188 AHLTSAKVIASGGIHALSDLQQL--------EKIGVAGTIVGKAYYNGDIS-------LKELREFEG-----------------  239  HisA_Sm
187 EHVK-VNLIASGGVSSRADLEAL--------AQLGLYGAIAGKALYNHHIS-------MSDIVEVEQIAY-------------  240  HisA_Lm
187 NKFPKSNIIASGGIANIKNVTDL--------QASGIHDIIVGRALYDGDVT-------LAQLKEVDG-----------------  238  HisA_Lnm
189 KESS-LNVIISGGVKDNSDIQR--------ATRSDFYGIIVGKAYYEKKIN-------LEKEFRNAN-----------------  239  HisA_Ll
188 NSYP-FQVIASGGIASLKDLMDLSSL----KTKISLYGVITGKALYEGKLD-------LAKAISSI-----------------  241  HisA_Li
187 QATT-MNIIASGGVTTEKDVQQL--------KTMNLYGAIIGKALYDGSIK-------LQDILEGEMDAR------------  240  HisA_Oi
190 VRSA-IPVIAAGGVATMQDVYDLYP----VSRSANLQGAITGKAIYEGSLD-------LAEAMRWIDSRTA------------  248  HisA_Dd
189 KATT-IPVIASGGIRHQQDIQRL--------ASLNVHAAIIGKAAHQAS---------FWEGLE-----------------  234  HisA_Sa
190 KETG-KSVIASGGVSSVADLEALAR-----NEADGVSGAIIGKALYTNQFT-------LSEALERVKRK------------  245  HisA_Bs
191 KATG-KQVIASGGVSSLQDLDALRN-----EKQNGIAGAIVGKALYTNQFT-------LEDALKEES-------------  244  HisA_Bh
188 EKAG-MKVTASGGATTSDDLHKLRS-----LEKYGVDSVIIGKALYECNFPCQELWYAYEQGLGIDGEFSTARKKECCS------  260  HisA_Cht
```

3.2.1.3 Analysis of the SPB motifs

As proposed in Section 2.2.5 an obvious explanation for the bifunctionality of PriA relies on its broad specificity for the substrates PRA and ProFAR. Therefore, differences are likely to be found in the SPB motifs of PriA. The first difference that becomes apparent lies precisely in the SPB motif of the N-terminal end of PriA. As it can be seen, the first glycine of the glycine triplet at the C-terminal end of β strand 3 is invariant in all sequences other than in actinomycetes sequences (Gly79, Gly80 and Gly81 in *T. maritima*) where it is replaced by either threonine or serine. The serine residue is conserved in all actinomycetes other than in the two *Corynebacterium* species, where a threonine replaces the glycine. It is tempting to speculate that this idiosyncrasy of the *Corynebacterium* PriA proteins is related to the presence of a separate TrpF protein in these organisms (see Sections 2.1 and 4.2.2.3).

How can such a subtle change of glycine for serine account for the dual substrate specificity of PriA? This question has no straightforward answer without the elucidation of the structure of PriA. Nevertheless, some ideas can be delineated beforehand. By analogy to the structure of tHisF (Lang *et al.*, 2000) the third glycine residue of the triplet is proposed to be involved in PriA in the binding of one of the phosphate moieties of ProFAR. Although the replaced residue in PriA is Gly79 and not Gly81 the overall hydrophobicity of the phosphate-binding domain is critical for binding of the substrate and it is likely to be affected by this subtle change. Moreover, the possibility that the first glycine residue of the triplet in PriA could be structurally and/or functionally equivalent to the third glycine of the triplet in tHisA cannot be ruled out (see Todd *et al.*, 2001).

The putative effect of this replacement is further highlighted by the topological constraints of serine and threonine with respect to glycine. The former two amino acids have similar uncharged polar side chains whereas glycine is the simplest possible amino acid with no side chain. The replacement therefore could not only affect the charge of the N-terminal substrate-binding pocket of PriA, but also its size and shape, since serine and threonine are more bulky than glycine. It is proposed therefore that *via* this change PriA has gained some of the plasticity required for binding both PRA and ProFAR. This could be in relation to the observation that the N-terminal SPB motif in PriA resembles more closely the conserved SPB motif identified in TrpF sequences (Wilmanns *et al.*, 1991). Although TrpF only has the C-terminal SPB motif to which the phosphate moiety of PRA is known to bound, it could be that PriA can make use of either SPB sites to accommodate PRA.

Another interesting observation dealing with the composition of the SPB motifs in PriA involves the invariant aspartic acid in the C-terminal end of β strand 8 of TrpF (Wilmanns *et al.*, 1991; Henn-Sax *et al.*, 2002). The role of this residue is unknown but its conservation highlights its putative importance. Moreover, the phosphate moiety of PRA binds to a hydrophobic residue in tTrpF, which is next to the Asp residue in question (Henn-Sax *et al.*, 2002). As seen in the alignment there is no Asp or similar residue in the β strand 8 of the HisA sequences. In contrast, the PriA sequences appear to have either an aspartic acid or a closely related glutamic acid in an equivalent position to the Asp residue in TrpF. This makes the C-terminal SPB motif of the PriA sequences look more similar to the TrpF SPB site. Although these observations can only be

corroborated by direct experimentation, the identification of such differences can direct future mutagenesis work.

3.2.1.4 Analysis of the catalytic residues

As mentioned previously, the catalytic residues of tHisA are conserved in the *S. coelicolor* PriA enzyme. As the alignment shows there are only two exceptions where the residue Asp127 of *T. maritima* is not conserved; these are in the PriA proteins from the corynebacteria species. In these sequences the Asp residue has been replaced by alanine, although there is an Asp residue just two positions away towards the N-terminal end that could be structurally and/or functionally equivalent. When looking at the replacements reported by Jürgens *et al.* (2000) that led to the evolution of TrpF activity in the scaffold of tHisA, none of them occured in the PriA sequences. Only one of them lays at a C-terminal end of a β strand, namely, the catalytically essential Asp127 in *T. maritima* (Jürgens *et al.* 2000; Henn-Sax *et al.*, 2002). Jürgens *et al.* (2000) claimed that replacement of this single residue by a valine in the scaffold of tHisA "is sufficient to convert the specificity of an enzyme for substrates from different metabolic pathways". Despite the fact that the mechanism of reaction of this mutant enzyme has resisted elucidation (Henn-Sax *et al.*, 2002), it has been speculated that the increased hydrophobicity in this region conferred by the valine residue may play a role in the PRA isomerase activity (Jürgens *et al.*, 2000). Although I do not agree entirely with the interpretation of their results (see Section 3.2.4.1), it is worthwhile to note that the flanking residues of the equivalent Asp residue in the PriA sequences are invariantly

hydrophobic residues, a pattern that is not necessarily conserved across the rest of the sequences (Figure 3.2).

3.2.2 Model structure of PriA

The conclusions drawn from the multisequence alignment discussed in the previous Section were reinforced by the construction of a model of the 3D structure of PriA. This model was constructed by Dr D. Lee during a visit by the author of this thesis to the biocomputational group of Prof C. Orengo at University College London. For this purpose, the molecular coordinates of the structure of HisF were used. As it can be seen in Figure 3.3, the model shows that the putative catalytic residues of PriA, Asp11, Asp130 and Thr166, are as expected pointing towards the internal face of the β-barrel. Moreover, the model predicts that the side-chain of residue Ser81, which replaces a glycine residue invariant in all other HisA sequences (see Section 3.2.1.3), protrudes towards the active site. Therefore, this supports the view that *via* possession of this residue PriA might have some of the plasticity required for binding PRA and ProFAR.

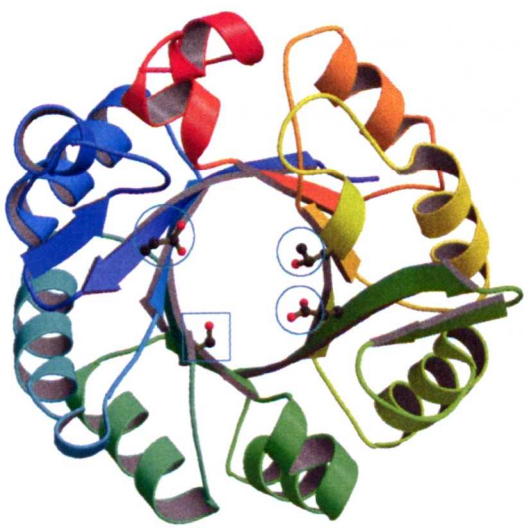


Figure 3.3 Model structure of PriA. The model structure of PriA was done using as template the coordinates of the structure of HisF from *T. maritima* (PDB 1thf). The model shows the side-chains of the presumed catalytically important residues Asp11, Asp130 and Thr166 (circles) and the actinomycetes-specific residue Ser81 (square) present in the SPB site at the N-terminus of PriA. As it can be observed, the side-chain of Ser81 protrudes into the active site of the β-barrel, which contrasts with the invariant glycine of all HisA sequences. See also Section 3.2.1.3.

As introduced in Section 1.1.5, the specificity for substrates can evolve through different mechanisms (Todd *et al.*, 2001 and references therein). What might be an antecedent to the broad substrate specificity of PriA is the strategy adopted by the $(\beta\alpha)_8$-barrel endo-β-*N*-acetylglucosaminidase H from *Streptomyces plicatus* (Rao *et al.*, 1995). This glucosaminidase lacks the fifth and sixth α helices of the barrel in such a way that the large protein moiety of its substrate can be accommodated (Van Roey *et al.*, 1994; Rao *et al.*, 1995; Todd *et al.*, 2001). Another possibility involves the use of accessory domains as exemplified by members of the glycosyl hydrolases superfamily (cited in Todd *et al.*, 2001). In members of this superfamily β-galactosidase domains appear to have a structural role by providing loops that alter the size of the substrate-binding pocket in such a way that small sugar molecules are bound (Juers *et al.*, 1999). Although none of these peculiarities are apparent from the multisequence alignment (Figure 3.2) or the model structure of PriA (Figure 3.3), these possibilities cannot be ruled out until the elucidation of the structure of PriA.

3.2.3 Expression and purification of PriA, eTrpF and eHisA

The availability of a purification protocol of PriA is essential for its biochemical and structural characterisation. The following Sections report on the optimisation of a protocol for the purification of recombinant PriA to homogeneity. The quality of the protein obtained proved to be suitable for crystallisation trails and enzyme assays. A simple purification protocol of HisA (eHisA) and TrpF (eTrpF) from *E. coli* is also reported.

3.2.3.1 Expression of GST fusions of PriA, eTrpF and eHisA

The fact that GST-PriA fusions produced from pGEX-4T-1 (pGEXpriASc) complement *trpF* and *hisA* deletions in *E. coli* implies that soluble and active PriA is produced in this system (see Section 2.2.4.1 and Table 2.2). Therefore, the production of recombinant PriA from pGEXpriASc was pursued. A similar approach was followed for the production of eHisA (pGEXhisAEc) and eTrpF (pGEXtrpFEc), which would serve as controls in future experiments (similar to the experiments reported in Section 2.2.4). All together different parameters were systematically tested in order to optimise the expression levels of the *priA*, *trpF* and *hisA* fusions. The parameters that were experimentally varied were time-point of induction with IPTG (from OD_{750} 0.5 to 0.9); concentration of IPTG used for induction (from no induction to 100 mM IPTG); and temperature and period of incubation (from 2 hours at 37 °C to 20 hours at 20 °C). In addition, a variety of *E. coli* strains were used as host including W3110 *trpC*(*F*del), HfrG6, BL21(DE3), Rosetta and C41 (see Section 7.2 in Methods).

Although the GST-PriA fusion was shown to be expressed consistently from pGEXpriASc at high levels, it proved to be insoluble in all the conditions tested (data not shown). This observation contrasts with the high levels of soluble protein that were obtained from the pGEXhisAEc and pGEXtrpFEc systems. The GST-eHisA and GST-eTrpF proteins were overexpressed in *E. coli* strains Hfr G6 and W3110 *trpC*(*F*del), respectively, to avoid contamination from the host in future enzyme assays. In these systems, soluble eHisA and eTrpF GST fusions are obtained even at fairly standard conditions, i.e. induction with 1mM IPTG at an OD_{750} of 0.6 and overexpression for 2-3 hours at 37 °C. It is interesting to note that Henn-Sax *et al.* (2002) reported that for some

unknown reason IPTG inhibits expression of the *hisA* gene of *E. coli* on a pET derivative, which seems not to occur in this system. In addition, a hexahistidine-tagged version of eTrpF was overexpressed from the construct pMS401. This construct is a pJB122 derivative (Altamirano *et al.*, 2000) that contains a modified version of *trpF* from *E. coli*, which expresses high levels of soluble PRAI (Patrick and Blackburn, personal communication).

3.2.3.1.1 Expression of soluble PriA

The discrepancy between the complementation of *trpF* and *hisA* by *priA* in *E. coli* (see Section 2.2.4 and Table 2.2) and the insolubility of PriA can be explained in two ways. It could be that either very low levels of soluble PriA are required for complementation in *E. coli* or solubility of PriA is lost during its extraction. In agreement with the latter possibility, it has been reported that aggregates are often produced after expression due to the insolubility of the recombinant protein in the buffer used for preparation of the crude extracts (Lindwall *et al.*, 2000). To test this possibility a sparse matrix composed by thirty different buffers was used for generation of crude cell extracts as previously described (see Section 7.3.2 in Methods). The increase in solubility of overexpressed PriA was analysed by SDS-PAGE electrophoresis. Unfortunately, none of the buffers tested gave better results (data not shown), suggesting that low levels of soluble PriA are enough for complementation in *E. coli*.

It was speculated that the presence of GST could be affecting the folding and therefore the solubility of PriA. To test this hypothesis, the *priA* gene from *S. coelicolor* was cloned into different expression vectors. Vectors that would lead to a his-tagged

version of PriA at the N-terminal end were chosen, because the active site of PriA is predicted to be located towards the C-terminal end. This might reduce the likelihood of a deleterious effect on the folding of PriA (Dabrowski and Kur, 1998; Henn-Sax *et al.*, 2002). Overexpression of *priA* carried out from the vectors pProEX and pET-15b shows somehow contrasting results (for cloning details see Table 7.1 in Methods). For some unknown reason, the level of expression of the *priA* gene from pProEX (pProEXpriASc) was very low as judged by SDS-PAGE electrophoresis (data not shown). This might be related to the inhibition effect of IPTG reported by Henn-Sax *et al.* (2002). In contrast, *priA* was expressed at decent levels from the pET-15b construct (pETpriASc) albeit the majority of PriA produced remain in the insoluble fraction. After several rounds of optimisation, the best balance between soluble and insoluble PriA can be seen in Figure 3.4. To sum up, the optimal conditions for the overexpression of *priA* using the pETpriASc system includes the use of *E. coli* C41 as host (Miroux and Walker, 1996) bearing the auxiliary plasmid pRIL, which encodes rare Arg, Ile and Leu tRNAs, induction with as low as 1mM IPTG at an OD_{600} of 0.7 and expression overnight (16-20 hours) at 20 °C.

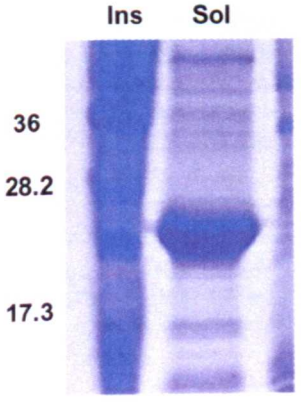


Figure 3.4. Production of soluble 6Xhis-PriA from pETpriASc. The solubility of PriA was analysed by SDS-PAGE electrophoresis. The gel shows the best ratio of soluble (Sol) and insoluble (Ins) protein after several rounds of optimisation (see text for optimal conditions). The sample of the soluble fraction was ten times concentrated. The approximate position of the molecular markers is shown next to the gel in the left-hand side (17.3, 28.2 and 36 kDa). The predicted size of the his-tagged version of PriA is 26 kDa.

3.2.3.2 Purification of 6Xhis-PriA, GST-eTrpF and GST-eHisA fusions

The 6Xhis-PriA fusion produced from pETpriASc was purified taking advantage of its engineered nickel affinity (see Section 7.3.2 in Methods). However, it was discovered that PriA elutes with washing buffer containing as little as 10 mM imidazole. For some unknown reason 6Xhis-PriA binds to the resin very loosely, making its purification through this means virtually impossible. The results of an imidazole gradient carried out to optimise its concentration can be seen in the SDS-PAGE shown in Figure 3.5. The best results were obtained when the resin was washed twice with buffer containing 10 mM imidazole followed by a single wash with buffer containing 20 mM imidazole. The 6Xhis-PriA fusion is easily eluted from the resin with buffer containing 100 mM imidazole. 6Xhis-PriA protein prepared in this fashion showed consistently PRA isomerase activity which is not due to contamination carried from the expression host as shown by the lack of activity of the pET-15b negative control treated in the same way. Although a loss in yield was preferred over the presence of impurities during the optimisation of this protocol, further steps were required for the purification of 6Xhis-PriA to homogeneity.

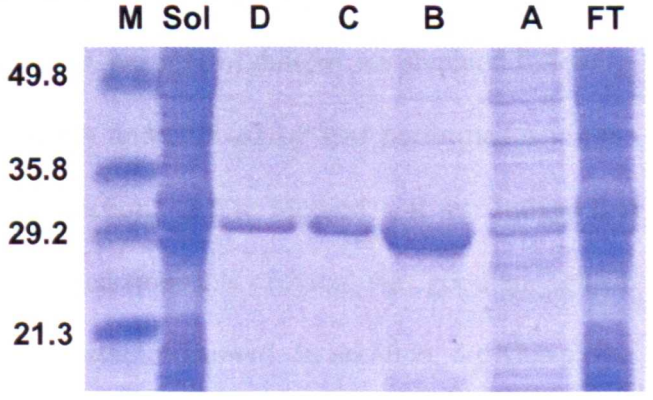


Figure 3.5 Purification of 6Xhis-PriA through nickel affinity. Imidazole gradients were used to optimise the washing steps: 10mM (A), 50 mM (B), 100 mM (C) and 200 mM (D). PriA elutes at very low imidazole concentration (20 mM). The flow-through (FT) and the soluble (Sol) fraction are also shown. The size of the molecular markers (M) is shown next to the gel in the left-hand side.

The impurities remaining after the Ni-affinity purification were removed by size-exclusion chromatography. The elution buffer contained 0.5 mM DTT and 200 mM NaCl because it has been reported that salt stabilizes HisA of *Salmonella typhimurium* (Margolis and Goldberger, 1966). In the system used 6Xhis-PriA elutes consistently after 98 minutes (or 195 ml of buffer at a flow rate of 2 ml/min) as a discrete peak consisting of three fractions (31-33) that is preceded by a shoulder (Figure 3.6A). Fractions 31 to 33 proved to have PRA isomerase activity (see below Section 3.2.4) whereas the fractions of the preceding shoulder were inactive. Interestingly, Henn-Sax *et al.* (2002) reported a similar elution pattern while purifying eHisA, the significance of which is unknown. Analysis of the fractions 31 to 33 through SDS-PAGE electrophoresis revealed that they contain a single protein with the predicted size for the 6Xhis-PriA fusion (Figure 3.6B). Therefore, these fractions were pooled and concentrated to ~10 mg/ml.

The molecular size of purified 6Xhis-PriA calculated from SDS-PAGE electrophoresis and size-exclusion gel filtration fits well with the predicted 26 kDa for the 6Xhis-PriA fusion. Furthermore, confirmation of the identity of PriA was done by N-terminus sequencing of purified 6Xhis-PriA (H. Wright, personal communication). Approximately 15 mg of purified 6Xhis-PriA per litre of culture are obtained following this protocol. The purification of GST-eHisA and GST-eTrpF was performed using the MicroSpin GST purification module (see Section 7.3.2 in Methods). It was judged that the purity of the protein preparations obtained from this protocol was good enough for serving as controls in the enzyme assays (data not shown). In addition, a 6Xhis-eTrpF protein from pMS401 was purified by nickel affinity (data not shown).

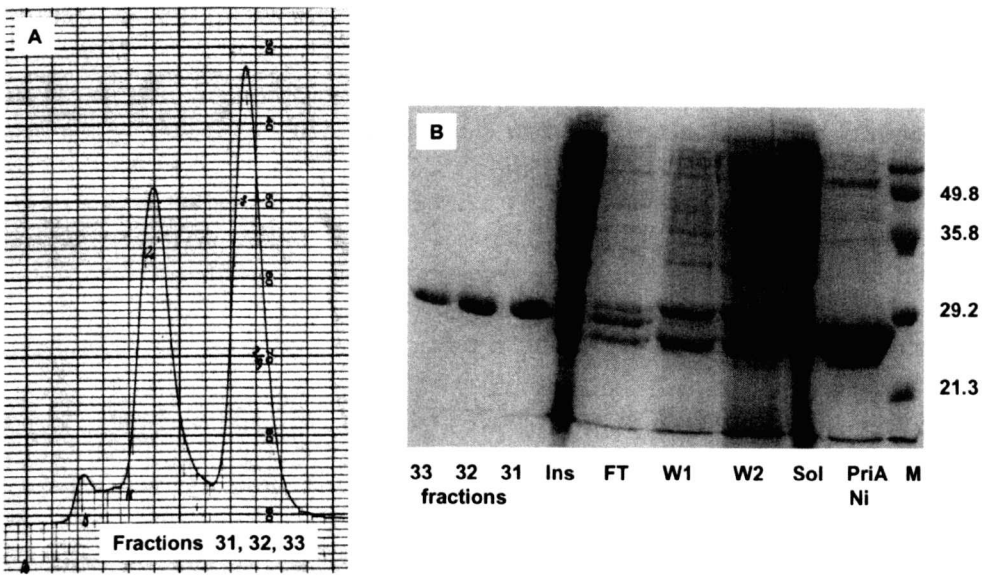


Figure 3.6 Purification of 6Xhis-PriA through size-exclusion gel filtration. A. Chromatogram showing the 6Xhis-PriA peak (fractions 31-33), which elutes after 98 minutes. PRA isomerase activity can be detected in fractions 31 to 33 whereas no activity is detected in the fractions composing the preceding shoulder. **B.** SDS-PAGE electrophoresis analysis of a typical result of the protocol followed: insoluble fraction (Ins), flow-through (FT), pool of the two 10 mM imidazole washes (W1), 20 mM imidazole wash (W2), soluble fraction (Sol) and 6Xhis-PriA purified through nickel affinity (PriA Ni). The impurities remaining were successfully removed by size-exclusion chromatography (fractions 31, 32, 33).

3.2.4 PRA isomerase activity of 6Xhis-PriA

Biochemical assays for the isomerase activities of TrpF upon PRA (Kirschner *et al.*, 1987; Eberhard *et al.*, 1995; Hommel *et al.*, 1995; Jürgens *et al.*, 2000; Henn-Sax *et al.*, 2002) and HisA upon ProFAR (Klem and Davisson, 1993; Davisson *et al.*, 1994; Lang *et al.*, 2000; Jürgens *et al.*, 2000; Henn-Sax *et al.*, 2002) have been previously reported. However, in assays intended for determination of Michaelis-Menten kinetic parameters several difficulties were associated with the instabilities of PRA and ProFAR (Martin *et al.*, 1971; Kirschner *et al.*, 1987; Henn-Sax *et al.*, 2002). If the substrates PRA (Kirschner *et al.*, 1987) and ProFAR (Martin *et al.*, 1971) were produced chemically they needed to be synthesized and purified freshly prior to the enzyme assays. Although

these protocols are available from the literature they require the use of HPLC. Unfortunately, this preparative equipment was not available in our laboratories at the time these experiments were undertaken.

An alternative to the chemical synthesis of the substrates was to produce them enzymatically *via* coupled-enzyme assays involving the enzyme that catalyses the preceding reaction in the pathway. This approach has been exploited in the determination of the kinetic parameters of HisA (Davisson *et al.*, 1994; Lang *et al.*, 2000; Jürgens *et al.*, 2000; Henn-Sax *et al.*, 2002) and TrpF (Eberhard *et al.*, 1995; Hommel *et al.*, 1995; Jurgens *et al.*, 2000; Henn-Sax *et al.*, 2002). In the TrpF system PRA was produced enzymatically *in situ* from anthranilic acid and a large molar excess of phosphoribosyl pyrophosphate (PRPP) by the activity of PRA transferase or TrpD. It was then assumed that anthranilic acid was completely converted into PRA avoiding the need for its purification. In the HisA system ProFAR was produced in a similar fashion. However, determination of the kinetic values of HisA was even more complicated because in addition to the preceding enzyme there was the need for purified HisF to convert PRFAR into a more traceable product since the UV spectra of ProFAR and PRFAR are not significantly different (Klem and Davisson, 1993; Lang *et al.*, 2000; Jürgens *et al.*, 2000; Henn-Sax *et al.*, 2002). The number of purified enzymes required for these experiments was beyond the scope of this thesis considering the time constraints imposed.

It was only possible to determine the PRA isomerase activity of PriA qualitatively. PRA can be synthesized *in situ* spontaneously from a mixture containing equimolar quantities of anthranilic acid and ribose-5-phosphate in 50% ethanol (Patrick

and Blackburn, personal communication; Kirschner *et al.*, 1987). PRA isomerase activity was then measured by detection of the production of 1-(o-carboxyphenylamino)-1-deoxyribulose-5-phosphate (CDRP) at 350 nm (see Section 7.3.1 in Methods). Making use of this simple assay the PRA isomerase activity of the 6Xhis-PriA, GST-eTrpF and GST-eHisA enzymes was determined. As expected, the PriA and eTrpF fusions show PRA isomerase activity whereas the GST-eHisA enzyme proved to be inactive. This *in vitro* result confirmed that PriA is capable of performing the TrpF reaction on its own, nullifying the possibility that an unidentified protein associates with PriA *in vivo* conferring the TrpF activity deduced from the phenotype of the mutant *S. coelicolor* WH101.

With this simple enzyme assay on hand, the PRA isomerase activity of PriA was routinely detected during the purification procedure at different steps. It was found that PriA was very unstable and lost activity very easily even at room temperature. For this reason the entire purification procedure was carried out at 4 °C. It was also found that PRA isomerase activity of 6Xhis-PriA was lost after digestion with thrombin intended to remove the hexahistidine-tag, even though the digestion efficiency was very poor as judged by SDS-PAGE electrophoresis (data not shown) and western blot analysis with a histidine-tag antibody (H. Wright, personal communication). Therefore, since it can be expected that the histidine-tag would not affect the folding and activity of PriA (Dabrowski and Kur, 1998; Henn-Sax *et al.*, 2002) a histidine-tagged version of PriA was used for the crystallisation experiments.

3.2.4.1 Comparison of the catalytic efficiency of GST-PriA and the mutant enzymes of Jürgens and co-workers (2000).

Although a quantitative comparison between the bifunctional naturally occurring PriA and the artificially produced HisA-TrpF enzymes of Jürgens and co-workers (2000) cannot be done due to the lack of PriA steady-state kinetic values, circumstantial evidence suggests that PriA is far more efficient. The observation that the *E. coli* W3110 *trpC*(*F*del) colonies bearing the mutated HisA enzymes with TrpF activity took "several days to appear" implies that the efficiency of these enzymes was very poor (Jürgens *et al.*, 2000). This contrasts with the ease by which *priA* complements the *trpF* deletion of the same strain overnight even without IPTG induction. The relationship between catalytic efficiency and ability to complement the enzyme function *in vivo* has been previously experimentally demonstrated (Yano *et al.*, 1998; Henn-Sax *et al.*, 2002).

The poor catalytic efficiency of the HisA-TrpF mutant variants is also inferred from the steady-state kinetic parameters reported by Jürgens *et al.* (2000); the highest catalytic PRA isomerase efficiency k_{cat}/K_m ($mM^{-1} \cdot s^{-1}$) reported for any of the mutant variants was 0.49 compared to $13.3 \cdot 10^3$ of wild type tTrpF. Moreover, the only mutant variant that proved to have both TrpF and HisA activities showed the poorest parameters: an identical catalytic PRA and ProFAR isomerase efficiency k_{cat}/K_m ($mM^{-1} \cdot s^{-1}$) of 0.046 compared to $13.3 \cdot 10^3$ and 1.2 of wild type tTrpF and tHisA, respectively. Based on this data, Jürgens *et al.* (2000) concluded that the turnover number (k_{cat}) was the trait selected for in their experiments. They reasoned that PRA accumulates in the *trpF* minus mutant used for selection *in vivo* implying that small values of K_m seem not to be as crucial for cell growth as were increased values of k_{cat}. It is puzzling, therefore,

that after the presentation of this argument it is stated, "a single amino acid exchange is sufficient to convert the *specificity* of an enzyme for substrates". This statement seems to be at odds with the residues replaced in the mutant variants, which are not even remotely related to residues identified as important for binding of the substrates such as those within the SPB motif, although it has become recurrent in enzyme directed-evolution studies that residues that do not contact the substrate directly are frequently replaced and have an overall influence on the evolution of substrate specificity (Perona and Craik, 1997; Yano *et al.*, 1998). In contrast, it is interesting to note that the mutant variant that is claimed to have acquired TrpF activity by exchange of a single residue is generated by replacement of one of the catalytically crucial Asp residues (see also Section 3.2.1.2). It could be that the apparent TrpF activity detected in this mutant variant corresponds in reality to very low levels of residual activity as suggested by the *catalytic promiscuity* hypothesis (Jensen, 1976; O'Brien and Herschlag, 1999; Gerlt and Babbit, 2001b) which was uncovered only when the HisA reaction cannot be accomplished and an excess of PRA is present possibly forcing its non-specific binding. In other words, it might appear that the results of Jürgens *et al.* (2000) are an artefact intrinsic to the system understudy. In contrast, the ease by which PRA isomerase activity is detected in PriA suggests that it is likely that the trait selected for (or retained) during the course of natural evolution is substrate affinity. This would imply that PriA has an ancient-like broad-substrate specificity and might be a suitable model for studying enzyme and pathway evolution.

3.2.5 Crystallisation of 6Xhis-PriA

With the availability of the purification protocol and activity assay of PriA, a very fruitful and active collaboration with Dr H. Wright of the Warwick Protein Structure group was established in order to crystallise and elucidate PriA's structure. The screening of the crystallisation conditions was performed using the hanging drop vapour diffusion method (see Section 7.3.3 in Methods). After screening several commercial trials and optimising those conditions that showed initial promising results, it was found that PriA crystallised in 100 mM sodium citrate buffer, pH 5.0 containing 2 M ammonium sulphate as precipitant. Thoma *et al.* (1999) have reported that tHisA crystallizes in four different conditions all of which involve PEG as precipitant within the pH range from 5.6 to 6.6. We did not find any indication of crystallization of PriA under similar conditions. Crystals of PriA (Figure 3.7) were frozen using 35% ethylene glycol in the mother liquor and were shown to diffract X-rays to 2.4 resolution (data not shown) using in-house facilities. The unit cell dimensions of the crystals were about a = b = 63 Å and c = 105 Å. Two native data sets have been collected so far at synchrotron X-ray sources at a resolution of 2 Å and structure determination is currently in progress (H. Wright, personal communication).

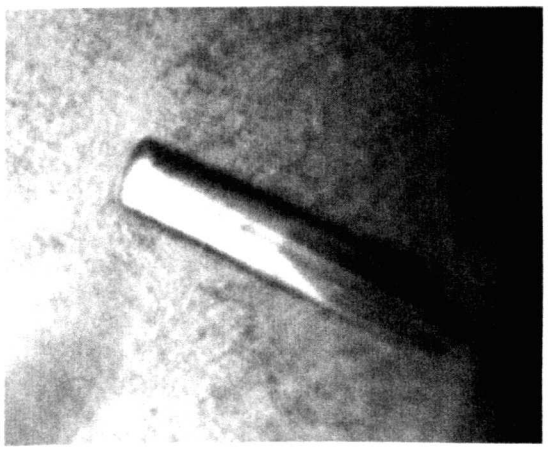


Figure 3.7 Crystals of 6Xhis-PriA. The crystal shown is 0.3 x 0.08 x 0.08 mm in size. For crystallization conditions refer to text.

3.3 Conclusions

The degree of similarity of the PriA homologues from actinomycetes suggests that this unusual enzyme might be present across this group of bacteria. A detailed computational analysis of the sequence of PriA has pinpointed some peculiarities at the amino acid-level that might be related to the broad-substrate specificity of this enzyme. Of particular relevance seems to be a replacement of a glycine residue by a serine or threonine residue in actinomycetes (serine 81 in *S. coelicolor* PriA) that putatively modifies the active site of the β-barrel to allow the binding of both PRA and ProFAR.

A protocol for the expression and purification of PriA that overcomes its low heterologous solubility and instability is reported. The availability of this protocol endowed the beginning of the biochemical and structural characterization of PriA. Although no kinetic parameters were determined, it was found qualitatively that PriA shows PRA isomerase activity in accordance with the physiological role proposed in Chapter II. Crystallization conditions of PriA were reported and structure determination was underway at the time of submission of this thesis.

Chapter IV

Chapter IV

"Genomic context analysis of the *his* and *trp* genes of actinomycetes: regulatory and evolutionary deductions"

4.1 Introduction

4.1.1 Genomic organisation of the *his* and *trp* genes: the operon paradigm

Soon after Jacob, Monod and co-workers reported their operon model for the co-ordinated expression of the *lac* catabolitic genes (Jacob *et al.*, 1960), the operon paradigm was adapted to the pioneering studies on tryptophan biosynthesis (reviewed by Yanofsky, 2001 and 2003) and histidine biosynthesis (reviewed by Alifano *et al.*, 1996). Because of this precedence the expectation is that the *his* and *trp* genes will be found in microbial genomes in operons, e.g. see the textbooks by Dale (1998) and Wagner (2000). This expectation has been reinforced by the fact that an operon seems to be advantageous to the cell since it allows co-ordinated expression of functionally related genes through the use of regulatory mechanisms such as transcription repression and/or attenuation. For reviews of these regulatory mechanisms see Yanofsky (2001 and 2003) and Henkin and Yanofsky (2002).

While the *his* and *trp* operons have been demonstrated and extensively characterised in model organisms such as *E. coli* and *B. subtilis* (Alifano *et al.*, 1996; Yanosky, 2001 and 2003; Berka *et al.*, 2003 and references therein) it has become apparent even before the current availability of a large collection of microbial genome-

sequences that different organisations of the *his* genes (Limauro *et al.*, 1992; Fani *et al.*, 1995) and *trp* genes (Crawford, 1989 and Doolittle *et al.* 1992) are frequently encountered. This divergence goes beyond gene order and it is also observed in the regulatory mechanisms by which expression of these genes is controlled. For instance, despite *E. coli* and *B. subtilis* showing similar organisation of their *trp* genes, the regulation of each operon involves different attenuation mechanisms (Yanofsky, 2001 and 2003). The notion that the organisation of the *his* genes (Fani *et al.*, 1998) and *trp* genes (Dandekar *et al.*, 1998) is not conserved was confirmed when more microbial genome-sequences became available. This notion turned out to be in agreement with the observation of a general lack of gene order conservation that has been reported after comparison of multiple microbial genome-sequences (Mushegian and Koonin, 1996; Watanabe *et al.*, 1997; Overbeek *et al.*, 1999; Itoh *et al.*, 1999; Huynen *et al.*, 2000; Wolf *et al.*, 2001; see also section 1.1.3.1).

4.1.2 Genomic organisation of the *his* and *trp* genes in *S. coelicolor*

In contrast to what it is observed in the model organism *E. coli*, where all the *trp* genes form an operon, the *trp* genes in *S. coelicolor* are found in three different loci. This observation is further complicated by the presence of paralogous *trp* genes probably involved in secondary metabolic pathways (see also Chapter V). However, in spite of this lack of gene clustering, the *trpE3*, *trpC1*, *trpB* and *trpA* genes are close to each other within a region in the chromosome included in cosmid Sc4G6 (Redenbach *et al.*, 1996; Bentley *et al.*, 2002). The *trpC1*, *trpB* and *trpA* genes are part of a cluster that includes an ORF of unknown function (*trpX* or SCO2038) located downstream *trpC1*

and just before *trpB*. Furthermore, the *trpE3* gene is found upstream the *trpCXBA* cluster separated only by three small genes (see Figures 2.3 and 4.3).

Previous studies in our laboratory showed that *trpC1* is transcribed from an independent promoter whereas the *trpXBA* genes form an operon (Hu, 1995; Hu *et al.*, 1999). Furthermore, it is well known that *trpC1*, *trpB*, and *trpA* are involved in primary metabolism since mutation of these alleles result in tryptophan auxotrophic strains (Hopwood *et al.*, 1973; Hopwood *et al.*, 1985 and references therein). These mutant strains were used to clone the *trpC1BA* genes and the *trpD* gene, which is elsewhere on the chromosome, by complementation (Hu *et al.*, 1999). It seems therefore that the *trp* genes present in the Sc4G6 locus are not organised as an operon although they should be co-regulated in order to be able to synthesise tryptophan. One scenario, which is explored in this chapter, is that the *trp* genes contained in cosmid Sc4G6 are part of a *superoperon* involving internal promoters and more than one operon. The term *superoperon*, which is synonymous to *supraoperon*, is used to define regulons, i.e. sets of genes co-ordinately regulated, that are physically linked on the chromosome. It is proposed herein that the putative Sc4G6 *trp* superoperon may be analogous to the complex regulatory organisation described for other gene systems such as the so-called *aromatic supraoperon* of *B. subtilis* and relatives. The latter system is formed by a co-regulated cluster of genes that include: the *trp* genes (other than *trpG*); the *aroF*, *aroB*, and *aroH* genes, which are upstream the *trp* genes; and the *hisC* (called *hisH* in *E. coli*), *tyrA*, and *aroE* genes just downstream of the *trp* genes (see also section 2.2.5; Berka *et al.*, 2003; Panina *et al.*, 2003 and references therein).

The locus containing the *trpE3* and *trpC1XBA* genes in *S. coelicolor* is very close to the *priA* gene, the latter being only seven genes upstream. The *priA* gene is part of a cluster of *his* genes that appear to form a *his* operon similar to the one found in *E. coli* (Limauro *et al.*, 1990 and 1992). However, the product of *priA* is also involved in tryptophan biosynthesis as demonstrated in the two previous Chapters (Barona-Gómez and Hodgson, 2003). This cluster of *his* genes lacks the *hisG* and *hisE* genes, which are elsewhere on the chromosome, and the unidentified $hisB_p$ gene. Furthermore, the *hisI* gene is not part of this *his* operon since there are three genes (one transcribed in opposite direction) between *hisF*, the last gene of the *his* operon, and *hisI*. Nevertheless, the *hisI* gene is just upstream *trpE3* potentially forming an operon and representing a putative functional and physical link between the *his* and *trp* genes (see Figures 2.3 and 4.3). Overall this means that five out of the seven enzymic functions required for synthesis of tryptophan from chorismate are potentially encoded within the Sc4G6 locus, although the functionality of *trpE3* still needs to be demonstrated (see Chapter V). Similarly, this locus includes the majority of the enzymic functions required for synthesis of histidine. The *S. coelicolor* locus where the majority of the *his* and *trp* genes are found in the chromosome, as shown in Figure 2.3, and its orthologous counterparts in other organisms (Figure 4.3), will be referred to as the *HT cluster* from now onwards. It is worth noting that the HT clusters discussed in this Chapter are composed not only by *trp* and *his* genes but also by a variety of other genes, some of them of unknown function.

4.1.3 Conservation of gene order and its use to predict gene function

The central paradigm of genomic context analysis rests in the concept of *gene order conservation* (for a review see Osterman and Overbeek, 2003). Gene order is traditionally defined as conserved if sets of genes transcribed in the same direction with intergenic regions no exceeding 300 base pairs (i.e. directons) are present in two or more genomes of distantly related organisms (Overbeek *et al.*, 1999). Since it is well acknowledged that bacterial genomes are very unstable, as suggested by the general lack of gene order conservation (see Sections 1.1.3 and 1.1.3.1 and references therein), the discovery of conserved clusters (or strings) of genes implies that such conservation might reflect a functional constraint. Therefore, "there is, for all practical purposes, little doubt that the respective genes (of a conserved cluster) comprise an operon and are functionally linked" (Wolf *et al.*, 2001). Within this context related functionality refers to the gene products as being involved in the same metabolic or cellular function but not necessarily that they share the same functional category (functional categories as defined for Clusters of Orthologous Groups (COG); see http://www.ncbi.nlm.nih.gov/COG/).

Evidence supporting the central corollary of genomic context analysis, i.e. that conserved gene order can hint at gene function, has accumulated in the last few years as a consequence of the boom in bacterial genome-sequencing projects (Mushegian and Koonin, 1996; Watanabe *et al.*, 1997; Dandekar *et al.*, 1998; Overbeek *et al.*, 1999; Itoh *et al.*, 1999; Huynen *et al.*, 2000; Wolf *et al.*, 2001; Zheng *et al.*, 2002; Yanai *et al.*, 2002; Snel *et al.*, 2002). The principle has been successfully exploited to assign function correctly to several gene products, for example, in efforts aimed to reconstruct metabolic

pathways (Osterman and Overbeek, 2003 and references therein). Indeed, this principle was intuitively exploited in Section 2.2.2.2 of Chapter II for the discovery of PriA and reconstruction of the tryptophan biosynthetic pathway in *S. coelicolor*.

4.1.4 Genomic context analysis of *priA*

The purpose of this chapter is to investigate to what extent the physiology associated with the bifunctionality of PriA discovered in *S. coelicolor* is spread across the bacterial world. This aim was pursued by a comparative genomics analysis focused on the genomic context of *priA*. A large number of bacterial genomes that were completely sequenced while this work was in progress were systematically exploited (almost one hundred by April 2003! See complete list at http://www.tigr.org/tigrscripts/CMR2/CMRGenomes.spl). These include the genomes of eight representatives of the actinomycetes clade with different life-styles and taxonomic positions (Table 4.1 and Figure 4.1). The availability of these genome sequences represents the opportunity of looking at the evolution of the Gram-positive organisms with a high (G + C) content, otherwise called actinomycetes, for the first time from a holistic perspective. It has been previously anticipated that "multiple genomes separated by intermediate evolutionary distances such as representatives of different genera within the same bacterial family could be particularly helpful for making the best use of gene order conservation" for predicting gene function and inferring genome evolutionary histories (Wolf *et al.*, 2001). This corollary is derived from the fact that closely related organisms would show no significant differences since they have diverged too recently (Himmelreich *et al.*, 1997)

whereas gene order in distantly related organisms is not conserved (see sections 1.1.3.1 and 4.1.2 and references therein).

In addition to the genome sequences of *S. coelicolor* M145 (Bentley *et al.*, 2002), *M. tuberculosis* H37Rv (Cole *et al.*, 1998) and *M. leprae* TN (Cole *et al.*, 2001) the list of actinomycetes that have been fully sequenced includes: the gastrointestinal commensal *Bifidobacterium longum* NCC2705 (Schell *et al.*, 2002); the avermectin-producing streptomycete *Streptomyces avermetilis* ATCC 31267 (Ōmura *et al.*, 2002; Ikeda *et al.*, 2003); the glutamic-acid-producing species *Corynebacterium glutamicum* ATCC 13032 (Bathe *et al.*, 1996) and *Corynebacterium efficiens* YS-314^T (Fudou *et al.*, 2002); and the unusual endopathogen *Tropheryma whipplei* TW08/27 (Bentley *et al.*, 2003). The genome of the latter shows a dramatically reduced size of less than one million base pairs and, as expected from an organism with an endopathogenic life-style that has undergone reductive evolution, *Tp. whipplei* is a metabolic cripple lacking the majority of the amino acid biosynthetic genes including the *his* and *trp* genes (Bentley *et al.*, 2003). Therefore, the genome of *Tp. whipplei* was not included in the analysis. Some of the characteristics of the chromosomes of the actinomycetes that were compared are summarised in Table 4.1. Their approximate taxonomic position within a 16s rRNA-based phylogenetic tree of the actinomycetes is shown in Figure 4.1.

Several other actinomycetes whose genomes have been partially sequenced or whose annotation is in progress were also used (e.g. different genomes from the *Mycobacterium* complex and the pathogen *Corynebacterium diphtheriae*). Although the genomes of the latter organisms were not directly used in the comparative analysis, mainly because their sequences have not been annotated, they were often investigated to

determine the presence or lack of a *trpF*-like gene in any given genus. However, the genome sequence of *Thermobifida fusca* YX (formerly known as *Thermonospora fusca*; Zhang *et al.*, 1998), a moderate thermophilic cellulose-degrading organism, was included because although its genome has not been fully sequenced, the genomic sections already available include the *his* and *trp* genes on an annotated fashion. Finally, when more than one genome-sequence of a given species was available, only the genome of one strain is reported since it was found that different strains do not show different gene profiles nor organisation of their *his* and *trp* genes (e.g. *M. tuberculosis* strains H37Rv and CDC1551).

Table 4.1 General characteristics of the chromosomes of the actinomycetes compared.

Species* (Family/Suborder)	Strain	Size of the chromosome (topology)	G + C content	Coding density	Average gene length
Streptomyces coelicolor (Streptomycineae)	M145	8, 667, 507 bp (Linear)	72.12%	88.9%	991 bp
Streptomyces avermitilis (Streptomycineae)	ATCC 31267	9, 025, 608 bp (Linear)	70.7%	86.2%	1, 034 bp
Thermobifida fusca[£] (Streptosporangineae)	YX	~ 3, 700, 000 bp (?)	?	?	?
Mycobacterium tuberculosis (Corynebacteriaceae)	H37Rv	4, 411, 532 bp (Circular)	65.61%	90.8%	1,012 bp
Mycobacterium leprae (Corynebacteriaceae)	TN	3, 268, 203 bp (Circular)	57.79%	49.5%[¤]	1,011 bp
Corynebacterium efficiens (Corynebacteriaceae)	YS-314[T]	~ 3, 147, 000 bp (Circular)	?	?	?
Corynebacterium glutamicum (Corynebacteriaceae)	ATCC 13032	~ 3, 309, 000 bp (Circular)	?	?	?
Bifidobacterium longum (Bifidobacteriaceae)	NCC2705	2, 256, 646 bp (Circular)	~ 60%	~ 86%	?
Tropheryma whipplei (Micrococcineae)	TW08/27	925,938 bp (Circular)	43.6%	84.4%	998 bp

* For references of the corresponding genome-sequencing paper, when available, refer to text. The family or suborder is as in the taxonomy browser of the NCBI (http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=201174/).
£ The genome of *Tb. fusca* has not been fully sequenced.
? These features have not been published.

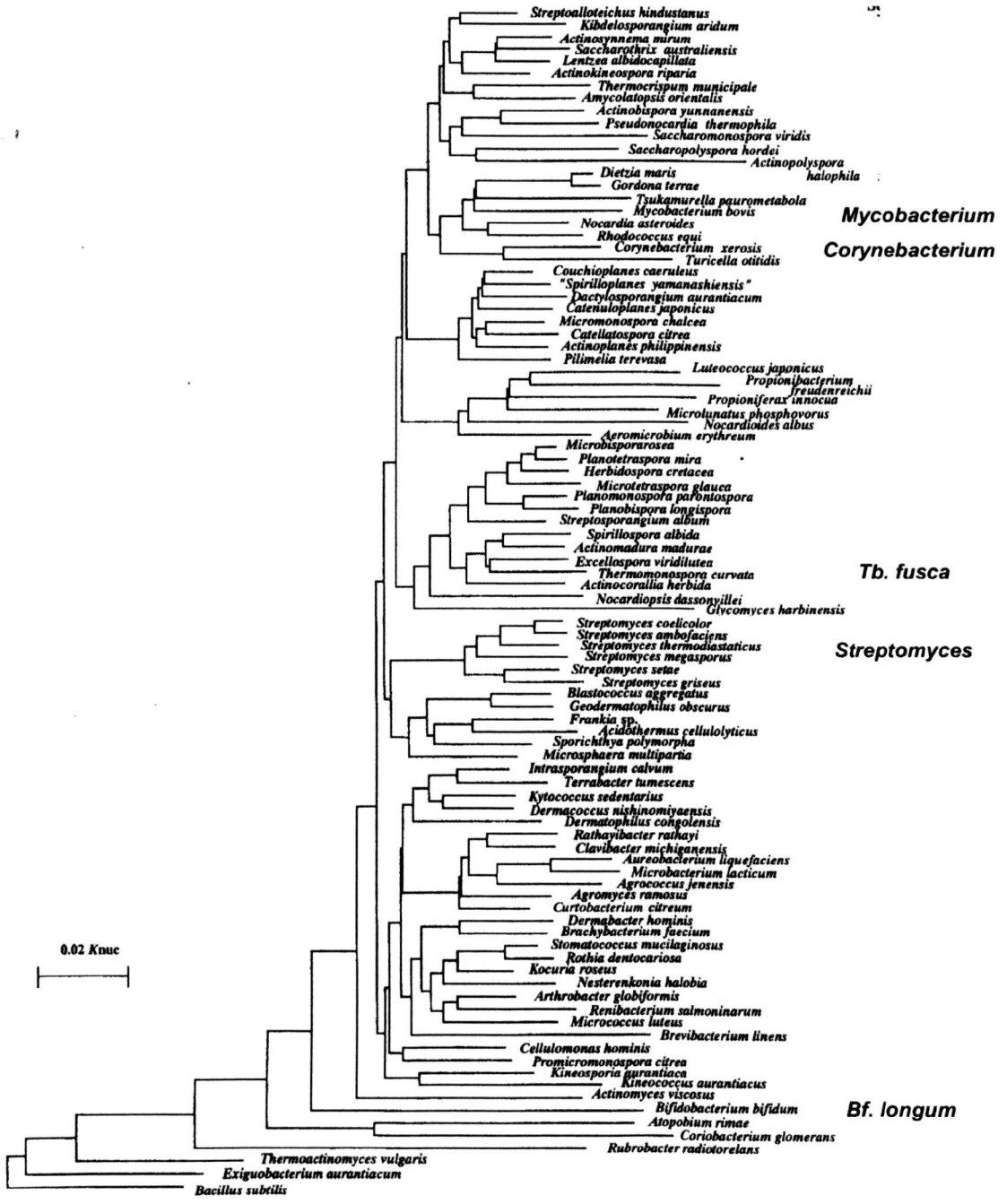


Figure 4.1 Phylogenetic tree of the actinomycetes clade. The tree was constructed using 16s rRNA sequences of 90 genera of actinomycetes. The approximate taxonomic position of the organisms compared in this Chapter is shown in the right-hand side in blue. The figure was taken from *The Atlas of Actinomycetes*, published by the Society for Actinomycetes of Japan (Miyadoh *et al.*,1997).

4.1.4.1 Rational for the discovery of PriA proteins in bacterial genomes

The rational behind this comparative genomics study is founded in two assumptions. First, the lack of a *trpF* gene in any given microbial genome is indicative of a non-orthologous gene displacement similar to the one demonstrated for *S. coelicolor* in previous Chapter (Barona-Gómez and Hodgson, 2003). Second, it is assumed that the clustering of the *his* and *trp* genes within the Sc4G6 locus in *S. coelicolor* is driven by the functional constraint imposed by the bifunctionality of *priA*. At a first glance, the clustering of the *trp* and *his* genes at the surroundings of *priA* could be in accordance with either the selfish operon hypothesis (Lawrence and Roth, 1996) or the notion that genes tend to cluster as a consequence of the benefits associated with co-regulation (de Daruvar *et al.*, 2002; see also section 1.1.3.1). These two possibilities are discussed in the light of our results. Finally, a phylogenetic tree of PriA and HisA homologs was also constructed using the multisequence alignment shown in Figure 3.2. This was aimed at detecting possible evolutionary relatives of PriA. Some of the hypotheses derived from the comparative genomics analysis were tested *in vivo* in the laboratory by means of intergeneric complementation studies and mutational analysis.

4.2 Results and Discussion

4.2.1 Methodological considerations and definitions

The comparative genomics study reported herein is the result of sequence comparisons making use of standard bioinformatics tools routinely used in any

molecular biology laboratory. The programs, databases and sequences are publicly available from the Internet (see Section 7.4 in Methods). Moreover, the programs used can be run in any personal computer without specialised hardware or software. The simplicity of the approach used, based on the determination of reciprocal homologous-pairs at the amino-acid level as detected by BLASTP searches in the eight actinomycetes genomes investigated, highlights the power of comparative genomics, particularly when some *a priori* knowledge is on hand. A cutoff *e* value of 10^{-10} was used as a very conservative criterion to determine homology between two genes (Parkhill, 2002), although the expectation value detected was more often well below this threshold. Furthermore, in order to avoid confusion several evolutionary terms are defined in the following sections and used accordingly.

4.2.1.1 Definition of evolutionary terms used in the analysis

When referring to similarity of two genes or proteins within an evolutionary context (i.e. homology), several considerations and definitions apply and must be used carefully (Sonnhammer and Koonin, 2002). One of the distinctions more often overlooked is that between *orthologues* and *paralogues*. Two genes are orthologous to each other when they descend from the same gene present in the last common ancestor of *different* organisms, and they appear to perform the *same* cellular or metabolic function. This definition is the basis of the classification of genes into *clusters of orthologous genes* or COGs extensively used in bioinformatics (see http://www.ncbi.nlm.nih.gov/COG/). In contrast, two genes are paralogous when they result from a gene duplication event within the *same* organism and are more likely to be

involved in *different* cellular or metabolic functions. Different evolutionary scenarios can arise from the combination of these evolutionary possibilities but will not be introduced further (for additional reading see Itoh *et al.*, 1999; Parkhill, 2002; Sonnhammer and Koonin, 2002).

4.2.1.2 Definition of *clusters of orthologous genes*

The definition of *clusters of orthologous genes* used in this analysis is similar to the definition used for COGs (see above). More specifically, after the criteria proposed by Itoh *et al.* (1999), a pair (or cluster) of orthologous genes is defined when: (1) orthologous ORFs between two genomes compared are the most similar ORF reciprocally; (2) similarity of the pair shows statistical significance; and (3) if a particular ortholog shows more similarity to certain paralogues within the genome, all of the paralogues are regarded as being orthologous to the counterpart of the other genome. Although these criteria were proposed for defining orthologous gene pairs when comparing only two genomes, it can be extended to COGs as the number of genomes compared is more than two. Thus, a *cluster of orthologous genes* is defined as a set of genes conformed by homologous genes that appear to have the same cellular or metabolic function in two or more organisms.

4.2.1.3 Definition of *orthologous strings of genes*

When performing genomic context analyses a COG is only relevant when in addition to homology and common functionality the orthologous genes appear in the chromosome of different organisms at *homologous* positions. A homologous

chromosomal position is defined as common gene neighbourhood or conserved genomic context, i.e. that the genes in the proximity of the gene analysed are conserved in two or more genomes. In addition to this criterion, the conserved genes need to be transcribed in the same direction and to be separated between them by no more than 300 base pairs (Overbeek *et al.*, 1999; see also section 4.1.3). For the sake of clarity such "orthologous clusters of genes" will be termed *orthologous strings of genes* (OSG) in order to avoid confusion with clusters of orthologous genes defined in Section 4.2.1.2. Moreover, the term *string* avoids confusion because the term cluster, particularly within the *Streptomyces* community, is often associated with a group of genes involved in the same metabolic pathway (e.g. clusters of secondary metabolic genes), which are potentially co-regulated forming a superoperon rather than a single operon (see also Section 4.1.2). In contrast, the likelihood that an OSG forms an operon is statistically very high (see Section 4.1.3 and references therein) and therefore the terms OSG and operon are indistinctively used during this analysis despite the lack of experimental evidence supporting a polycistronic transcription.

4.2.1.4 Definition of *über-operons*

An OSG or operon is the highest level of conservation and as discussed in previous sections is rarely found. However, gene context conservation of a higher order than operons has previously been reported and appears to be more common (Lathe *et al.*, 2000; Rogozin *et al.*, 2002). The set of genes that is conserved at a higher level of organisation than operons has been coined the term *über-operon* (Lathe *et al.*, 2000) and different levels of gene order conservation are possible. Based on the terminology of

Itoh *et al.* (1999) über-operons can be identical; similar (translocations, deletions and two insertions are allowed); deconstructed (two or more orthologues are present but the organisation is not conserved); and unknown (gene order conservation cannot be estimated). Due to the large number of genes comprising an über-operon it is more likely that these genes form superoperons rather than single operons. Furthermore, in contrast to what is expected from the gene content of an operon, where all the genes are functionally related, the genes forming an über-operon are not necessarily involved in the same metabolic or cellular function.

4.2.2 Phylogenetic co-occurrence of (the lack of) a *trpF* gene.

The question addressed in this Section is how common the lack of a *trpF* gene is across bacteria. The actinomycetes were first investigated and then the analysis was extended to the bacterial complement. On one hand, if the lack of *trpF* is a consequence of an ancient evolutionary experience with, therefore, broader ecological and physiological implications this phylogenetic co-occurrence should be spread at least across the actinomycetes and possibly beyond. On the other hand, if the lack of *trpF* is a consequence of a recent evolutionary event this phylogenetic profile should be unique to the genus *Streptomyces* or even species-specific. Answering this question has a bearing in the discussions presented in Chapter III on the evolution of the $(\beta\alpha)_8$-barrels since it will hint to how close to the ancestor of the $(\beta\alpha)_8$-barrels PriA might be. This analysis was aided by the web-server STRING (http://www.bork.embl-heidelberg.de/STRING/), which integrates multiple forms of genomic context analysis and indicates the reliability of the predictions (Snel *et al.*, 2000; von Mering *et al.*, 2003).

4.2.2.1 The *Mycobacterium* genus lacks a *trpF* gene

At early stages of this investigation it became apparent that the genomes of *S. coelicolor*, *M. tuberculosis* and *M. leprae* were similar in that no *trpF* gene could be identified (see Section 2.2.1 and references therein). The genome of another strain of *M. tuberculosis*, the clinical isolate CDC1551, has been sequenced confirming this view (Fleischmann *et al.*, 2002). Furthermore, different incomplete genome sequences of *Mycobacterium* species have become available during the last couple of years including those of *M. bovis*, *M. smegmatis*, *M. marinum* and *M. avium* (for a list of genomes whose sequencing is in progress see http://wit.integratedgenomics.com/GOLD/). Investigation of the partial genome-sequences of these organisms by BLAST searches failed to detect a *trpF*-like gene suggesting that the lack of a *trpF* gene is conserved across the genus *Mycobacterium*. Therefore, it is tempting to speculate that a non-orthologous gene displacement similar to that of *S. coelicolor* as previously reported had occurred in the *Mycobacterium* lineage (see also Sections 4.2.3.3 and 4.2.3.3.1).

4.2.2.2 Other actinomycetes lack a *trpF* gene

A similar scenario to the one described for *Mycobacterium* species was encountered in other actinomycetes. No hint of a *trpF*-like gene was found in *Bifidobacterium longum* NCC2705 (Schell *et al.*, 2002). This scenario was confirmed by BLAST searches in a second strain of *Bf. longum*, strain DJO10A, albeit the genome of this strain has not been fully sequenced. Furthermore, the genome sequence of *S. avermitilis* (Ōmura *et al.*, 2001; Ikeda *et al.*, 2003) and the incomplete genome sequence of *Tb. fusca* also lack a *trpF* gene. The case of *Bf. longum* is particularly interesting

because the bifidobacteriales are close to the root of the phylogenetic tree of the Gram-positives with a high (G + C) content (Figure 4.1), implying that *Bf. longum* diverged very early in evolution. Therefore, the lack of a *trpF* gene in this most diverged organism suggests that, whether a PriA protein is confirmed in this genus, it may go back in evolution to the last common ancestor of the actinomycetes. This would support the view that the bifunctionality of PriA is an ancestral feature and not a consequence of a recent event.

4.2.2.3 *Corynebacterium* has a *trpF* gene acquired by lateral gene transfer

As mentioned in Section 2.1 it was known from before the genome-sequencing era that *C. glutamicum* has a functional *trpF* gene within a *trp* operon (Matsui *et al.*, 1986). Based on the organisation of the *C. glutamicum trp* operon, and the sequence similarity between these *trp* genes to those found in enteric bacteria, it has been proposed that *C. glutamicum* acquired this *trp* operon relatively recent by lateral gene transfer (LGT) from an enteric bacterium as donor (Crawford, 1989). In the course of this investigation the genomes of *C. glutamicum* and *C. efficiens* were fully sequenced. Additionally, an incomplete version of the genome sequence of *C. diphtheriae* became available (see http://www.sanger.ac.uk/ Projects/C_diphtheriae/). BLAST searches of these genomic sequences showed that not only *C. glutamicum* has a *trpF* gene within an enteric-like *trp* operon, but so do *C. efficiens* and *C. diphtheriae*. The notion that this *trp* operon was acquired by LGT is further supported by the phylogenetic proximity of *Corynebacterium* to the genus *Mycobacterium* (Figure 4.1), where neither a *trpF* gene nor an enteric-like *trp* operon is present. Thus, the discovery of a *trpF* gene in members

of the *Corynebacterium* genus does not contradict the view that a lack of a *trpF* gene is a distinctive phylogenetic co-occurrence profile of the actinomycetes.

4.2.2.4 A *trpF* gene is conserved in the bacterial complement

The phylogenetic co-occurrence of *trpF* in the bacterial complement was investigated making use of STRING (see Section 7.4 in Methods). At the time the searches were done the STRING web-server included a total of 89 fully sequenced genomes of which 65 were from bacteria. Among the latter the genomes of *S. coelicolor*, *M. tuberculosis* (strains H37Rv and CDC1551), *M. leprae* and *C. glutamicum* were already included in the database. As expected, the phylogenetic co-occurrence of *trpF* in these actinomycetes, i.e. the lack of a *trpF*-like gene, was confirmed other than in *C. glutamicum*. Furthermore, the analysis also showed that *trpF* is conserved in the bacterial complement investigated. However, as seen in Figure 4.2, there are other examples of organisms that also lack a *trpF* gene, but whose genomes are also lacking other *trp* genes. Moreover, the sub-group of these organisms consists of endopathogens (e.g. *Streptococcus pyogenes* and *Chlamydia pneumoniae*) suggesting that the lack of *trp* genes in the genomes of these organisms is a reflection of their lifestyles and a consequence of reductive evolution. It is well acknowledged that endopathogens are prone to lose metabolic capabilities and to rely upon resources supplied by the host (e.g. see Cole *et al.*, 2001). Thus, the endopathogenic scenario for the *trp* genes is of a different nature to the lack of *trpF* in actinomycetes and therefore it can be stated that the *trpF* phylogenetic profile of actinomycetes is a distinctive feature of this lineage.

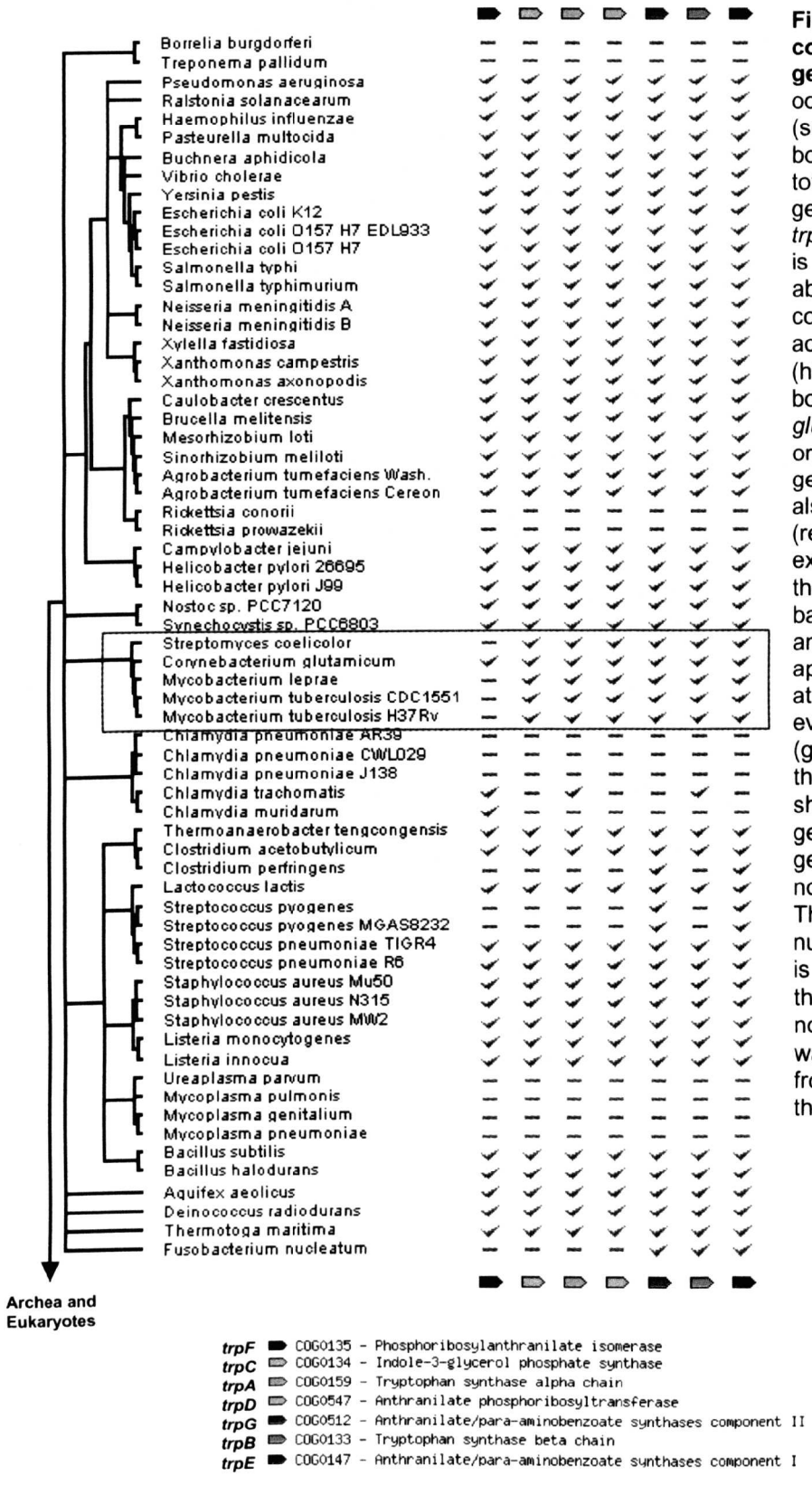


Figure 4.2 Phylogenetic co-occurrence of the *trp* genes in eubacteria. The occurrence of the *trp* genes (see nomenclature at the bottom of the figure) in a total of 65 bacterial genomes is shown. The *trpF* gene (PRA isomerase) is shown in red and its absence (orange dash) is confirmed for the actinomycetes clade (highlighted with a blue box) other than in *C. glutamicum*. Other organisms also lack a *trpF* gene but these bacteria are also lacking other *trp* genes (refer to text for explanation). Therefore, the complement of the bacterial genomes analysed contain, or appear to have contained at one point during evolution, a *trpF* gene (green mark). Please note that the scheme does not show the position of the *trp* genes in the corresponding genomes, i.e. whether or not they cluster together. The corresponding COG numbers for the *trp* genes is given at the bottom of the Figure within the nomenclature. The Figure was downloaded directly from the results obtained at the web-server STRING.

4.2.3 Genomic context analysis of *priA* reveals a conserved HT cluster in actinomycetes

Inspection of the genomic context of the *priA* or *hisA* homologues in the eight actinomycetes genomes compared showed that a HT cluster (possibly an über-operon rich on *his* and *trp* genes) is highly conserved (Figure 4.3). In six out of the eight genomes investigated a similar HT cluster of *S. coelicolor*, as described in section 4.1.2 and shown in Figures 2.3 and 4.3, was detected. In addition to *S. coelicolor* the organisms where a conserved HT cluster was found are *S. avermitilis*, *M. tuberculosis*, *M. leprae*, *C. glutamicum* and *C. efficiens*. The conserved HT cluster typically contains two orthologous strings of genes or operons, i.e. an OSG at the 5' end rich in *his* genes and an OSG at the 3' end mainly composed by *trp* genes but which can start with a *hisI* gene. More precisely, the *his* operon shows the organisation $hisDCB(X)_nHA(X)_nFI$ and the OSG at the 3' (or *trp* operon) the gene order $hisI(X)_ntrpE(X)_nC(X)_nBA$, where *n* genes *X* of predicted or unknown function are found intercalated, but unrelated to the biosynthesis of histidine and tryptophan. In *Mycobacterium* and *Corynebacterium* the *hisI* gene is the last gene of the *his* operon whereas in *Streptomyces* this gene is the first one of the *trp* cluster. In general terms, the *Corynebacterium* HT cluster shows the highest number of deletions whereas the highest number of insertions is detected in *Streptomyces*. The *Mycobacterium* HT cluster stands half the way in this respect between *Corynebacterium* and *Streptomyces* and shows the highest number of *his* and *trp* genes per base pair.

In contrast, at least at a first glance, a HT cluster appears not to be conserved in *Bf. longum* and *Tb. fusca* and to have been deconstructed (or never constructed) during

course of evolution. These two cases deserve further investigation due to the lack of a *trpF* gene that has been reported in *Bf. longum* (Schell *et al.*, 2002) and *Tb. fusca* (see Section 4.2.2.2). Therefore, if the assumptions made in Section 4.1.4.1 are correct, one should expect to detect an HT cluster in these two organisms as a consequence of the constraints imposed by the putative bifunctionality of the orthologous PriA protein. This would be true for *Tb. fusca* whether the lack of a *trpF* gene in this organism is confirmed once its genome is fully sequenced. However, the *his* and *trp* genes that are part of the conserved HT cluster in other actinomycetes are spread in the genomes of *Bf. longum* and *Tb. fusca* in at least three different loci (Figure 4.3). Nevertheless, it is interesting to note that despite these marked differences in the physical position of the *his* and *trp* genes of *Bf. longum* and *Tb. fusca*, the *his* and *trp* operons appear to be conserved and, moreover, to show similarities to the organisation of the conserved HT cluster detected in *Mycobacterium*, *Streptomyces* and *Corynebacterium*. It seems therefore that in these organisms a physical association between the *his* and *trp* genes existed at one point in evolution (see also Section 4.2.3.5). In-depth discussions of the HT clusters shown in Figure 4.3 in a genus-by-genus basis are presented below.

4.2.3.1 The HT cluster of actinomycetes shows conservation of a higher order than operons

In *Streptomyces*, *Mycobacterium* and *Corynebacterium*, i.e. in the genera where a conserved HT cluster was discovered, the *his* and *trp* operons are separated between them by a few genes (from two to nine genes depending upon the organism) including consistently one gene that is transcribed in opposite direction to the rest of the cluster.

Therefore, the conserved HT cluster of actinomycetes seems not to be a single OSG but at least two discrete operons that might be co-regulated forming a superoperon, i.e. a complex regulatory gene array including more than one operon and several promoters. Furthermore, despite the fact that the intercalated *X* genes are not conserved other than in the cases discussed below (see Section 4.2.3.1.1), these genes are in most of the cases predicted to encode membrane-associated proteins of predicted or unknown function (the latter called *mpu* genes after membrane proteins of unknown function) which might have analogous functions in the different organisms. The significance of the co-clustering of these *mpu* genes with the *his* and *trp* genes is unknown but in some cases it may hint to unsuspected roles of these genes in the regulation or transport of histidine and tryptophan biosynthesis.

Another explanation for the co-clustering of genes encoding for membrane-associated proteins (including the *mpu* genes) with the *his* and *trp* genes is that the conserved HT clusters of actinomycetes are über-operons. As defined in Section 4.2.1.4 an über-operon is not necessarily conformed by genes that are functionally related. Therefore, it could be that this genomic organisation appears conserved because it would allow genes that are unrelated to the *his* and *trp* genes to benefit from the same levels and patterns of expression that are typical of these amino acid biosynthetic genes. The advantage to the *mpu* genes and other genes in clustering with *his* and *trp* genes may rely on the use of concerted regulatory mechanisms that control the expression of the *his* and *trp* genes such as regulation of gene expression in a growth-phase and growth-rate dependent fashion (Hu *et al.*, 1999; see also Section 1.3.1.1). This process has been previously described in the genomic context of the genes encoding for

ribosomal proteins and the term *genomic hitchhiking* has been used to define it (Rogozin *et al.*, 2002).

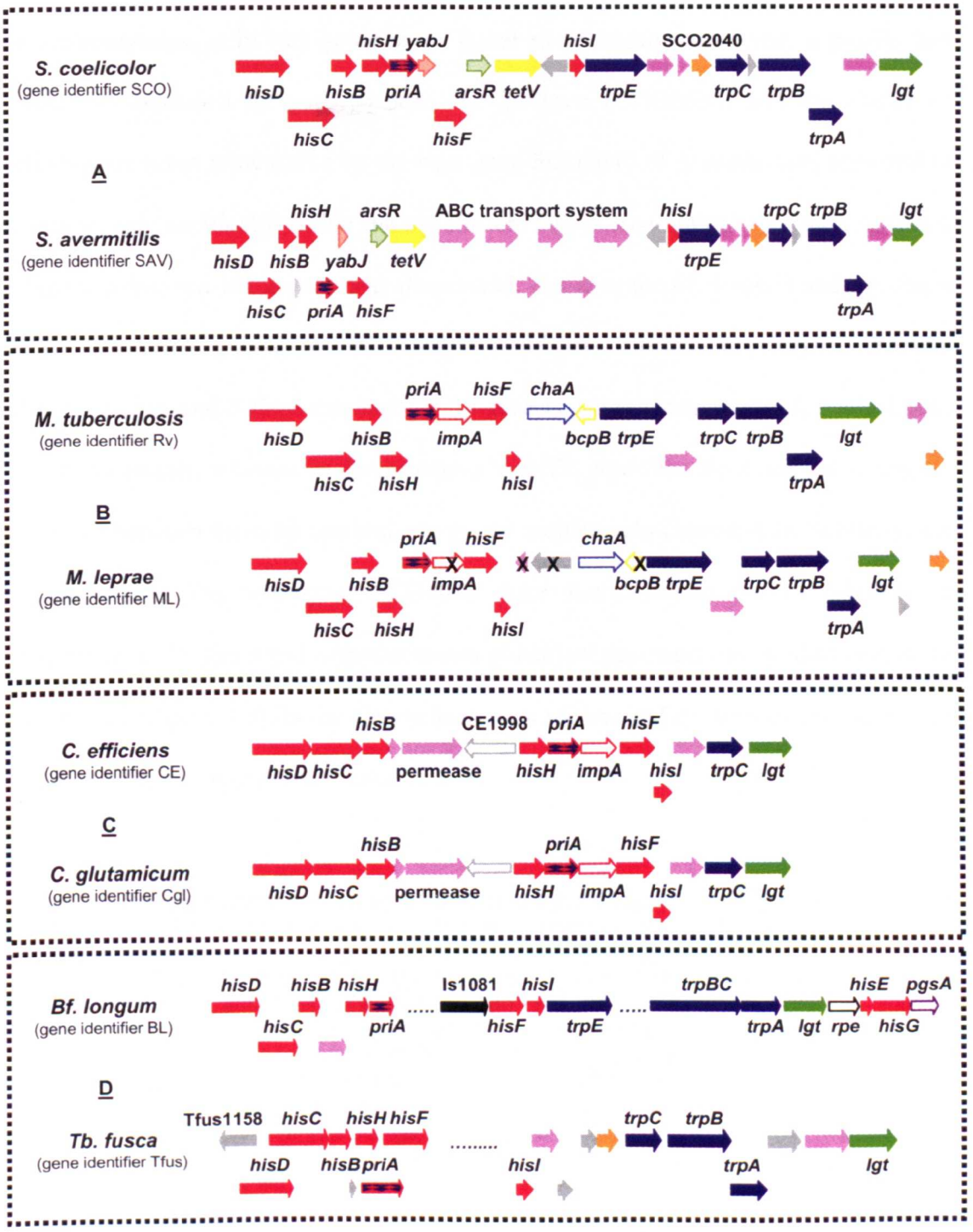


Figure 4.3 Genomic context analysis of *priA* (see legend in next page).

4.2.3.1.1 The *lgt* gene and orthologues of the *mpu* gene SCO2040 are conserved in the *trp* cluster of actinomycetes.

Of all the membrane-associated genes that are part of the conserved HT cluster of actinomycetes, only two genes were found to be highly conserved, a profile that cannot be explained by pure chance. The *lgt* gene (COG0682) and the cluster of orthologous genes represented by the *mpu* gene SCO2040 of *S. coelicolor*, appeared to be conserved specifically within the *trp* cluster. This is true not only for the organisms where a conserved HT cluster was discovered, but also for *Bf. longum* and *Tb. fusca* where at a first glance no HT cluster could be detected. In *Corynebacterium*, *Mycobacterium* and *Bifidobacterium* the *lgt* gene is just downstream *trpA*, the last gene of the *trp* cluster, whereas in *Streptomyces* and *Tb. fusca* the *trpA* and *lgt* genes are separated between them by one and two genes, respectively (Figure 4.3). Similarly, the orthologues of the *mpu* gene SCO2040 were found within the *trp* clusters of *Streptomyces*, *Tb. fusca* and *Mycobacterium* either just upstream *trpC* or downstream of the *lgt* gene (Figure 4.3). In the *Corynebacterium* species and *Bf. longum* orthologues of this gene were not found to be conserved.

Figure 4.3 Genomic context analysis of *priA* (see figure in previous page). Analysis of the neighbourhood of *priA* (red and blue) reveals that a conserved cluster of *his* (red) and *trp* (blue) genes exist in the actinomycetes. **A** HT cluster of *Streptomyces*. **B** HT cluster of *Mycobacterium*. **C** HT cluster of *Corynebacterium* and **D** *his* and *trp* operons of *Bf. longum* and *Tb. fusca*. The so-called *mpu* genes and other genes that encode for membrane-associated proteins, such as the ABC transport system of *S. avermitilis* and the permease of *Corynebacterium*, are shown in light magenta. Genes that are unique to the corresponding species or genus within the context of the HT cluster are shown in solid grey. The only two conserved genes of the actinomycetes HT cluster, other than the *his* and *trp* genes, are the *lgt* gene (bright green) and the orthologues of the *mpu* gene SCO2040 (bright orange). Other genes of predicted function that are discussed within the text are *yabJ* (light red), *arsR* (light green), *tetV* (yellow), *impA* (red no fill), *chaA* (blue no fill), *bcpB* (yellow no fill), CE1998 and its Cgl orthologue (grey no fill), Is1081 (black), *rpe* (black no fill), *pgsA* (dark magenta no fill). The genes that are predicted to be pseudogenes in *M. leprae* are marked with a cross. The physical separation of the different loci shown in section D is depicted by suspension points. See also the legend of Figure 2.3.

The *lgt* gene encodes for the enzyme prolipoprotein diacylglyceryl transferase (EC 2.4.99.-), which is well known to be involved in the maturation of lipoproteins in *E. coli* (Gan *et al.*, 1993; Sankaran and Wu, 1994) and *B. subtilis* (Leskelä *et al.*, 1999). The Lgt enzyme is responsible for the transfer of a diacylglyceryl moiety from a molecule of phosphatidylglycerol to a conserved cysteine residue of an apolipoprotein (Sankaran and Wu, 1994). In *E. coli* the *lgt* gene seems to be essential (Gan *et al.*, 1993) whereas in *B. subtilis* it has previously been shown to be dispensable for growth but essential for efficient protein secretion (Leskelä *et al.*, 1999). Interestingly, the *lgt* gene in *E. coli* forms a two-gene operon with the *thyA* gene whose product, the enzyme thymidylate synthase (EC 2.1.1.45), is involved in the synthesis of the nucleotide thymidine-5'-phosphate (Gan *et al.*, 195). Since it is difficult to imagine how an enzyme involved in the modification of apolipoproteins could be related to the biosynthesis of tryptophan, it is proposed that the genomic association of the *lgt* gene with the *trp* genes has an answer in the genomic hitchhiking phenomenon previously described (see Section 4.2.3.1).

The conserved *mpu* gene SCO2040 and its orthologues are more difficult to analyse. Other than in the actinomycetes mentioned above homologues of this gene were only found in the genome sequences of *Xanthomonas* species, i.e. *X. axonopodis* and *X. campestris* (da Silva *et al.*, 2002), and *Synechocystis* sp. PCC 6803 (Kaneko *et al.*, 1996). However, nothing is known about these genes in these organisms. Moreover, a genomic association of the SCO2040 homologues with the *trp* genes in these organisms was not found. Nevertheless, although it is acknowledged that at this stage it is impossible to draw any conclusions without uncertainty regarding the putative

function of the SCO2040 family of *mpu* orthologues, the observation of their conservation within the *trp* cluster of actinomycetes suggests that it might be functionally related to the biosynthesis of this amino acid in these organisms, probably at the level of regulation or transport. This observation is particularly valuable because when studying genes whose function is completely unknown it is desirable to have any sort of guide that would aid in defining an experimental approach, a situation that is achieved through this comparative analysis.

4.2.3.2 Investigation of the *Streptomyces* HT cluster.

The two strepromycetes compared show similar HT clusters in terms of the organisation of their *his* operons and *trp* clusters (Figure 4.3). Moreover, these OSG are identical, i.e. the *X* genes are conserved. The level of conservation of the *his* and *trp* genes between these streptomycetes supports the view introduced in Section 4.1.2 that the *trp* genes within the Sc4G6 fragment (and in its orthologous loci from other streptomycetes) are involved in primary metabolism. The lack of a *trpF* gene in *S. avermitilis* (see Section 4.2.2.2), in addition to the level of conservation of the *Streptomyces* HT cluster strongly suggest that the *hisA* homologue of *S. avermitilis* is in fact a *priA* gene. Furthermore, since in *S. coelicolor* the *trpE* paralogue specific for primary metabolism (i.e. the gene encoding the enzyme anthranilate synthase) still needs to be unambiguously characterised, this conservation suggests that *trpE3* (or SCO2043), and its orthologues from streptomycetes, is involved in primary metabolism (see also Chapter V).

The only difference between the HT clusters of the two streptomycetes analysed is a transport system composed by two ABC cassettes in tandem unique to *S. avermitilis* that is inserted between the *his* operon and the *trp* cluster. Although this transport system appears to have affinity for branched-chain amino acids as inferred from sequence similarity, there is no experimental evidence supporting this affinity. Therefore, it might be involved in transport of tryptophan or histidine as part of the system in *S. avermitilis*. Other than in this distinction the *S. coelicolor* and *S. avermitilis* HT clusters are identical. Interestingly, the three genes between *trpE* and *trpC* are *mpu* genes of which the last one (i.e. the pair SCO2040 and SAV6174) is conserved across some members of the actinomycetes (discussed in previous Section). Furthermore, the next two genes just downstream of *trpA* show similarity to *dsbG* of *E. coli* and to the *lgt* gene. Both of these genes are functionally associated with the membrane since either they are present in the periplasm (DsbG) or have a general role in the modification of membrane apolipoproteins (Lgt). The DsbG protein is a periplasmic disulfide isomerase with narrow substrate specificity (Bessette *et al.*, 1999) and the Lgt protein, as discussed above, is involved in maturation of apolipoproteins.

4.2.3.2.1 Regulatory deductions from the comparative analysis of the *Streptomyces* genomes

In *S. coelicolor* only three genes separate the *his* operon from the *trp* cluster, all of which are conserved in *S. avermitilis*. These include a homologue of the DNA-binding repressor *arsR* found in the arsenical resistance operon, which shows an N-terminal helix-turn-helix domain (SCO2047 and SAV6161); a homologue of the

tetracycline resistance determinant *tetV*, which is a integral membrane efflux protein (SCO2046 and SAV6162); and a gene of unknown function that is transcribed in opposite direction to *arsR* and *tetV* (SCO2045 and SAV6169). Another conserved gene that may be part of a putative regulatory network is the orthologous pair SCO2049 and SAV6159. These orthologues are annotated as genes of unknown function encoding for small proteins of around 14 kDa. However, a more detailed inspection of the domain composition of the predicted proteins using the web-server CDART (Conserved Domain Architecture Retrieval Tool, http://www.ncbi.nih.gov/BLAST/) revealed that these orthologues are homologous to the product of *yabJ* from *B. subtilis*.

The YabJ protein is a member of a highly conserved family of proteins that include YjgF from *E. coli*, YER057c from yeast and UK114 from goat (Sinha *et al.*, 1999 and references therein). The function of these proteins is not fully understood but they have been implicated in the regulation of several primary metabolic pathways (Goupil-Feuillerat *et al.*, 1997; Enos-Berlage *et al.*, 1998; Rappu *et al.*, 1999). The YabJ protein has been shown to be involved in the regulation of purine biosynthesis *via* an unidentified mechanism of which there is circumstantial evidence suggesting an interaction between YabJ and the adenine-feedback-regulated repressor PurR (Rappu *et al.*, 1999). These indications have recently been complemented by structural studies of YabJ. Unexpectedly, YabJ is structurally homologous to chorismate mutase (Sinha *et al.*, 1999), an enzyme that catalyses the arrangement of chorismate to perphenate at the branch chain point of the biosynthetic pathway for aromatic amino acids (see Section 5.1.1 and Figure 5.1). However, it is not clear whether the structural similarities detected between these two proteins reflect a functional link or simply that these proteins share a

common ancestry. The structure of YabJ also suggests that the biochemical role(s) of the members of this family of proteins would necessarily involve binding of a small-molecule metabolite (e.g. a substrate) or a macromolecule (e.g. a regulator) to a deep, wide conserved cleft at the subunit interface (Sinha *et al.*, 1999). Although these structural data do not clarify the mechanism of action of YabJ it suggests that its function might be in the regulation of the biosynthesis of small-molecule metabolites as opposed to a catalytic function.

The chromosomal position of SCO2049 and SAV6159 also brings some insights into the possible role(s) of these orthologues in *S. coelicolor* and *S. avermitilis*. In fact, this was the original observation that motivated a more detailed analysis of these orthologues. In the streptomycetes compared this gene pair is just downstream *priA* and appears to be translationally-coupled with *hisF* since the last codon of *yabJ* overlaps the initiation codon of *hisF* (Figure 4.3). Since translational coupling implies production of stoichiometric amounts of the two proteins, it is tempting to speculate that the streptomycete YabJ homologues are involved in the regulation of histidine and/or tryptophan biosynthesis and that they might play a crucial role in controlling the bifunctionality of PriA. This view is supported by the fact that *hisA* and *hisF* are very rarely found separated in bacterial genomes (Alifano *et al.*, 1996; Fani *et al.*, 1998) probably as a consequence of their functional associations (Lang *et al.*, 2000; see also next Section). Hence, the chromosomal positioning of SCO2049 and SAV6159 between *priA* and *hisF* could reflect a strong functional constraint. However, it is puzzling that this same position is occupied in the HT clusters of *Mycobacterium* and

Corynebacterium species by an *impA* gene whose product is predicted to be involved in the biosynthesis of *myo*-inositol (see Sections 4.2.3.3 and 4.2.3.4).

4.2.3.2.1.1 Mutational analysis of SCO2049

During the construction of the auxotrophic strain WH101 (see Section 2.2.3 and Figure 2.5) an error occurred while designing the mutagenesis oligonucleotides. This led to a PCR product that, when used for mutagenesis of cosmid Sc4G6 meant the *priA* gene plus the 5' end of SCO2049 were replaced with the *aac(3)IV* plus *oriT* disruption cassette of pIJ773 (see also Section 7.1.3.2 in Methods and Appendix 2). This mistake was employed profitably by transferring the resulting cosmid into *S. coelicolor* M145 by conjugation with *E. coli* ET12567/pUZ8002. As expected, colonies that had undergone double crossovers as inferred from kanamycin sensitivity proved to be tryptophan and histidine auxotrophs. The resulting strain was called WH102 and presented the opportunity to look at the phenotype associated with SCO2049. Determination of the growth curves of strains WH101 and WH102 in YEME medium (Kieser *et al.*, 2000), i.e. with histidine and tryptophan present, showed that the double mutant WH102 was significantly impaired in growth when compared to WH101.

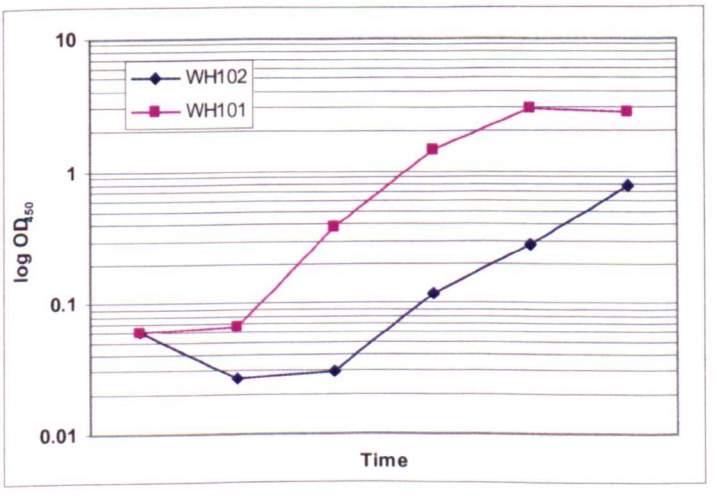


Figure 4.4 Growth curves of strains WH101 and WH102 in YEME. The growth of the double mutant WH102 (blue marks) is significantly impaired compared to WH101 (cyano marks).

Although the meaning of this result cannot be assessed without further experimentation, this result is circumstantial evidence that SCO2049 is an important gene for the growth of *S. coelicolor*, at least within a *priA* minus background. Interestingly, it is known that histidine biosynthesis is connected to both nitrogen metabolism and the *de novo* synthesis of purine nucleotides at the step where PRFAR and glutamine are converted into the products ImGP and 5'-(5-aminoimidazole-4-carboxamide) ribonucleotide AICAR *via* the activity of the enzyme ImGP synthase (Alifano *et al.*, 1996; see also Figure 4.6). This bifunctional enzyme consists of a cyclase moiety encoded by *hisF* and an aminotransferase subunit encoded by *hisH* (Klemm and Davisson, 1993). The resulting ImGP is an intermediary in the biosynthesis of histidine, while AICAR is used in the *de novo* biosynthesis of purine nucleotides (Alifano *et al.*, 1996). Therefore, it is tempting to speculate that the phenotype detected in WH102 could be related to a toxic effect derived from a metabolic imbalance at the interconnectivity that is expected to occur at the enzyme ImGP synthase. When considering the putative regulatory role of SCO2049 it may be that the pathways converging at this step are deregulated when this protein is absent.

4.2.3.3 Investigation of the *Mycobacterium* HT cluster

The HT clusters of *M. tuberculosis* and *M. leprae* appear to be identical, albeit the fact that some of the orthologous genes present in *M. tuberculosis* are predicted to be pseudogenes in *M. leprae* (Figure 4.3). In *Mycobacterium* the *hisI* gene is part of the *his* operon as inferred from the observation that this histidine biosynthetic gene is potentially translationally coupled with *hisF*. This contrasts with the streptomycetes and

Tb. fusca where *hisI* is separated from the *his* operon by insertion of other genes. In spite of the differences between the HT clusters of *Mycobacterium* and *Streptomyces*, their *his* operons and *trp* clusters appear to be similar and overall these lineages show highly conserved HT clusters (Figure 4.3). Since *Streptomyces* and *Mycobacterium* had diverged long enough back in evolution to expect gene order not to be conserved (Figure 4.1) similar functional constraints could be driving gene conservation in both lineages. Therefore, the failure to detect a *trpF* gene in the *Mycobacterium* lineage (see Section 4.2.2.1) plus the high degree of conservation between the *Mycobacterium* and *Streptomyces* HT clusters strongly hints at the existence of a *priA* gene in the *Mycobacterium* complex.

4.2.3.3.1 *Mycobacterium tuberculosis* has a *priA* gene as revealed by intergeneric complementation studies

The hypothesis of the existence of a *priA* gene in *Mycobacterium* was tested through intergeneric complementation studies. For this purpose the *M. tuberculosis* gene Rv1603, which is annotated as a *hisA* gene within the *M. tuberculosis his* operon (Cole *et al.*, 1998), was cloned into the expression vector pGEX-4T-1. The cloning strategy was identical to that followed in Section 2.2.4 for cloning of *priA*, *trpF* and *hisA* (see Table 7.1 in Methods). The construct obtained, named pGEXpriAMt, contains the putative *priA* orthologue from *M. tuberculosis* under the control of the IPTG inducible *lacI*-dependent promoter. Thus, expression of the insert from this construct is independent of the genetic background and therefore suitable for this type of experiments. The pGEXpriAMt plasmid was used to transform *E. coli* auxotrophs with

mutations in *trpF*, W3110 *trpC*(*F*del) (Darimont *et al.*, 1998), and *hisA*, Hfr G6 (Matney *et al.*, 1964). The resulting transformants were tested for tryptophan and histidine auxotrophy in modified VB and MA media, respectively (see Section 7.2 in Methods). Complementation of both defects was achieved after incubation overnight. However, in contrast to the results obtained with the *S. coelicolor priA* gene (see Section 2.2.4) it appeared that complementation with the Rv1603 required induction of expression with the appropriate concentration of IPTG. For some unknown reason complementation was only detected when very low levels of IPTG was used, but not when no IPTG was added or its concentration surpassed 0.1 mM. The optimal concentration of IPTG for expression of Rv1603 and subsequent complementation of *trpF* and *hisA* defects was around 0.01 mM. This result demonstrates that a PriA enzyme exists in *M. tuberculosis* implying that the appearance of PriA may have occurred in ancestral times. The results presented in this section have been published elsewhere (see Appendix 1; Barona-Gómez and Hodgson, 2003).

4.2.3.3.2 The *Mycobacterium* HT cluster is tightly organised

The *Mycobacterium* HT cluster shows the greatest density of *his* and *trp* genes of all organisms compared. Of all the genes present in the HT genomic region of *Mycobacterium* only four of them are not annotated as *his* or *trp* genes. The first gene, just downstream *hisI* and transcribed in the same direction is a *chaA* homologue. The ChaA protein in *E. coli* is a Ca^{++}/H^{+} antiporter involved in extrusion of inorganic ions through the membrane at alkaline pH (Ohyama *et al.*, 1994). The *chaA* gene must be important since in *M. leprae* it appears to be functional. The second gene, just upstream

trpE but in opposite direction to the rest of the genes in the cluster is a homologue of *bcpB*. This gene encodes for the protein BCP (from bacterioferritin co-migratory protein) and in *E. coli* has previously been shown to act as a general hydroperoxide peroxidase in a thioredoxin-dependent manner (Jeong *et al.*, 2000). However, this orthologue appears to be a pseudogene in *M. leprae* suggesting that it is dispensable (Figure 4.3). The third gene is an *mpu* gene that is just downstream *trpE* and potentially translationally coupled to it. It is likely therefore that this orthologue (Rv1610 and ML1270) is related to the biosynthesis of tryptophan in *Mycobacterium*. At this stage it is difficult to imagine how these three genes could be related to the biosynthesis of tryptophan and histidine. Moreover, since none of these genes is conserved in any other actinomycetes HT cluster their association with the *his* and *trp* genes cannot be explained in terms of the genomic hitchhiking phenomenon described in Section 4.2.3.1.

As discussed in Section 4.2.3.2.1 a puzzling observation of the *Mycobacterium* HT cluster is the presence of an *impA* gene between *priA* and *hisF*, the fourth gene of unrelated function to histidine and tryptophan biosynthesis. This observation contrasts with the scenario in *Streptomyces* where the position is occupied by a *yabJ* homologue. Although this disagreement poses some doubts on the potential regulatory role proposed for SCO2049 (see Sections 4.2.3.2.1 and 4.2.3.2.1.1) the *Mycobacterium* genomes analysed also have a *yabJ*-like gene somewhere else on the chromosome that could be performing a similar function. The product of *impA* (Rv1604) in *M. tuberculosis* is predicted to be an inositol monophosphatase (EC 3.1.3.25), a key enzyme in the biosynthesis of free *myo*-inositol from inositol-1-phosphate (Parish *et al.*, 1997; Nigou *et al.*, 2002). Inositol plays a crucial role in the physiology of *Mycobacterium* species

since it acts as precursor for synthesis of several cell wall constituents unique to this genus such as phosphatidylinositol mannosides (PIMs), some of which have been implicated in immunopathogenesis of tuberculosis (Parish *et* al., 1997; Jackson *et al.*, 2000; Nigou *et al.*, 2002).

The importance of the activity of the inositol monophosphatase enzyme for the physiology of *Mycobacterium* is supported by the redundancy of *impA* paralogues (COG0483) observed within *M. tuberculosis*. In this organism there are at least four genes that potentially encode inositol monophosphatases. One of these paralogues, the *suhB* gene (Rv2701c), has recently been characterised *in vitro* (Nigou *et al.*, 2002). These researchers reported that the product of the *suhB* gene preferentially hydrolyses inositol-1-phosphate as substrate, suggesting that SuhB might be the main inositol monophosphatase in *M. tuberculosis*. However, previous observations in *M. smegmatis* by Parish *et al.* (1997) suggested that the orthologue of Rv1604 in this organism (as judged from its homologous position between the *hisA* and *hisF* genes) should be also involved in the synthesis of inositol since mutation of this gene resulted in an altered cell envelope and a reduced level of synthesis of a particular type of PIM. Nevertheless, whether there is a functional link between the products of *impA*, *priA*, *hisF* and *hisH* genes, it is difficult to envisage a metabolic model accounting for such scenario and therefore further experimentation is required. The fact that *impA* is also conserved in the *Corynebacterium* HT cluster suggests that a functional constraint should exist and therefore hints towards a functional correlation between these genes. Another possibility is that the *impA* gene is in fact a new class of $hisB_p$ gene, a gene that remains to be identified in actinomycetes. This suggestion is supported by the fact that ImpA and

$HisB_p$ perform analogous dephosphorylation reactions upon similar substrates (both enzymes share the first three EC numbers 3.1.3.-). However, the *M. smegmatis impA* mutant strain constructed by Parish *et al.* (1997) does not support this view since this strain did not turn out to be a histidine auxotroph, although endogenous complementation of a phosphatase activity is likely to occur.

4.2.3.4 Investigation of the *Corynebacterium* HT cluster

The HT cluster of the *Corynebacterium* genus appears to have suffered major deletions of its *trp* genes that otherwise would be part of the *trp* cluster. In contrast, the *his* operon has been split into two putative operons after insertion of three genes between $hisB_d$ and *hisH* (Figure 4.3). The only *trp* gene that still remains within the *Corynebacterium* HT cluster is the *trpC* gene whereas all the *his* genes are conserved. Interestingly, and despite the fact that the *trpE* and *trpBA* genes have been lost, suggesting that major deletions occurred in this region, the next gene just downstream *trpC* is the *lgt* gene. This level of conservation is unexpected and, as discussed previously, it supports the view that in actinomycetes for some as yet uncharacterised phenomenon the *lgt* gene is physiologically associated with the biosynthesis of tryptophan (see also Section 4.2.3.1)

At the other end of the cluster the genes that appear in the middle of the *his* operon are a *mpu* gene unique to *Corynebacterium* (Cgl2018 and CE2000); a gene predicted to encode a permease of the major facilitator superfamily (Cgl2017 and CE1999); and a gene of unknown function highly conserved across actinomycetes that is transcribed in opposite direction. The annotators of the *C. efficiens* genome succeeded in

identifying the latter gene (CE1998) whereas in *C. glutamicum* this orthologue was missed. However, close inspection of the *C. glutamicum* sequence allowed the detection of the corresponding orthologue in this organism (Figure 4.3). Finally, as discussed above, an *impA* gene is conserved in *Corynebacterium* and *Mycobacterium* supporting the view that this gene is functionally related to the *his* and *trp* genes. Overall the HT clusters of the *Corynebacterium* species analysed here are identical although they show less similarity with the corresponding HT clusters of *Streptomyces* and *Mycobacterium*.

The above observations raises the question of why the *Corynebacterium* HT cluster was deconstructed during the course of evolution? As mentioned in Section 4.2.2.3 and references therein, the corynebacteria have an enteric-like *trp* operon elsewhere on the chromosome that seems to have been acquired by lateral gene transfer (Crawford, 1989). Moreover, it is known that *C. glutamicum* synthesises tryptophan *via* the products of this *trp* operon (Matsui *et al.*, 1986). It seems that the *Corynebacterium* ancestor was subject to selection for a feedback-regulated tryptophan operon. Once *Corynebacterium* underwent this lateral gene transfer, the HT cluster as it stands in other actinomycetes was freed from its original constraints allowing major deletions and insertions to occur. Jung *et al.* (1998) cloned by complementation of *E. coli* mutants the *hisA*, *impA* and *hisF* genes of *C. glutamicum* before the genome sequence of this organism became available. These researchers were not aware of the potential PRA isomerase activity of this enzyme and nothing was reported in this respect. It would be very interesting to determine the TrpF activity of the corynebacteria HisA homologues.

4.2.3.5 *Thermobifida fusca* and *Bifidobacterium longum* show aspects of an actinomycetes HT cluster

The lack of an HT cluster in *Tb. fusca* and *Bf. longum* could be explained in two different ways. It could be that in these lineages the HT cluster was never formed, suggesting that the functional constraints that drove the clustering of the *his* and *trp* genes in other actinomycetes, i.e. the existence of a PriA enzyme, did not exist in these organisms. Alternatively, it could be that at one point in evolution the HT cluster did exist in these lineages and for varying reasons such as the acquisition of other genes (for example, the *trp* enteric-like operon of *Corynebacterium*) it has been deconstructed to a certain degree. Our analysis favors the latter interpretation as supported by two lines of evidence. First, the lack of a *trpF* gene in these organisms (see Section 4.2.2.2) is a strong suggestion that the functional constraints assumed to be driving the clustering of the *his* and *trp* genes in other actinomycetes are also present in *Bf. longum* and *Tb. fusca*. Secondly, as introduced in Section 4.2.3, a closer inspection of the three loci including the *his* and *trp* genes in these organisms show 'reminiscences' of the existence of an ancestral HT cluster that has been partially deconstructed.

This situation is particularly noticeable in *Bf. longum* where the majority of the *his* genes are present within a single locus showing the same gene order found in other actinomycetes. However, this operon finishes with a *hisA* gene and strikingly lacks the *hisF* gene (Figure 4.3). It is very rare for the *hisA* and *hisF* genes not to be next to each other (Alifano *et al.*, 1996; Fani *et al.*, 1998). Moreover, in the cases where these genes are not adjacent, there is typically only one gene between them (e.g. the organisation of *priA*, SCO2049 and *hisF* in *S. coelicolor*). In contrast, the *hisF* gene of *Bf. longum* lays

in a second locus that remarkably includes the *hisI* and *trpE* genes forming what it appears to be an operon similar to the central part of the conserved HT cluster of actinomycetes. Thus, a question that needs to be answered is how were the *hisA* and *hisF* genes separated?

After a detailed inspection of the sequence upstream of *hisF*, I discovered a previously unnoticed ORF encoding a transposase of the mutator family homologous to Is*1081* from *M. bovis* (Figure 4.3). This discovery explains why the *hisA* and *hisF* genes are present in distant loci and suggests that an HT cluster may have existed in this organism until relatively very recent. The remaining *trp* genes lay in a third locus with unexpected features. Firstly, the products of *trpB* and *trpC* are fused within a single polypeptide. As far as we are aware this type of fusion has never been previously reported. Secondly, these *trp* genes appear to form an operon that includes the *lgt* gene; a gene similar to *rpe* putatively encoding for the enzyme ribulose-5-phosphate-3-epimerase; the *hisE* and *hisG* genes; and the *pgsA* gene, which encodes for the membrane-associated enzyme phosphatidylglycerol phosphate synthase, which catalyses the committed step in the synthesis of acidic phospholipids (EC 2.7.8.7; Gopalakrishnan *et al.*, 1986). Overall these observations support the conclusion that the *Bf. longum his* and *trp* genes used to form an HT cluster which unexpectedly also included the *hisE* and *hisG* genes at one point during the course of evolution. Furthermore, this observation strongly suggests that in *Bf. longum* a non-orthologous gene displacement similar to the *S. coelicolor* and *M. tuberculosis* displacement event had taken place in this early diverged organism, the implication being that PriA is a molecular fossil.

The *Tb. fusca* case is not that informative in this respect. The only similarity to an HT cluster that could be detected is the co-clustering of the *trpCBA* genes with the *hisI* gene. There is no evidence of a transposase to account for the deconstructed HT cluster. Nevertheless, it is worthwhile to note that the *lgt* gene and the *mpu* gene that we observed as conserved in the HT clusters of other actinomycetes (see Section 4.2.3.1.1) are also conserved in the *trp* cluster of *Tb. fusca*. Another interesting feature of this organism is the overlapping organisation of the *his* genes (Figure 4.3). The first gene of the *Tb. fusca his* operon, *hisD*, overlaps with Tfus1158 upstream and with *hisC* downstream. Furthermore, the coding sequence of the *hisH* gene overlaps almost entirely with *hisA*, which in turn overlaps with the 5' end of *hisF*. The meaning of this tight genomic organisation is unknown but it certainly has an effect on the coding density. However, all of the overlapping genes are in different frames ruling out the possibility of domain sharing, a tempting possibility that would be in accordance with the internal duplication observed in *hisA* and *hisF* (Fani *et al.*, 1994; Lang *et al.*, 2000; see also Chapter III).

4.2.4 Phylogenetic analysis of PriA and HisA proteins

An unanswered question is how widespread is PriA beyond the actinomycetes clade and across the bacterial complement. This question was addressed by constructing a phylogenetic tree making use of the multisequence alignment shown in Figure 3.2. Overall the tree shows phylogenetic coherence (Figure 4.5). Furthermore, it is confirmed that the actinomycetes PriA proteins cluster together forming an evolutionary discrete group. However, it is interesting to note that the *Corynebacterium* proteins appear to

have diverged significantly from the rest of the sequences as judged from their position in an outer branch within the actinomycetes lineage. This observation suggests that it is possible that the HisA homologue in *Corynebacterium* might have lost already its PRA isomerase activity as a consequence of the presence of a *trpF* gene elsewhere on the chromosome (see Section 4.2.2.3).

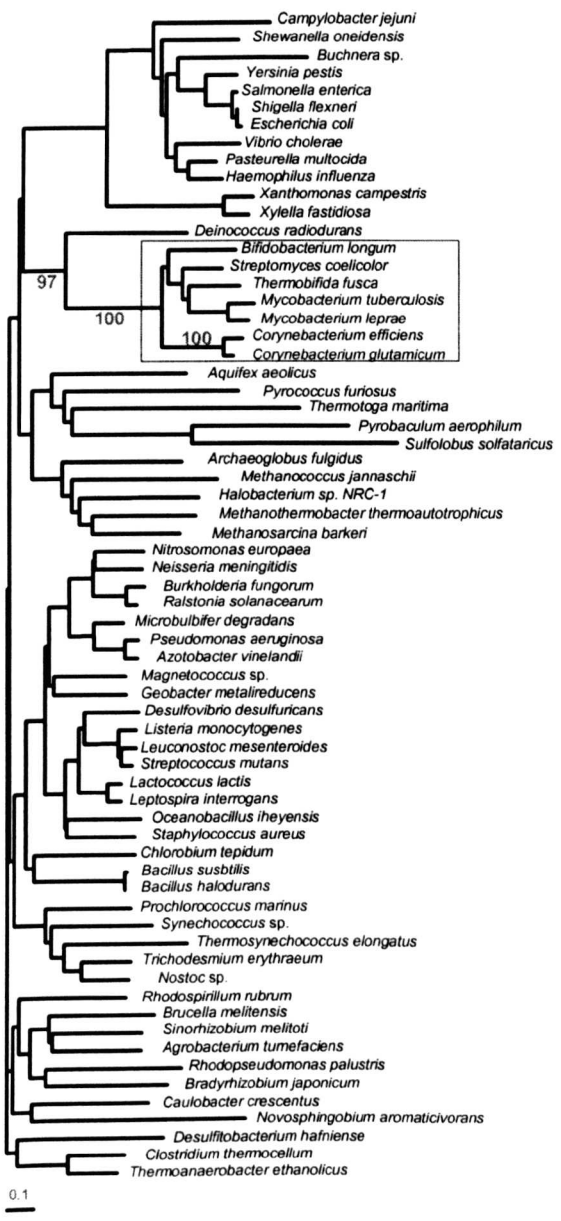


Figure 4.5. Phylogenetic tree of HisA and PriA proteins. The tree was constructed using the neighbour-joining method. The bootstrap values for the the actinomycetes clade (shown within a blue box) are provided in the relevant branching points. As it can be seen the *Corynebacterium* proteins have diverged significantly. The HisA homologue from *D. radiodurans* may also cluster with the actinomycetes. Refer to text for further explanation.

As judged from the tree shown in Figure 4.5, it may appear that the HisA protein from *Deinococcus radiodurans* is related to the PriA enzymes from actinomycetes, suggesting that this novel protein may be spread beyond the actinomycetes. However, despite the fact that the *D. radiodurans* HisA protein may seem to cluster with the actinomycetes sequences, some considerations should be taken into account. For instance, *D. radiodurans* is known to have a high (G + C) content of around 65% (White *et al.*, 1999). This could cause a GC bias in the codon usage making proteins to appear more similar than what they really are. Additionally, the *D. radiodurans* HisA enzyme is the only sequence included in the analysis from this bacterial lineage. It is likely that if more sequences from relatives of *Deinococcus* are included the topology of the tree would be different. On the basis of the data available it is only safe to conclude that PriA has spread across the actinomycetes.

4.2.5 Co-evolution of histidine and tryptophan biosynthetic enzymes

Back in the early 1970s, a common origin for certain histidine and aromatic amino-acid biosynthetic enzymes was proposed (cited in the review by Jensen and Gu, 1996; see also Section 2.2.5), such that an evolutionary relationship was established between glutamine aminotransferases involved in the biosynthesis of histidine and the aromatic amino-acids phenylalanine and tyrosine. In 1976, Jensen proposed a common origin for certain histidine and tryptophan biosynthetic enzymes along with his patchwork hypothesis on enzyme and pathway evolution (Jensen, 1976). Jensen's proposal was based on the similarities found between certain Trp and His enzymes at the level of mechanism of reaction involving analogous substrates. The relationships

proposed included the enzymes PRA isomerase (*trpF*) and ProFAR isomerase (*hisA*), the phosphoribosyl transferase enzymes encoded by *trpD* and *hisG*, and the aminotransferase component of anthranilate synthase (*trpG*) and the product of *hisH*.

With the advent of more sensitive methods for the comparison of sequences and the increase in the number of 3D structures reported some of these relationships have been corroborated and further expanded. In addition to PRA isomerase and ProFAR isomerase extensively discussed in Chapters II and III, it has recently been demonstrated that the structures of ImGP synthase, encoded by *hisF* and *hisH* (Chaudhuri *et al.*, 2001; Douangamath *et al.*, 2002), and anthranilate synthase, encoded by *trpE* and *trpG* (Knochel *et al.*, 1999; Morollo and Ech, 2001; Spraggon *et al.*, 2001), are highly conserved suggesting a common evolutionary origin. In addition, the structure of HisF shows great similarity with the structure of TrpC (i.e. IGP synthase) suggesting a common origin (Lang *et al.*, 2000; Gerlt and Babbitt, 2001a and b). Furthermore, a common ancestry has also been suggested for the phosphoribosyl transferase enzymes involved in the biosynthesis of tryptophan and histidine, i.e. PRA transferase (*trpD*) and ATP phosphoribosyl transferase (*hisG*), respectively (Mayans *et al.*, 2002). However, the structure of HisG still needs to be elucidated.

How does this wealth of evolutionary evidence fit with our results? One tempting interpretation of the discovery of a conserved HT cluster in actinomycetes, with *priA* at the core of this complex gene array, is that the majority of the His and Trp enzymes diverged from a set of bifunctional common ancestral genes. This possibility is exemplified by the existence of the bifunctional PriA enzyme. As discussed in Chapter II, it is likely that a TrpF protein could evolve from PriA. If this scenario is retro

projected in evolution, and the evolutionary evidence discussed above supporting a common ancestry for several other His and Trp enzymes is taken onboard, it seems plausible that the majority of the His and Trp enzymes diverged from ancestral proteins with broad-substrate specificities. It has recently been proposed that ancestral enzymes adopting the TIM-barrel fold, such as those involved in amino acid biosynthesis, "...thanks to its intrinsic binding pocket, provided a protein scaffold that lowered the activation barriers of functions by concentrating the reactants through low-specificity binding" (Anantharaman *et al.*, 2003). These enzymes would not have been Trp nor His enzymes, although more sequence conservation with the latter is expected, but an ancestral protein capable of performing the two reactions required within each pathway. More specifically, it is proposed that in the ancestral cell only one set of His-Trp proteins existed that were responsible for the synthesis of both IGP (from glutamate and chorismate) and ImGP (from PRPP and ATP) for tryptophan and histidine biosynthesis, respectively (see Figure 4.6). The remaining Trp enzymes, e.g. the enzyme tryptophan synthase (*trpBA*), could have arisen from enzymes taking part in other biosynthetic pathways such as threonine biosynthesis (Parsot, 1987).

The evolutionary model proposed, accounting for the co-evolution of the first enzymatic steps of the biosynthetic pathways of tryptophan and histidine, is in accordance with the patchwork hypothesis of the evolution of metabolic pathways *via* the recruitment of enzyme function (Jensen, 1976). Although the scenario depicted here has been previously implied, our data is compelling evidence that the His and Trp enzymes involved in the early steps of these pathways co-evolved. The suggestion that an HT cluster in the early diverged actinomycete *Bf. longum* may had included at some

point a *hisG* gene (see Section 4.2.3.5) but not a *trpD* gene supports this scenario. Analogous to the clustering of *his* and *trp* genes driven by *priA*, this ancestral-like HT cluster may have been formed as a consequence of the functional constraint imposed by the bifunctionality of the ancestral product of *hisG*. Although *Bf. longum* has a *trpD* gene elsewhere on the chromosome, as seen in other actinomycetes, this gene could be a recent acquisition or the result of a different convergent evolutionary trajectory involving the class II nucleoside phosphorylases (Sinha and Smith, 2001; Mayans *et al.*, 2002). Hence, it seems plausible that the ancestor of HisG used to be a bifunctional enzyme with dual specificity for phosphoribosyl ATP and PRA explaining the co-clustering of *hisG* with the *trpCBA* cluster of *Bf. longum*. In a similar vein, one constant of the conserved HT cluster of actinomycetes is the lack of a *trpG* gene. This observation contrasts with the invariant conservation of a *hisH* gene, suggesting that both biosynthetic pathways may share the aminotransferase activity of HisH.

One apparent flaw of the model shown in Figure 4.6 is the lack of accountability for the enzyme functions encoded by the *hisI* gene, which are involved in the early steps of ImGP biosynthesis. In some organisms, such as enteric bacteria, the *hisI* gene encodes a bifunctional enzyme with phosphoribosyl ATP (PR-ATP) pyrophosphohydrolase (C-terminus) and phosphoribosyl AMP (PR-AMP) cyclohydrolase activities (N-terminus) (reviewed by Alifano *et al.*, 1996). These enzymes essentiality facilitate the hydrolysis of PR-ATP and PR-AMP, respectively, reactions that may occur at environmental conditions without enzymatic catalysis after a period of time. The implication of such scenario is that the enzyme activities encoded by *hisI* may have been dispensable for the synthesis of ImGP within the ancestral cell.

Therefore, the enzymes encoded by *hisI* may not have been recruited until relatively recently as suggested by the patchwork hypothesis. Unlike the enteric bacteria the actinomycetes *hisI* gene includes only the PR-AMP cyclohydrolase enzyme function while an independent *hisE* gene encodes the enzyme PR-ATP pyrophosphohydrolase. It is interesting to note that the *hisI* gene in actinomycetes is conserved within the HT cluster. This level of conservation suggests that there may be a functional constraint driving the co-clustering of the *hisI* and the *trp* genes. The meaning of this observation is unknown but it may reflect a hitherto unidentified evolutionary relationship between *hisI* (encoding PR-AMP cyclohydrolase only) and a *trp* gene. However, as discussed above, the HisI protein could be unrelated to the Trp enzymes without being at odds with the evolutionary model proposed in this Section.

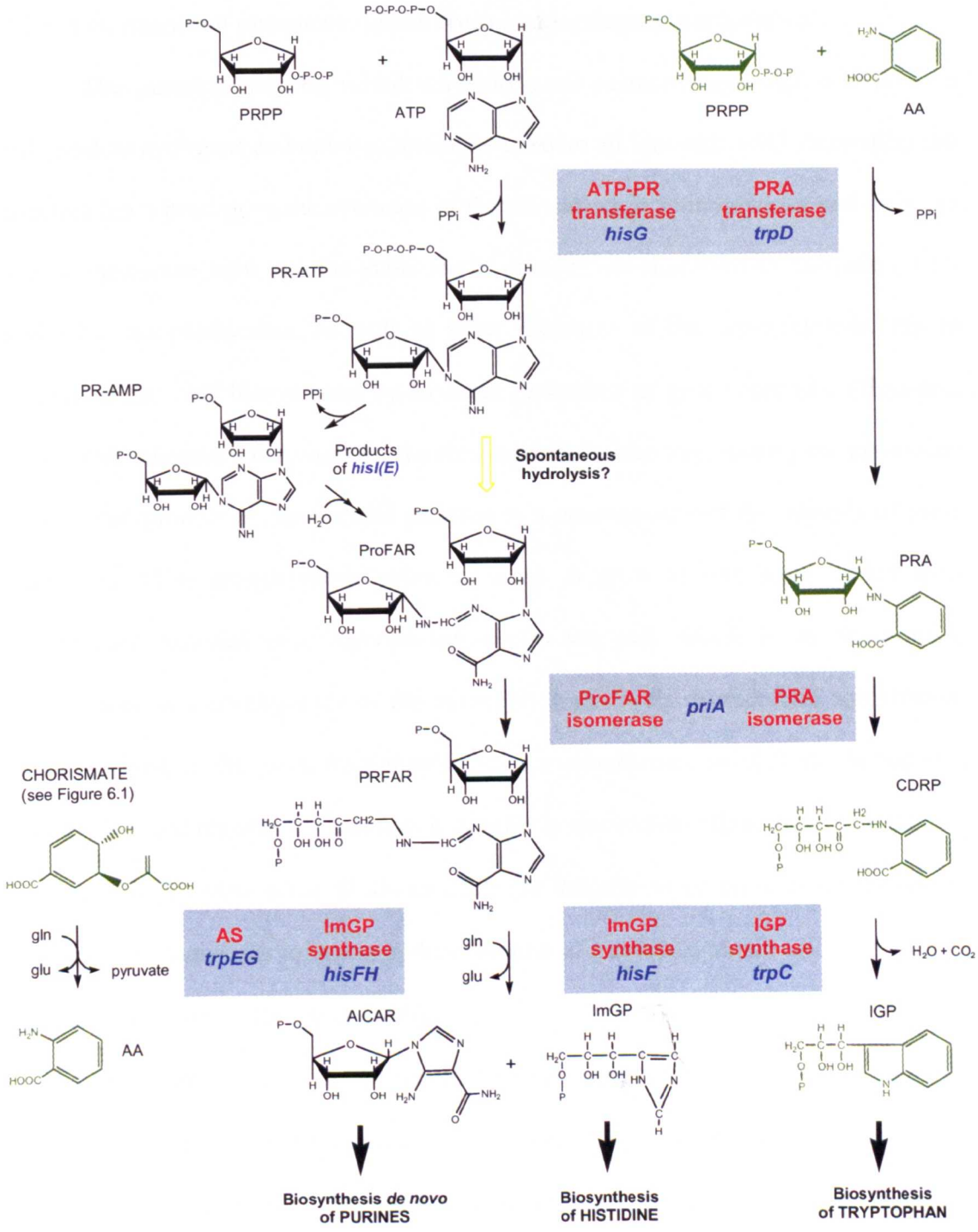


Figure 4.6 Co-evolution of histidine and tryptophan biosynthesis in actinomycetes. The model predicts that the early steps of the histidine (black structures) and tryptophan (green structures) biosynthetic pathways leading to ImGP (and AICAR) and IGP, respectively, were achieved by bifunctional His-Trp enzymes with broad-substrate specificity, such as PriA, present in the ancestral cell. The evolutionary relationships between His and Trp enzymes are highlighted by blue boxes. Within this model the steps catalysed by the products of *hisI(E)* are proposed to had undergone spontaneously (yellow arrow). Refer to text for further explanation of the pathways metabolites, genes and enzymes.

4.2.6 A comment on operon evolution from a physiological perspective

The question may be raised why did some actinomycetes fail to evolve an independent *trpF* gene and opted to retain the putative ancient-like *priA*? Answering this question has a bearing on the evolution of the HT cluster in actinomycetes and therefore it may shed some light into the evolution of operons. As discussed in Sections 1.3.1.1 and 2.2.5, streptomycetes, as well as other members of the actinomycetes, fail to regulate amino acid biosynthesis by feedback repression of gene expression (Hodgson, 2000). Our laboratory believes that the reason of the failure to regulate gene expression by the final product of the cognate pathway is a consequence of the lifestyle of these organisms. Most actinomycetes grow, or used to grow at one point during their evolutionary histories (e.g. *Mycobacterium*), in the soil, which is an oligotrophic environment as a consequence of the extremely high levels of microbial competition. Soil is derived, in the main, from plants and as a consequence is relatively carbon rich and nitrogen and phosphate poor, thus it is unlikely that soil-dwelling bacteria will come in contact with amino acids. If amino acids are unlikely to be present, it may not be worth for actinomycetes to invest in the evolution of individual amino acid biosynthetic gene control systems (Hodgson, 2000).

The traditional view on the *raison d'être* of operons is that they have evolved as a strategy for co-ordinated regulation of gene expression by feedback repression and/or attenuation (see Sections 1.1.5.1 and 4.1.1). Interestingly, autotrophic organisms that are very unlikely to come in contact with amino acids very rarely contain operons of amino acid biosynthetic genes. For example, the *trp* genes in *Synechocystis* sp. strain PCC6803 are fully dispersed throughout the chromosome (Kaneko *et al.*, 1996). If actinomycetes

do not regulate amino acid biosynthetic genes at the gene level it may appear therefore that there is no pressure to cluster genes into operons. If the expression of amino acid biosynthetic pathways is not inactivated by the cognate amino acid, and a physiological regime where constitutive expression of amino acid biosynthetic genes seems to be the rule, expression of enzymes that have roles in more than one biosynthetic pathway, such as PriA, would always be available. Therefore, from a physiological point of view there is no obvious reason for genes to cluster in actinomycetes. The situation is different in copiotrophs associated with animals such as enteric bacteria. Amino acids are likely to be supplied in the gut on an irregular basis and in varying concentrations and therefore it is worthwhile for these organisms investing in regulation of individual amino acid biosynthetic genes. Thus, what is the constraint driving the clustering of the *his* and *trp* genes in actinomycetes?

The selfish operon hypothesis may aid in solving this conundrum (Lawrence and Roth, 1996; Lawrence, 1999). This model postulates that the organization of bacterial genes into operons is beneficial to the constituent genes in that proximity allows horizontal co-transfer of all the genes required for a selectable trait. In other words, this hypothesis assumes that only sets of genes that code for all the proteins required for a particular metabolic or cellular function will be fixed after they have been horizontally transferred. Thus, since *priA* and probably other genes (see above) are involved in the biosynthesis of both histidine and tryptophan, it could be that the clustering of the *his* and *trp* genes observed in actinomycetes brings the putative benefits postulated by this hypothesis. However, an observation arguing against this possibility is the fact that not all the enzyme functions required for the synthesis of tryptophan and histidine (e.g. the

TrpD and HisG proteins) are present, or appear to have been present in *Bf. longum*, within the HT cluster.

4.3 Conclusions

A comprehensive comparative genomic context analysis was presented in this Chapter. The availability of modern biocomputational tools and the large collection of bacterial genome-sequences were exploited in order to understand the evolution of the *his* and *trp* genes in the Gram-positives with a high (G + C) content. It was shown that a cluster of genes including the majority of the *his* and *trp* genes, the so-called HT cluster, is highly conserved across the actinomycetes, and may have been present in the last common ancestor of this group of bacteria. A model by which the early enzymes of the tryptophan and histidine biosynthetic pathways co-evolved is proposed. This analysis also led to the hypothesis that a PriA enzyme existed in *Mycobacterium* species, a situation that was proved by experimentation to be correct. A preliminary mutational analysis and characterisation for a *yabJ* homologue present in the HT cluster of streptomycetes was also reported, suggesting that this gene may be involved in the regulation of histidine and tryptophan biosynthesis. Finally, a putative functional link with tryptophan biosynthesis was proposed for the conserved SCO2040 family of proteins and the product of the *lgt* gene in actinomycetes. It is anticipated that the analysis presented herein may aid in guiding future efforts in the laboratory.

Chapter V

Chapter V

"Chorismate-utilising enzymes in *S. coelicolor*: towards the identification of the gene encoding for anthranilate synthase"

5.1 Introduction

The biochemistry of the enzymes that utilise as substrate the final product of the shikimate or common aromatic biosynthetic pathway (i.e. chorismic acid) has been extensively studied in the last few decades (for reviews of the different enzymes see Floss, 1997; Dosselaere and Vanderleyden, 2001; Crosa and Walsh, 2002; Knaggs, 2001 and 2003). One of the reasons for the high level of attention that this family of enzymes has received relates to the central role of chorismate at the core of the network of the small-molecule metabolic pathways. This network includes: the biosynthesis of the aromatic amino acids tryptophan, phenylalanine and tyrosine; the biosynthesis of the cofactors folic acid, menaquinone and ubiquinone; the biosynthesis of siderophores, i.e. specialised metabolites involved in iron acquisition; and secondary metabolites with ecologically important antibiotic activities (see Figure 5.1).

Since all of these biosynthetic pathways use chorismate as the starting substrate for the synthesis of their final-products it is reasonable to expect that, on one hand, the chorismate-utilising enzymes are evolutionary related and therefore show similarities, while, on the other hand, show sufficient-differences to catalyse the specific reaction within the pathway and be regulated in order to ensure the metabolic

balance within the cell. Indeed, the mechanistic and regulatory elements that each of these enzymes possess to convert chorismate can differ substantially making their elucidation a fundamental scientific challenge. For an example of the evolutionary relatedness and the complexity of the mechanisms of this family of enzymes see Parsons *et al.* (2002).

5.1.1 Biosynthesis of phenylalanine, tyrosine and tryptophan

The possible metabolic fates of chorismate are shown in the scheme of Figure 5.1. As discussed above, chorismate is the starting point for the biosynthesis of tryptophan, phenylalanine and tyrosine. The latter two amino acids share the first step in their biosynthesis, i.e. the synthesis of prephenate from chorismate *via* the activity of the enzyme chorismate mutase (CM, EC 5.4.99.5). From a structural perspective there are two types of CMs that differ in the fold that they adopt: the AroH class and the AroQ class. For a more detailed description of this classification see Dosselaere and Vanderleyden (2001). Furthermore, CM isoenzymes can show different structural organisation and allosteric regulatory features, i.e. they can be either monofunctional or be encoded within multidomain proteins such as the P-protein involved in phenylalanine biosynthesis in Gram-negative bacteria (Subramaniam *et al.*, 1994; Gu *et al.*, 1997). Two CMs from streptomycetes, i.e. *S. aureofaciens* and *S. venezuelae*, have been previously purified and characterised although no sequence was reported (Görisch, 1987; Hodgson, 2000 and references therein).

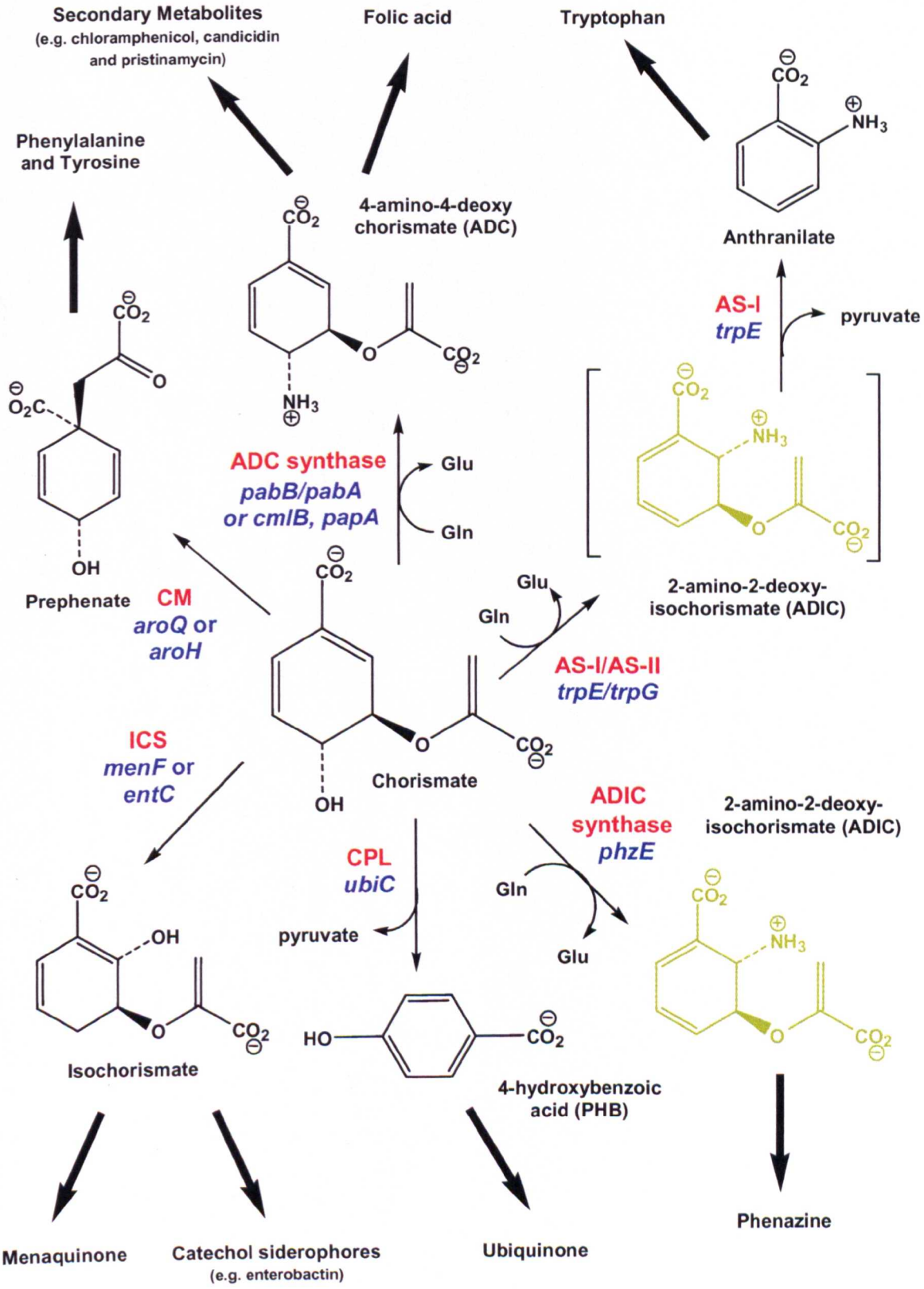


Figure 5.1 Chorismate-utilising enzymes (*see legend in following page*).

The remaining aromatic amino acid, i.e. tryptophan, is produced from chorismate after its conversion into anthranilate. This reaction involves the multidomain complex anthranilate synthase (AS, EC 4.1.3.27), which is found in bacteria as a heterotetramer of the products of the *trpE* and *trpG* genes in the form $TrpG_2$:$TrpE_2$ (Morollo and Eck, 2001; Spraggon *et al.*, 2001). The *trpG* gene encodes a glutamine amidotransferase of the so-called *triad* family whose function is to provide ammonia under physiological conditions from glutamine. However, the reaction can also proceed less efficiently without an amidotransferase at alkaline pH, implying that the product of *trpG* is not essential (for a review see Zalkin and Smith, 1998). The product of *trpE* is a bifunctional enzyme that associates with Mg^{2+} ions and TrpG from which ammonium is channelled in order to convert chorismate into anthranilate in a two-step process involving the formation of the non-diffusible intermediate 2-amino-2-deoxyisochorismate (ADIC, see Figure 5.1). One implication of this two-step mechanism is that TrpE contains two independent domains that can be physically separated to yield independent ADIC synthase and ADIC pyruvate-lyase activities (Morollo and Bauerle, 1993; see also Figure 5.1).

Figure 5.1 Chorismate-utilising enzymes (*see figure in previous page*). The different enzymatic conversions on which chorismate serve as substrate are schematically shown. The reactions shown are the first steps of a series of biosynthetic pathways composed of several steps represented by big arrows. The genes (blue) and enzymes (red) are as follow: AS-I, anthranilate synthase subunit I (*trpE*) and AS-II, anthranilate synthase subunit II (*trpG*); ADCS, 4-amino-4-deoxychorismate synthase (*pabB* for subunit I and *pabA* for subunit II in primary metabolism and the bifunctional *cmlB*, *papA*, etc, in secondary metabolism); ADICS, 2-amino-2-deoxyisochorismate synthase (encoded by *phzE* in *P. fluorescence*); CPL, chorismate pyruvate-lyase (*ubiC*); ICS, isochorismate synthase (*menF* in the synthesis of menaquinone and *entC* in the synthesis of catechol siderophores); CM, chorismate mutase (*aroQ* or *aroH*). The ADIC intermediary common to the synthesis of tryptophan and phenazines is highlighted in green. Refer to the text for further explanation.

5.1.2 The diversity of anthranilate synthase homologues

There are two closely related families of *trpE* paralogues whose products perform very similar reactions to the conversion catalysed by anthranilate synthase. This similarity is reflected by the fact that these three families of paralogues share the same EC number 4.1.3.27. One of these families is involved in the biosynthesis of phenazines, a group of secondary metabolites with important antibiotic activities whose genetics and biochemistry have been extensively characterised in *Pseudomonas* species (Mavrodi *et al.*, 2001; McDonald *et al.* 2001 and references therein). Phenazines had also been detected previously in *Streptomyces* and other actinomycete cultures (Geiger *et al.*, 1988; Gilpin *et al.*, 1995; Pusecker *et al.*, 1997; Kim *et al.*, 1997; Gebhardt *et al.*, 2002; Krastel *et al.*, 2002) although nothing has been reported on the genetics and biochemistry of the biosynthesis of these compounds in these organisms. The only piece of information on phenazine biosynthesis in actinomycetes is that it appears to be different from the system used by *Pseudomonas* species, as judged from circumstantial evidence derived from Southern hybridisation experiments (Mavrodi *et al.*, 2001).

It was originally assumed that in *Pseudomonas* species anthranilate served as the precursor for the synthesis of phenazine and other derivatives (Essar *et al.*, 1990). It was not until very recently that McDonald *et al.* (2001) demonstrated that the correct phenazine precursor is ADIC (see Figure 5.1). These researchers showed that a TrpE paralogue termed PhzE in *P. fluorescens*[1] uses chorismate as substrate and converts it into ADIC, the non-diffusible intermediary in the synthesis of anthranilate (see Section 5.1.1). The next step within the phenazine biosynthetic pathway is the conversion of ADIC into the anthranilate analogue *trans*-2,3-dihydro-3-

[1] As a word of caution please note that the names of the genes involved in the biosynthesis of phenazine and derivatives vary from species to species making their nomenclature rather confusing. Therefore it is recommended to use the name of the enzyme instead.

hydroxyanthranilic acid (DHHA) by the activity of the isochorismatase encoded by *phzD* thus ruling out the involvement of anthranilate as precursor in the synthesis of phenazine. Furthermore, it was shown that no product could be detected after incubation of the putative anthranilate-utilising Phz enzymes with anthranilate (McDonald *et al.*, 2001).

A subtle difference between AS of tryptophan biosynthesis and ADIC synthase of phenazine biosynthesis is found in their gene organisation. The TrpE and TrpG proteins involved in tryptophan biosynthesis are very rarely found fused within a single polypeptide encoded by a single gene (Crawford, 1989; Doolittle *et al.*, 1992). In contrast, the ADIC synthase involved in the biosynthesis of phenazine and its derivatives is always found as a bifunctional enzyme with ADIC synthase and amidotransferase activities encoded by the same gene (Mavrodi *et al.*, 1998 and 2001; Giddens *et al.*, 2002). Nevertheless, regardless of this observation the similarities between AS and ADIC synthase are remarkable to the point that they have proven to be interchangeable: AS can restore phenazine biosynthesis in ADIC synthase mutants and ADIC synthase can complement tryptophan auxotrophs deficient for AS (Essar *et al.*, 1990; Pierson *et al.*, 1995). In the first case some leakage of ADIC from the AS complex must exist whereas in the second case an independent ADIC pyruvate-lyase activity not related to ADIC synthase must convert ADIC into anthranilate (see Figure 5.1).

The second closely related paralogue of AS is involved in the biosynthesis of 4-amino-4-deoxychorismate (ADC); moreover, two very closely related proteins that are members of this family of ADC synthases have been reported. The first ADC synthase is a primary metabolic enzyme, which is involved in the biosynthesis of the primary metabolite *p*-aminobenzoic acid (PABA), a precursor of the synthesis of folic

acid (see Figure 5.1). This ADC synthase is encoded by the *pabB* and *pabA* genes, which are paralogues of *trpE* and *trpG*, respectively (Goncharoff and Nichols, 1984; Crawford, 1989). The main difference between AS and ADC synthase is that the latter lacks a pyruvate-lyase activity. A third gene product, PabC functions as a ADC pyruvate-lyase. Besides this difference the level of similarity between anthranilate synthase and ADC synthase is remarkable. Furthermore, it has been previously reported that in some organisms a unique amidotransferase participates in both anthranilate and ADC biosynthesis (Kaplan *et al.*, 1984; Crawford, 1989; Slock *et al.*, 1990), a nature that has been defined as *amphibolic*. This definition might appear to be at odds with the sense in which an *amphibolic character* is used in biochemistry textbooks, e.g. Voet and Voet (1995), where it is usually applied to enzymes or pathways that participate in both anabolic and catabolic pathways. However, the deamination of glutamine is in essence a catabolic reaction and therefore it makes sense to call TrpG amphibolic.

The second sub-family of ADC synthases appear to be involved in secondary metabolism (see Figure 5.1). So far this enzyme had only been found in antibiotic-producing streptomycetes, including *S. griseus*, *S. venezuelae* and *S. pristinaespiralis*, which produce candicidin, chloramphenicol and pristinamycin, respectively (Gil and Campelo-Diez, 2003; He *et al.*, 2001; Blanc *et al.*, 1997, respectively). Similarly to the organisational differences found between AS of tryptophan biosynthesis and ADIC synthase involved in phenazine biosynthesis the secondary metabolic ADC synthases appear to be within fusions, i.e. the *pabA* and *pabB* paralogous products are included within a single polypeptide. Interestingly, the ADC used in the biosynthesis of these antibiotics in these streptomycetes is synthesised *de novo* rather than derived from the ADC primary metabolic pool. This notion is supported by the observation

that inactivation of the genes encoding the secondary metabolic ADC synthases abolishes production of the antibiotics, while the mutant remains prototrophic. In contrast, when the genes encoding for the ADC synthases involved in the biosynthesis of PABA are inactivated the phenotype obtained is an auxotroph that requires PABA for growth with no effect on antibiotic production (see references cited above in addition to Criado *et al.*, 1993; Brown *et al.*, 1996; Chang *et al.*, 2001). However, despite these specific functional roles it is interesting to note that the candicidin *pabBA* gene from *S. griseus* was obtained by complementation of a *S. lividans pab* minus auxotroph (Gil and Hopwood, 1983; see also Section 5.2.4), implying a promiscuous activity.

5.1.3 Isochorismate synthase and chorismate pyruvate-lyase

As it can be seen in the scheme of Figure 5.1 the synthesis of the cofactors menaquinone (coenzyme Q) and ubiquinone (vitamin K2) also starts from chorismate. These isoprenoid cofactors are lipophilic, non-protein components of the membrane-bound electron-transfer chain. Most Gram-positive bacteria and anaerobic Gram-negative bacteria possess only menaquinone, whereas the majority of strictly aerobic Gram-negative organisms contain ubiquinone exclusively (Dosselaere and Vanderleyden, 2001). The biosynthesis of ubiquinones from chorismate involves the production of 4-hydroxy benzoic acid (PHB) *via* the activity of the enzyme chorismate pyruvate-lyase (CPL) enzyme (see Figure 5.1). Not much work on this chorismate-aromatising enzyme has been reported. However, it is known that the amino acid sequence of the *pabC* product, ADC pyruvate-lyase, involved in the biosynthesis of PABA (see Section 5.1.2) is not related to the sequence of the CPL enzyme encoded by the *ubiC* gene of *E. coli* (see Figure 5.1). Nevertheless, some

functional complementation of PabC by UbiC has been reported confirming that these enzymes are analogous (Nichols and Green, 1992).

Menaquinones are synthesised from chorismate thanks to the activity of the enzyme isochorismate synthase (ICS, EC 5.4.99.6; see Figure 5.1). In *E. coli* and *B. subtilis* there is a gene, termed *menF*, specific for this function (Dosselaere and Vanderleyden, 2001 and references therein). This observation is relevant because isochorismate is also the starting point for the synthesis of a broadly distributed family of catechol siderophores involved in iron acquisition represented by enterobactin of *E. coli* (Crosa and Walsh, 2002). The ICS devoted to biosynthesis of enterobactin is called *entC* and in contrast to *menF* its expression is iron-regulated (Brickman *et al.*, 1990). However, despite the specificity of the products of *menF* and *entC* paralogues, it has previously been shown that under the appropriate growth conditions the function of *menF* is complemented by the *entC* product and *vice versa* (Rowland *et al.*, 1996; Dahm *et al.*, 1998; Buss *et al.*, 2001).

5.1.4 The evolutionary relationships of the AS paralogues

Comparison of the amino acid sequences of the chorismate-utilising enzymes indicates that AS (*trpE*), ADIC synthase (*phzE*), ADC synthase (*pabB* and antibiotic paralogues) and ICS (*entC* and *menF*), which are mechanistically analogous, share significant sequence similarity (Dosselaere and Vanderleyden, 2001). Two independent phylogenetic analyses of ADIC synthases (*trpE* and phenazine paralogues) and ADC synthases (*pabB* and antibiotic paralogues), including their cognate amidotransferases, have been reported (Chang *et al.*, 2001; James *et al.*, 2002). Although the interpretation of the data by these researchers differs to some extent, probably as a consequence of the number of sequences used and the functions

assumed for each enzyme, at least through my eyes, their results appear as coherent and agree well overall. Remarkably, despite the fact that the functional differences of the AS paralogues appear to be minor, the sequence differences are not minor, a fact that is reflected by the phylogenetic trees reported by these authors (see Figure 5.2 and Section 5.2.1). A similar analysis by Gary Xie and Roy Jensen at Los Alamos National Laboratory has confirmed these results (G. Xie and R. Jensen, personal communication).

The aim of this Chapter is to help to resolve the evolutionary relationships and functions of each AS paralogue present in the genome of *S. coelicolor*. This analysis was done with particular emphasis on identifying the *trpE* gene that encodes the AS involved in tryptophan biosynthesis. For this purpose the phylogenetic tree reported by Chang and co-workers (2001) is used since all of the putative AS paralogues present in *S. coelicolor* are included (see Figure 5.2). Of these AS paralogs two of them were identified before the *S. coelicolor* genome sequencing-project: *trpE1* (Hu, 1995; Hu *et al.*, 1999) and *trpE2* (Chong *et al.*, 1998; Hojati *et al.*, 2002). During the development of this thesis the remaining *trpE* paralogues, i.e. *trpE3*, *trpE4* and *trpE5*, were reported by the genome-sequencing project (Bentley *et al.*, 2002), confirming that this organism has a total of five putative *trpE* genes encoding AS paralogues. In the following sections the products of the five paralogous *trpE* genes present in *S. coelicolor* are grouped with other chorismate-utilising enzymes according to their sequence similarities and functional characteristics. Some of the functions predicted in this way were tested through mutational analysis. Furthermore, preliminary efforts aimed at the study of the regulatory patterns of *trpE1* are reported.

5.2 Results and Discussion

5.2.1 Definition of the TrpE paralogues present in *S. coelicolor*

The five TrpE paralogues of *S. coelicolor* are shown in the phylogenetic tree of Figure 5.2 (Chang *et al.*, 2001). Some characteristics of the TrpE paralogues analysed in subsequent Sections are summarised in Table 6.1 and the similarity matrix shown in Table 5.2 demonstrates the paralogy of these TrpE-like proteins. It can be observed in this table that the greatest similarity is between the products of the *trpE2* and *trpE3* genes (43% identity). The similarities between the other TrpE paralogues range from 14.4% to 24.9% identity and therefore are less significant. The figures reported in the matrix of Table 5.2 were obtained after aligning a version of TrpE1 from which the amidotransferase domain was deleted from its C-terminus and the intact remaining TrpE sequences. This deletion from residue number 433 to 615 was done since, as judged from domain predictions *in silico*, it is anticipated that TrpE1 is a fusion protein including the amidotransferase domain or subunit II of AS at its C-terminal end. In contrast, this domain could not be detected in the sequences of any other TrpE paralogue (Table 5.1). The deletion ensured that the sequences compared are equivalent and correspond to homologous domains, avoiding the possibility of a bias in the comparisons.

Table 5.1 General characteristics of the TrpE paralogues of *S. coelicolor*.

Paralogue	Gene number	Cosmid	Length (aa)	Predicted domains	Comments
TrpE1	SCO2117	Sc6E10	634	AS subunits I and II	Complements *trpE* in *E. coli*. Shows a putative leader peptide and attenuator upstream. Its expression seems to be slightly feedback-regulated by tryptophan, although at very high concentrations. Fused with a TrpG paralogue.
TrpE2	SCO3214	ScE8	511	AS subunit I	Member of a *trp* operon with the organisation *trpE2GD2C2* that seems to be co-regulated with genes involved in the biosynthesis of CDA. Therefore, a TrpG homologue is associated with this paralogue although they are not fused.
TrpE3	SCO2043	Sc4G6	502	AS subunit I	Member of the HT cluster of genes discussed in Chapter IV and therefore highly conserved across actinomycetes. There is no obvious TrpG associated with this paralogue.
TrpE4	SCO1547	ScL11	353	AS subunit I	A putative *pabC* homologue appears to be associated with this paralogue as they appear to form what looks like a three-genes operon. There is no obvious TrpG or PabA homologue associated with this paralogue.
TrpE5	SCO7691	Sc4C2	471	AS subunit I	Member of a cluster of genes predicted to be involved in the biosynthesis of a catechol siderophore called coelibactin that still remains to be identified.

Table 5.2 Similarity matrix of the TrpE paralogues.

	TrpE1	TrpE2	TrpE3	TrpE4	TrpE5
TrpE1	1.000	0.162	0.166	0.197	0.144
TrpE2	----	1.000	0.430	0.236	0.249
TrpE3	----	----	1.000	0.215	0.242
TrpE4	----	----	----	1.000	0.171
TrpE5	----	----	----	----	1.000

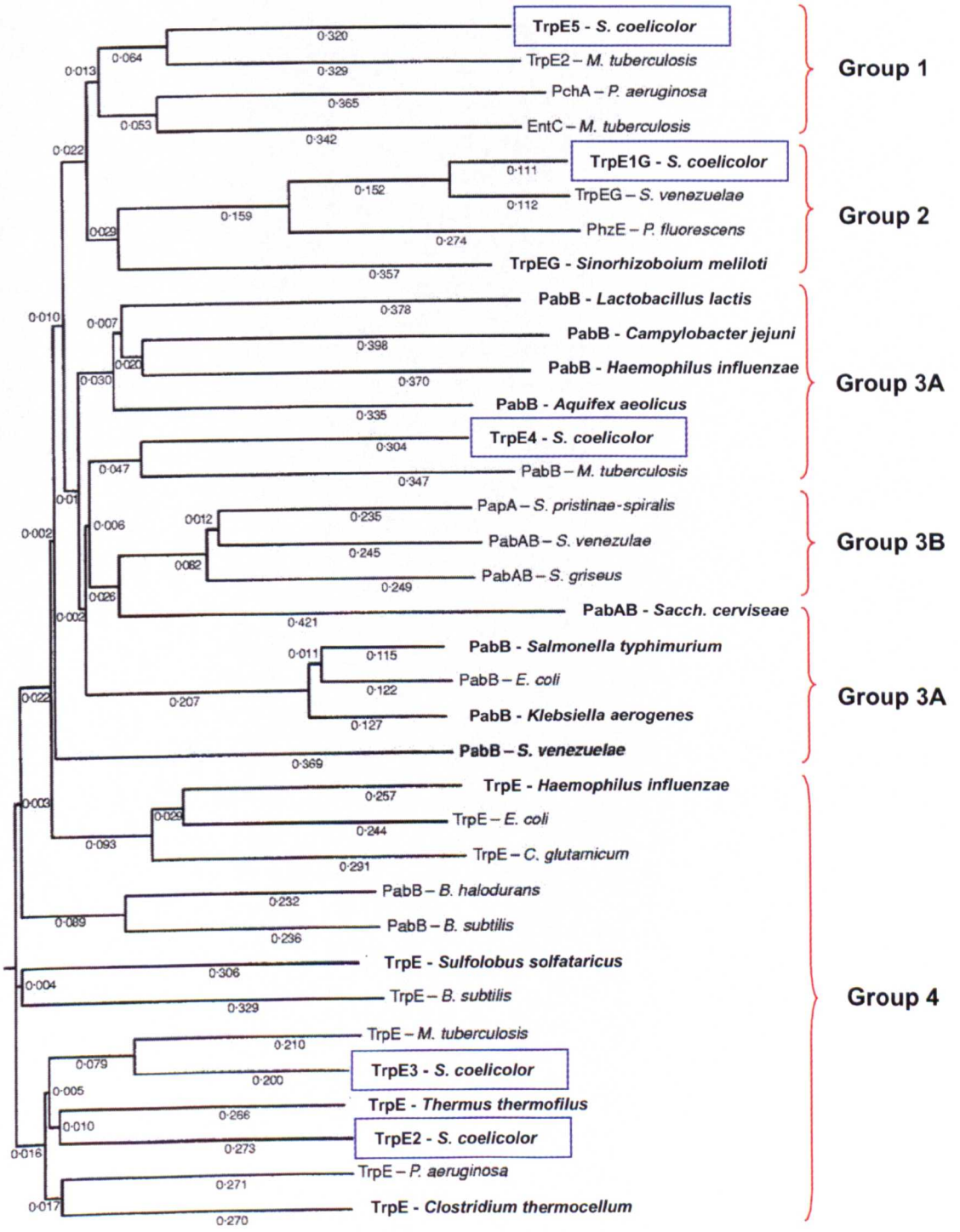


Figure 5.2 Phylogenetic tree of TrpE paralogues. The tree was taken from Chang *et al.* (2001). Refer to this reference for accession numbers of sequences. The tree shows all five TrpE paralogues of *S. coelicolor* (highlighted in blue boxes). Five different groups, which account for all five TrpE paralogues present on *S. coelicolor*, are proposed on the basis of sequence similarities and functional characterisation. Refer to text for further explanation.

5.2.2 Is the anthanilate synthase of *S. coelicolor* encoded by the *trpE1* gene?

As previously discussed in Section 5.1.4 the *trpE1* gene was discovered before the genome sequencing-project of *S. coelicolor*. Hu (1995) cloned this gene after Southern hybridisation against digested chromosomal DNA of *S. coelicolor* with a probe containing the *trpEG* gene of *S. venezuelae* (Lin *et al.*, 1998). Complementation of an *E. coli* mutant deficient in AS activity was subsequently used to identify the DNA fragment containing the *trpE1* gene (Hu, 1995). Sequence analysis of the *trpEG* gene from both *S. venezuelae* and *S. coelicolor* had shown that their products are homologous to the subunits I and II of AS and therefore these domains are fused within a single polypeptide (see Table 5.1). Furthermore, these genes have in common the location just upstream of a small ORF that resembles the *trpL* gene of *E. coli* (Paradkar *et al.*, 1991; Hu, 1995; Lin *et al.*, 1998). In *E. coli* and its relatives the *trpL* gene encodes a leader peptide rich in tryptophan residues involved in the regulation of the *trp* operon *via* transcriptional attenuation (Yanofsky, 2001 and 2003; Henkin and Yanosky, 2002 and references therein). It has been previously noted that the sequence of the putative transcript upstream of both *trpEG* genes, including the predicted *trpL* gene, has the potential for forming an RNA secondary structure that resembles that of a typical attenuator structure (Hu, 1995; Lin *et al.*, 1998).

Based on the features discussed above the *trpEG* genes of *S. coelicolor* and *S. venezuelae* were assumed to be involved in tryptophan biosynthesis. However, Lin *et al.* (1998) reported that a two-base deletion in the *trpEG* gene of *S. venezuelae* that stopped complementation of an *E. coli trpE* mutant did not lead to a tryptophan auxotroph when used to replace the wild type *trpEG* gene of *S. venezuelae*. Furthermore, despite extensive efforts aimed at demonstrating the expected feedback-repression by tryptophan of the *trpEG* gene *via* attenuation these researchers could

only detect a very low level of repression when using very high concentrations of tryptophan. These observations, in addition to the fact that the amino acid sequences of these *trpEG* gene products more closely resemble the sequence of ADIC synthases involved in phenazine biosynthesis (Group 1 in Figure 5.2) place some doubts on a role for this enzyme in tryptophan biosynthesis. The closest homologue of TrpE1 from a non-actinomycete that has been unambiguously functionally characterised is the ADIC synthase (50.2% identity) encoded by the *phzE* gene of *P. aeruginosa* PAO1 (Mavrodi *et al.*, 2001). Therefore, further experimentation is required if the role of TrpE1 in *S. coelicolor* is to be clarified.

5.2.2.1 Mutational analysis of *trpE1* (SCO2117)

The *trpE1* gene was deleted from the chromosome of *S. coelicolor* M145 using the PCR-targeting mutagenesis system REDIRECT® employed in Section 2.2.3 for the disruption of *priA* (see Section 7.1.3.2 in Methods and Appendix 2). Since a possible polar effect caused by the disruption of *trpE1* (SCO2117) was not expected due to the fact that this gene appears not to be within a putative operon a marked replacement using the apramycin *aac(3)IV* plus *oriT* cassette from pIJ773 was used. Colonies that proved to have undergone a double crossover as judged by kanamycin sensitivity and PCR screening were obtained immediately after conjugation between *S. coelicolor* M145 and *E. coli* ET12576/pUZ8002 cells bearing the disrupted Sc6E10 cosmid and selection with apramycin. This observation contrasts with the situation observed when the *priA* minus mutant WH101 was constructed (see Section 2.2.3) in that it appears that no selection against the disruption of *trpE1* was present, suggesting that whatever the role of *trpE1* is, it appears to be dispensable under the conditions used.

This preliminary observation fits well with the involvement of *trpE1* in ADIC biosynthesis other than in the biosynthesis of anthranilate. The medium used for this conjugation contained soya-flour (i.e. SFM) and it was previously noted that supplementation of SFM with tryptophan promoted growth of the *priA* mutant WH101, suggesting a tryptophan deficiency in SFM (see Section 2.2.3). When the resulting *S. coelicolor SCO2117::aac(3)IV* mutant, termed WPZ101, was tested for its growth requirements in minimal medium it proved to be a prototroph. The mutant strain WPZ101 is able to grow in minimal medium as well as other standard media without any apparent effect upon growth as a consequence of the disruption of *trpE1*. This result is in agreement with the phenotype reported by Lin *et al.* (1998) of the *S. venezuelae trpEG* mutant. Nonetheless, these researchers concluded that the *S. venezuelae* orthologue of *trpE1* was involved in tryptophan biosynthesis in *S. venezuelae* and proposed that there was a second set of *trpE* and *trpG* genes in this organism that were capable of compensating for the loss of *trpEG* (see also Section 5.2.2.2). In the light of the multiple AS paralogues detected in *S. coelicolor* (Tables 5.1 and 5.2) this possibility appears to be a plausible explanation for the prototrophic phenotype of the mutant strain WPZ101. However, involvement of the *trpEG* gene fusion in phenazine biosynthesis in *S. venezuelae* was not ruled out and no evidence was offered against this possibility (Lin *et al.*, 1998).

Despite the fact that other streptomycetes are known to produce phenazine secondary metabolites (Geiger *et al.*, 1988; Gilpin *et al.*, 1995; Pusecker *et al.*, 1997; Kim *et al.*, 1997; Gebhardt *et al.*, 2002) neither *S. coelicolor* nor *S. venezuelae* have been reported to produce any of these compounds nor a secondary metabolite derived from ADIC. A further complication of the proposal of the involvement of *trpE1* in phenazine biosynthesis is the apparent lack of other putative *phz* genes from the

genome of *S. coelicolor*, particularly within the vicinity of *trpE1*. A typical conserved organisation of *phz*-like genes is encountered in phenazine-producing *Pseudomonas* species (Mavrodi *et al.*, 1998 and 2001; McDonald *et al.*, 2001 and references therein) and *Erwinia herbicola* (Giddens *et al.*, 2002). Although nothing has been reported on the genetics of phenazine biosynthesis in streptomycetes, Mavrodi *et al.* (2001) have shown using Southern hybridisation experiments that the organisation of the putative *phz* genes in the phenazine-producing actinomycete *Brevibacterium iodinum* (and *Burkholderia* species) is different to that typically encountered in *Pseudomonas* species as they are not clustered. Therefore, it is tempting to speculate that *trpE1* in *S. coelicolor* is involved in ADIC biosynthesis and possibly in phenazine production potentially representing the first example of non-clustered secondary metabolic genes.

This possibility is being examined at the present time by the author of this thesis within a postdoctoral research project on functional genomics of the secondary metabolism of *S. coelicolor* at the Department of Chemistry, University of Warwick. Preliminary data on the secondary metabolite profile of WPZ101 suggests that differences between the *trpE1* mutant and wild type cultures can be detected (data not shown), although the identity of the metabolites produced is still under investigation and a phenazine-like compound cannot be confirmed at this stage. Furthermore, the notion that TrpE1 is a secondary metabolic ADIC synthase is supported by the observation that this protein has been identified by the *S. coelicolor* proteomics program at the John Innes Centre as a protein whose expression pattern resembles that of a secondary metabolic enzyme, i.e. expression at late exponential phase or early stationary phase (A.R. Hesketh, personal communication).

5.2.2.2 Preliminary efforts aimed at investigating the regulation of *trpE1*

As discussed in previous Sections the regulation of the putative *S. coelicolor* *trpE1* orthologue from *S. venezuelae* has been studied to some detail (Lin *et al.*, 1998). The effect of addition of tryptophan on the expression of this *trpEG* fusion was determined both in *E. coli* and *S. venezuelae* tryptophan auxotrophs. The levels of AS activity were measured and two-fold repression at non-physiological concentrations of tryptophan was reported (Lin *et al.*, 1998). However, it has previously been suggested that when studying the regulation of *Streptomyces* primary metabolic enzymes the significance of two-fold and less repression is questionable (Hodgson, 2000). Similar experiments to those by Lin *et al.* (1998) have been previously reported for other streptomycete amino acid biosynthetic genes that also showed a predicted attenuator structure upstream (Craster *et al.*, 1999). These researchers reported that the regulatory role of the putative classical attenuator predicted by sequence analysis upstream of *ilvB* and *leuA* is negligible. In disagreement with Lin *et al.* (1998), this study suggests that an unknown regulatory mechanism must be operating in the *ilv* / *leu* system of *S. coelicolor* in place of classical attenuation (Craster *et al.*, 1999). Taking into account these antecedents it is interesting and relevant to the aim of this project to determine the regulatory role, if any, of the putative attenuator upstream of *trpE1*. It is anticipated that knowing the regulatory features of *trpE1*, i.e. whether it is inhibited by tryptophan, may shed some light into its functional role.

For this purpose two versions of *trpE1* were amplified by PCR from DNA cosmid Sc6E10 as template using primers containing suitable restriction sites engineered at their 5' ends (see Table 7.1 in Methods). The first PCR product was a 2.3 kb-fragment that contained the *trpE1* gene plus its upstream region including the

putative attenuator. The second PCR product was a 2.06 kb-fragment that included a promoterless version of *trpE1* without the putative attenuator structure. Subsequently, two different 5' end versions of *trpE1* were obtained from these PCR products by exploiting a *Bam*HI site that lies at the 5' end of *trpE1* and an engineered *Hind*III site introduced at the 5' end of the PCR products. The *Hind*III / *BamH*I cleaved products were successfully cloned into the cloning vector pMTL21 using the same sites and the correct sequences of the inserts were determined by sequencing of both strands. The recombinant *trpE1* DNA-fragments were recovered from the pMTL21 constructs by restriction digestion with *Hind*III and *BamH*I after passing the constructs through a methylation-deficient *E. coli* strain. These fragments were intended to be sub-cloned as *xylE* fusions in the *Streptomyces* promoter-probe pIJ2842 (Clayton and Bibb, 1990) using the unique *Hind*III and *BamH*I sites of this plasmid. The use of these sites ensures the correct orientation of the inserts.

The promoter-probe pIJ2842 is derived from the SCP2* plasmid and therefore is a low-copy number plasmid suitable for studying the putative regulatory role of the attenuator structure upstream of *trpE1*. This is because for this type of regulatory studies it is recommended using a low-copy number vector since high-copy number plasmids have been previously shown to give unreliable results (discussed by Kieser *et al.*, 2000, pp. 341). The pIJ2842 vector includes a promoterless *xylE* reporter gene whose expression is easy to assess through a chromogenic assay. The *xylE* gene encodes for a catechol 2,3-dioxygenase that converts the colourless catechol into 2-hydroxymuconic semialdehyde, which is yellow. Unfortunately, the large size of pIJ2842 (around 12 kb) and its low-copy number makes the handling of this plasmid difficult. In order to avoid these limitations large-scale preparations of pIJ2842 using caesium chloride density gradient centrifugations were performed. Despite several

attempts of sub-cloning the two 5' end versions of *trpE1* into pIJ2842 were done, no positive clones were obtained. This negative result, which precludes the possibility of investigating the regulatory pattern of *trpE1*, is likely to be due to technical problems rather than to a problem with the approach followed.

5.2.3 Investigation of *trpE2* (SCO3214) and *trpE3* (SCO2043)

The products of the *trpE2* and *trpE3* genes are not only the closet *S. coelicolor* TrpE paralogues (Table 6.2) but also represent the closest homologues of ASs from other organisms (Group 4 in Figure 5.2). The closest homologue of both TrpE2 and TrpE3 from a non-actinomycete that has been functionally characterised is the AS subunit I of *P. putida* (39.2% and 42.9% identity, respectively; Essar *et al.*, 1990). Therefore, it is likely that either TrpE2 or TrpE3 is the AS involved in the biosynthesis of tryptophan in *S. coelicolor*. As discussed in Section 5.1.4 the TrpE2 paralogue was discovered before the genome-sequencing project (Chong *et al.*, 1998; Hojati *et al.*, 2002). Chon *et al.* (1998) cloned the *trpE2* gene (SCO3214) as part of a genomic region that included the genes that direct the production of CDA (see also Section 1.3.2). This secondary metabolite is a lipopeptide whose structure includes two tryptophan residues, one of which is in its D configuration (Kempter *et al.*, 1997). It has recently been shown that production of this lipopeptide is directed by a complex set of non-ribosomal peptide synthetases in addition to several sets of auxiliary genes responsible for synthesis of precursors (Hojati *et al.*, 2002). Among these, it has been presumed that the tryptophan residues incorporated into the peptide backbone of CDA are produced *de novo* by an auxiliary set of *trp* genes. These *trp* genes appear to be arranged in what looks to be an operon with the organisation *trpE2GD2C2aroH*, which also includes an *aro* gene potentially encoding a plant-like or class II 2-keto-3-

deoxy-D-arabinoheptusonate-7-phosphate (DHAP) synthase (EC 2.5.1.54). This *trp-aro* operon lies at the boundaries of the genomic region devoted to the production of CDA. Although no mutational analysis of *trpE2* or of any of these genes has been reported it has been shown that expression of *trpE2*, as well as the expression of the entire *trpE2GD2C2aroH* operon, is co-ordinated with the expression of the *cda* genes during stationary phase (Huang *et al.*, 2001; Ryding *et al.*, 2002). These observations suggest that TrpE2 is functionally linked to the biosynthesis of CDA and may direct the production of tryptophan for this secondary metabolite. Therefore, although TrpE2 might actually be an AS, a situation that still needs to be confirmed, it is unlikely that this protein is the AS involved in the biosynthesis of tryptophan during primary metabolism.

One consequence of this conclusion is that the putative AS encoded by *trpE3* (SCO2043) appears to be a good candidate to be involved in the biosynthesis of tryptophan during primary metabolism. This possibility is supported by the observation that the *trpE3* gene is highly conserved within the so-called HT cluster of actinomycetes extensively discussed in Chapter IV. This contrasts with the rest of the *S. coelicolor* putative *trpE* genes, which are not always conserved among the set of actinomycete genomes that have been analysed. Therefore, the level of conservation of *trpE3* across the actinomycetes suggests that the function of this gene is central to these organisms, such as the biosynthesis of a proteinogenic amino acid.

5.2.3.1 Mutational analysis of *trpE3* (SCO2043)

The hypothesis of the involvement of the product of *trpE3* in the biosynthesis of tryptophan was tested by mutational analysis. This gene was disrupted following the same approach as for the disruption of *trpE1* (see Section 7.1.3.2 in Methods and

Appendix 2). The *trpE3* mutant, termed WH103 is also an *aac(3)IV* plus *oriT* marked strain, raising the possibility of polar effects on the *trp* cluster (see Figures 2.3 and 4.3). However, analogous to strain WPZ101, the WH103 *trpE3* mutant strain was obtained immediately after the conjugation experiment and when it was screened for its growth requirements it was found that WH103 is also a prototroph that can grow in minimal medium without tryptophan.

One obvious explanation for the phenotype recorded for WH103 is that the lack of *trpE3* is compensated by the putative ADIC synthase encoded by *trpE1*. As discussed in Section 5.1.2 the ADIC synthase involved in phenazine biosynthesis and AS of *Pseudomonas* species have proven to be interchangeable, i.e. these proteins can complement each other (Essar *et al.*, 1990; Pierson *et al.*, 1995). Furthermore, from the results reported by Hu (1995) it is known that *trpE1* can complement an *E. coli* strain with a *trpE*-mutated gene. As part of my current postdoctoral research a double *trpE1* and *trpE3* mutant was constructed. Preliminary characterisation of this mutant, termed WH104, revealed that this strain is not an auxotroph. This negative result has not been corroborated, for example, by assaying for the levels of AS activity in cell-free extracts obtained from cultures of wild type and the mutant strains.

5.2.4 Identification of the *pabB* gene of *S. coelicolor in silico*

The two remaining AS paralogues that have not been investigated hitherto are the products of *trpE4* and *trpE5* (see Table 5.1). The TrpE5 protein has been predicted to be involved in the biosynthesis of a siderophore called coelibactin that still needs to be identified (Bentley *et al.*, 2002; G. Challis, personal communication). This enzyme is believed to be an isochorismate synthase as inferred from sequence similarity to EntC (19.4% identity) from the biosynthesis of enterobactin (Ozenberger *et al.*, 1989).

No amidotransferase activity is commonly associated with the function of this family of chorismate-utilising enzymes (Group 1 in Figure 5.2) and therefore is difficult to imagine how TrpE5 could be involved in the biosynthesis of anthranilate or compensating for the lack of *trpE3* and *trpE1*.

This view contrasts with the product of *trpE4*, which groups within the phylogenetic tree shown in Figure 5.2 with the sequences of ADC synthases (Group 3), particularly with those involved in the biosynthesis of PABA during primary metabolism (Sub-group 3A) rather then with the secondary metabolic enzymes (Sub-group 3B). However, the PabB and PabA proteins of *S. coelicolor* remain to be unambiguously identified. A *pab* minus mutant of *S. lividans* (strain JG10) that required PABA for growth has been previously isolated by chemical mutagenesis (Gil and Hopwood, 1983). Assuming that the similarities between *S. coelicolor* and *S. lividans*, two strains that are very closely related, are conserved in terms of chorismate metabolism one putative implication of the isolation of JG10 is that the ADC synthase of *S. coelicolor* may not be compensated by any other AS paralogue. Although this is in disagreement with the reciprocal overlapping promiscuous activities extensively discussed in this Chapter, this observation suggests that the prototrophic phenotypes of the *trpE* mutants discussed in previous Sections may not be due to compensation by the putative PabB protein of *S. coelicolor*, possibly encoded by *trpE4*. This would also imply that neither *trpE1* nor *trpE3* encode an ADC synthase.

The fact that the *pab* mutation on JG10 was complemented by a *pabB* paralogue from *S. griseus* encoding for an ADC synthase involved in secondary metabolism (Criado *et al.*, 1993), suggests that this mutation is related to *pabB* and not to *pabC*. Moreover, Arhin and Vining (1993) cloned the putative *pabB* and *pabA* genes from *S. lividans* by complementation of an *E. coli* PABA auxotroph. The

construct obtained from this experiment, bearing the *pabA* and *pabB* genes of *S. lividans* was also shown to rescue the PABA auxotrophy of JG10. Although this is strong evidence supporting the view that the mutation in JG10 is exclusively related to the *pabB* gene, it is not evidence suggesting that the products of the genes complementing the *pab* mutation of JG10 are involved in biosynthesis of this cofactor during primary metabolism. Detailed analysis of the sequence of the putative PabB isolated from *S. lividans* (Arhin and Vining, 1993) and the predicted TrpE4 protein of *S. coelicolor* shows only low levels of similarity (19.8% identity). Although this level of identity may be enough for suggesting a common function for two proteins from distantly related organisms, it is well below the expected level of identity for these two closely related streptomycetes. Moreover, the sequence reported by Arhin and Vining (1993) shows more similarity to other TrpE proteins from *S. coelicolor*, mainly, TrpE2 (25.6 % identity) and TrpE3 (28.2% identity), although the *pabB* gene of *S. lividans* was associated with and ORF that shows high similarity to *pabA* genes from other organisms (Arhin and Vining, 1993).

BLAST searches using this PabA sequence against the complete set of proteins of *S. coelicolor* revealed that there are three putative glutamine amidotransferases that are highly similar to this protein within the proteome of *S. coelicolor*. However, none of them seem to be associated with TrpE4. The proteins identified by this means were the TrpG protein of the CDA *trp* operon, the C-terminus domain of TrpE1 and SCO3851 (37.5%, 21.3%, 33.3% identity, respectively). Based on the close relatedness between *S. coelicolor* and *S. lividans*, and whether the PabB protein isolated by Arhin and Vining (1993) is in fact the ADC synthase involved in primary metabolism, this result is in disagreement with the possible role of TrpE4 as an ADC synthase involved in the biosynthesis of PABA in *S. coelicolor*. A possible

explanation to the phenotypes of the mutant strains reported in this Chapter is that TrpE4 could be an AS. However, the possibility that an ADC synthase is encoded by *trpE4* is supported by the observation that a *pabC* homologue lays just downstream of this gene (see Section 5.1.2). This observation, in addition to the low levels of sequence identity detected between TrpE4 and the putative PabB protein of *S. lividans*, raises some doubts about the assumed primary metabolic function of the genes isolated by Arhin and Vining (1993). It is worth noting that the putative *pabB* and *pabA* genes of *S. lividans* were isolated using an *E. coli pab* mutant rather than the JG10 strain (Arhin and Vining, 1993) and that the latter can be complemented by a secondary metabolic *pabB* paralogue from *S. griseus* (Criado *et al.*, 1993). Although the unambiguous identification of the *pabB* genes from *S. lividans* and *S. coelicolor* requires further experimentation, this analysis favours TrpE4 to be the real PabB protein. I could be that the putative *pabB* and *pabA* homologues isolated by Arhin and Vining (1993) are in fact orthologues of *trpE2* and *trpG* involved in the biosynthesis of CDA.

5.3. Conclusions

A genomic analysis of proteins potentially encoding for anthranilate synthases in *S. coelicolor* was presented, showing the complexity of the paralogous relationships between chorismate-utilising enzymes in this organism. The first steps towards the identification of the TrpE paralogue responsible for the conversion of chorismate into anthranilate *via* the activity of the bifunctional enzyme anthranilate synthase in tryptophan biosynthesis were presented. Two out of the five AS paralogues present on the genome of *S. coelicolor*, namely, *trpE1* (SCO2117) and *trpE3* (SCO2043), were mutated independently and in combination. The resulting mutant strains turned out to be prototrophs, suggesting extensive duplication of genes encoding AS paralogues. Several explanations to this phenotype were discussed although no definitive solution to this conundrum was given. It is anticipated that the analysis presented herein will serve to settle the basis for further investigation on the complex relationships of chorismate-utilising paralogous enzymes in *Streptomyces*.

Chapter VI

Chapter VI

"Characterisation of proline metabolism in *S. coelicolor*: some implications for secondary metabolism"

6.1 Introduction

6.1.1 Proline biosynthesis in *S. coelicolor*

Proline biosynthesis typically occurs from glutamate *via* the intermediates γ-glutamyl phosphate, glutamate-γ-semialdehyde and Δ^1-pyrroline-5-carboxylate (P5C; reviewed by Adams and Frank, 1980). The enzymes involved in the pathway are γ-glutamyl kinase (GK, EC 2.7.2.11, *proB*), which phosphorylates glutamate; glutamate-γ-semialdehyde dehydrogenase (GS-DH, EC 1.2.1.41, *proA*), which converts γ-glutamyl phosphate into glutamate-γ-semialdehyde; and P5C reductase (P5C-R, EC 1.5.1.2, *proC*), which converts P5C, the product of the spontaneous dehydration of glutamate-γ-semialdehyde, into proline (see pathway in Figure 6.1). This pathway has been shown to be operative in *S. coelicolor* by previous workers in our laboratory (Hood *et al.*, 1992). In conformity with the lack of amino acid biosynthesis gene regulation observed in streptomycetes (see Section 1.3.1.1), it was reported that P5C-R activity was not repressed by proline but, moreover, twofold stimulation of P5C-R activity was detected when the cells were grown in the presence of proline (Hood *et al.*, 1992).

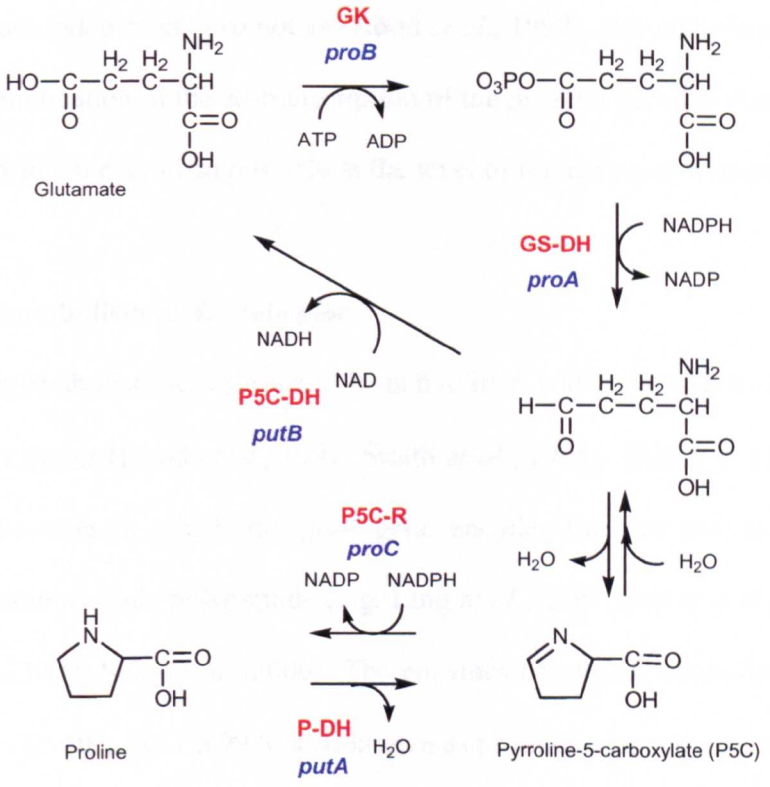


Figure 6.1 Proline metabolism in *S. coelicolor*. The intermediates, enzymes (red) and genes (blue) involved in the biosynthesis and catabolism of proline in *S. coelicolor* are shown. GK, γ-glutamyl kinase (*proB*); GS-DH, glutamate-γ-semialdehyde dehydrogenase (*proA*); P5C-R, Δ'-pyrroline-5-carboxylate reductase (*proC*); P-DH, proline dehydrogenase (*putA*); and P5C-DH, Δ'-pyrroline-5-carboxylate dehydrogenase (*putB*). Refer to text for further explanation.

Hood *et al.* (1992) also cloned by complementation of a *S. coelicolor* proline auxotroph the *proBA* genes. Similar to the enzyme assay data of P5C-R it was reported that neither *proB* nor *proA* were regulated at the level of gene expression as revealed by promoter probe studies. Furthermore, sequencing of the *proB* (SCO2587) and *proA* (SCO2585) genes by Hood *et al.* (1992), and later on by the genome-sequencing project, revealed that these genes are separated by a small ORF (SCO2586 or *proX*) with no homology with anything in the databases. The plasmid containing the genomic region with the *proBXA* genes was termed pDHP14 and it is a pIJ702 derivative. Finally, it was demonstrated by promoter-mapping experiments that the *proBX* and *proA* genes were

transcribed from independent promoters (Hood *et al.*, 1992), although they may form an operon. One implication of the co-transcription of the *proBXA* genes is that *proX* may be involved in proline metabolism possibly at the level of regulation or transport.

6.1.2 Proline catabolism in *S. coelicolor*

Proline catabolism was shown to be active in *S. coelicolor* and to consist of two independent enzymes (Hood *et al.*, 1992; Smith *et al.*, 1995). This is in contrast to α, β and γ proteobacteria in which the *putA* gene encodes the two enzymes of proline catabolism within a single polypeptide (e.g. Ling *et al.*, 1994; Surber and Maloy, 1998; Vílchez *et al.*, 2000; Soto *et al.*, 2000). The enzymes of proline catabolism are proline dehydrogenase (P-DH, EC 1.5.99.8; also known as proline oxidase), a membrane-bound protein that converts proline into P5C; and P5C dehydrogenase (P5C-DH, EC 1.5.1.12), which was found to be a cytoplasmic protein capable of converting glutamate-γ-semialdehyde, the spontaneous product of P5C hydration, into glutamate (see the biochemical pathway in Figure 6.1). Previous work in our laboratory has shown that P5C-DH is three hundredfold inducible at the level of transcription by proline supplemented exogenously (Smith *et al.*, 1995).

As discussed in Section 1.1.2 sub-functionalisation of a multidomain protein into independent proteins has been previously reported as one of the possible means by which proteins may evolve. Therefore, the observations by Hood *et al.* (1992) and Smith *et al.* (1995) discussed above are circumstantial evidence suggesting that either in *S. coelicolor* PutA underwent sub-functionalisation, implying that *putA* may be present in two different genes on the chromosome of this organism, or that *putB* and *putA* in

proteobacteria were fused at one point during the course of evolution. One implication of this scenario is that different gene names specific for each domain present on the proteobacterial PutA protein are required for *S. coelicolor*. It is proposed therefore to call the gene encoding the enzyme P5C-DH *putB* and to use *putA* exclusively for the gene encoding P-DH.

6.1.3 Proline as precursor for synthesis of secondary metabolites

When studying proline metabolism in *S. coelicolor* it is relevant to note that this organism produces a secondary metabolite that uses proline as precursor for synthesis, namely, the complex of red-pigmented tripyrrole prodiginines or RED (Wasserman *et al.*, 1966; see also Section 1.3.2). It has been shown by precursor labelling experiments that RED is derived from one unit of proline, one unit of glycine, one unit of serine and several units of acetate (Wasserman *et al.*, 1973; Figure 6.2). Furthermore, based on sequence comparisons functions have been assigned to the majority of the *red* genes, and a complete (convergent) pathway for the synthesis of RED, involving condensation of 4-methoxy-2,2′-bypyrrole-5-carboxaldehyde and 2-undecylpyrrole at a late stage, has been proposed (Cerdeño *et al.*, 2001). In recent years the biogenesis and chemical synthesis of the pyrrole alkaloids of the prodiginine family (also known as prodigiosins) have gained increasing attention since the discovery of the potential clinical use of these metabolites as immunosuppressants (reviewed by Manderville, 2001 and Fürstner, 2003). However, nothing is known about the mechanisms that mediate the interaction between proline metabolism (Figure 6.1) and the RED biosynthetic pathway (Figure 6.2), nor the biological role of these metabolites on the producing organism.

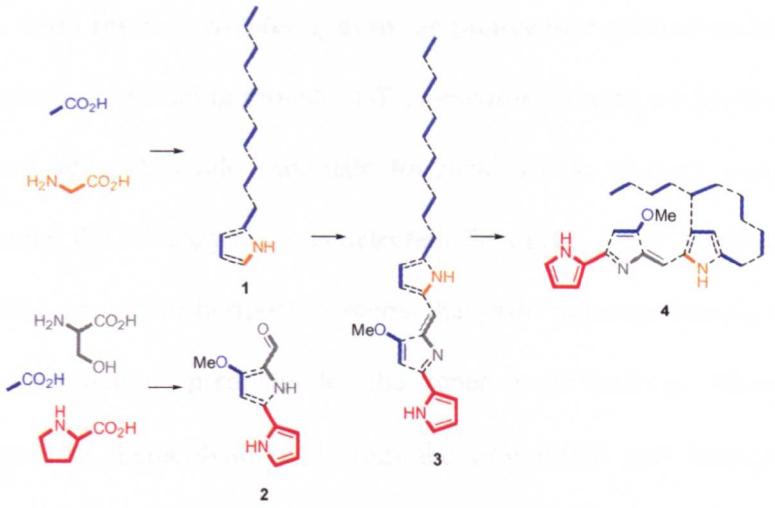


Figure 6.2 Biosynthetic origins of RED. The scheme shows the biosynthetic origin of the RED complex. The precursors are proline (red), serine (grey), glycine (orange) and acetates (blue). Two convergent pathways lead to 2-undecylpyrrole (**1**) and 4-methoxy-2,2´-bipyrrole-5-carboxyaldehyde (**2**). These two pyrroles are condensed on a late stage to give place to the linear tripyrrole undecylprodiginine (**3**), which subsequently can be cyclised to form *meta*-cycloheptylprodiginine (**4**). The latter two metabolites form the RED complex of *S. coelicolor*. Figure kindly provided by G. Challis (see also Cerdeño *et al.*, 2001).

Previous workers from Hodgson's laboratory have suggested that RED can act as a natural 'sink' for proline present in excess, e.g. as a consequence of loss of proline catabolic functions (Hood *et al.*, 1992). This suggestion arose from the observation that the *put*⁻ mutant *S. coelicolor* Pum7r, with a pleiotropic effect on proline degradation and proline uptake from the medium, overproduced RED (Hood *et al.*, 1992). Although this observation may have a bearing on the evolution of secondary metabolic pathways as what may had been the driving force for the appearance of this type of metabolic pathways (see Hodgson (2000) for an interesting discussion on the evolution of secondary metabolism) it is difficult to imagine that such an endeavour as it is to produce this compound is undertaken by the cell only for the purpose of contingent proline detoxification. A total of 33 kb of DNA are used to encode the twenty-three *red* genes that direct the synthesis of RED in *S. coelicolor*!

6.1.4 Insights from the *S. coelicolor* genome sequence into proline metabolism

The genome-sequencing project of *S. coelicolor* brought with it the discovery of a putative *proC* gene. A single candidate for *proC* whose product shares significant identity with other P5C-R enzymes was detected. This gene, i.e. SCO3337, lies far apart from the *proBXA* genes. Furthermore, it seems that *proC* is independently transcribed as judged from the functions predicted for the genes in its vicinity, which seem to be unrelated to proline metabolism, and from the orientation and intergenic distances between them (data not shown). Identification of this putative *proC* gene was encouraging since previous attempts in Hodgson's laboratory aimed at identifying this gene by complementation of eight *S. coelicolor* proline auxotrophs were unsuccessful (D. Hood and D.A. Hodgson, unpublished results). Moreover, the predicted amino acid sequence of SCO3337 includes a ten-residues sequence (i.e. TQKVAVLGTG) that matches with an N-terminal sequence obtained from a purified protein from *S. coelicolor* cultures, which was shown to have P5C-R activity (D. Hood and D.A. Hodgson, unpublished results).

Although this observation strongly suggests that SCO3337 is a *proC* gene, the fact that only one putative *proC* gene was identified, in contrast to the scenario described for the *trpE* paralogues in Chapter V, raises the question of why no proline auxotroph deficient in P5C-R has been isolated following traditional mutagenesis programs hitherto. Thus, the general aim of this Chapter was to characterise the function of this putative *proC* gene, with particular emphasis upon the effect that proline biosynthesis has on RED production. The discovery of a putative *proC* gene opened the possibility of constructing an artificial proline biosynthetic operon including the

complete set of *pro* genes. This would allow the study of the effect of precursor synthesis upon secondary metabolite production (see also Sections 6.1.3 and 6.2.5).

Analysis of the genome sequence also identified putative *put* genes encoding the enzymes P-DH and P5C-DH. Two different genes potentially encoding for the latter enzyme function were identified, namely, the genes SCO3835 and SCO5520. Based on two lines of evidence the SCO5520 gene was favoured over SCO3835 as to be the real *putB* gene. First, the best reciprocal matches of SCO5520 in *M. tuberculosis* (Rv1187) and *S. avermitilis* (SAV2723) are within a conserved cluster of genes, which interestingly includes several genes involved in the biosynthesis of branched-chain amino acids (see Figure 6.3). The fact that these genomic regions are conserved suggests that the cognate genes are involved in a fundamental metabolic process such as the degradation of amino acids for energy generation. Secondly, immediately upstream of SCO5520, and putative orthologues from *M. tuberculosis* and *S. avermitilis*, a gene with homology to *putA* can be detected. The annotators of the *S. coelicolor* genome sequence did not predict this function but a domain analysis of the protein product of SCO5519 showed that it includes a P-DH domain. Furthermore, the intergenic region between the putative *putA* (SCO5519) and *putB* (SCO5520) genes is of only 47 nucleotides, suggesting that they may be co-transcribed.

Although several other genes have been predicted to encode for amino acid oxidases, the physical link between *putA* (SCO5519) and *putB* (SCO5520) on the chromosome strongly suggests that this pair of genes are the proline catabolic genes. Further evidence supporting this prediction is the observation that an N-terminal sequence obtained from the protein that was shown to have P5C-DH activity by Smith *et*

al. (1995) matches with residues 173-182 of the product of SCO5520 (D.D.S. Smith, and D.A. Hodgson, unpublished results). An unanswered question is where on the chromosome, or within which of the PutA or PutB proteins, is the regulatory domain implied from the proline induction results of Smith *et al.* (1995). An analogous regulatory domain is part of the PutA protein in *E. coli*, which is a nice example of a moonlighting protein (see Section 1.1.2). This Chapter also reports on preliminary mutational experiments aimed at characterising the proline catabolic genes in *S. coelicolor*.

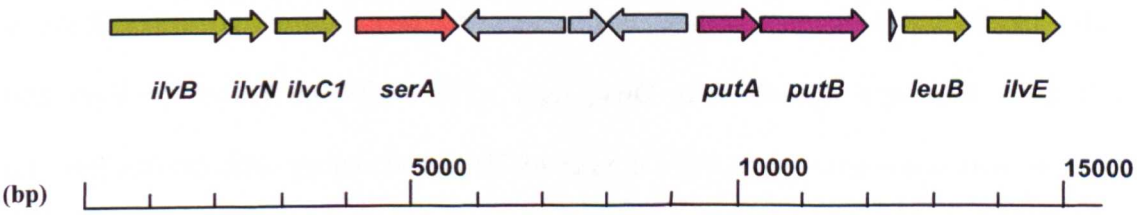


Figure 6.3 The *put* catabolic genes of *S. coelicolor*. The region where the putative *put* genes (magenta) are found in this organism is relatively conserved in *M. tuberculosis* and *S. avermitilis*, particularly in terms of the occurrence of the branched-chain amino acid biosynthetic genes *ilv* and *leu* (green). A single gene involved in serine biosynthesis (red) and genes of unknown function or unrelated to the metabolism of amino acids (grey) are also shown.

6.2 Results and Discussion

6.2.1 Proline biosynthesis: cloning of the *proC* homologue SCO3337

The first step that was undertaken in order to characterise the putative P5C-R activity of SCO3337 was to clone this gene in *E. coli*. The general strategy followed was similar to that described in previous chapters: cloning by PCR amplification using primers with restriction sites engineered at their 5' ends. The cosmid ScE7 was used as

template (Redenbach *et al.*, 1996). Various versions of *proC* were cloned (see Table 7.1 in Methods). Two promoterless versions of *proC* suitable for heterologous expression and purification of P5C-R, as well as for intrageneric complementation studies, were successfully cloned and confirmed by sequencing. The *proC* gene was cloned into the *Nde*I and *Bam*HI restriction sites of pET-3a to form pETproCSc. As judged by SDS-PAGE analysis overexpression of *proC* on pETproCSc in *E. coli* led to the production of an approximately 30 kDa-protein (data not shown), which agrees well with the predicted size of 29.8 kDa of the product of SCO3337. The putative *proC* gene was cleaved from this construct and sub-cloned into the *Streptomyces* expression vector pIJ4123 (Takano *et al.*, 1995) using the same restriction sites. The resulting construct, termed pFBP-1c, was used to assess the effect of a high *proC* gene dosage expressed from the thiostrepton-inducible promoter p_{tipA} (Holmes *et al.*, 1993) upon the production of RED (see Section 6.2.5). *Eco*RI and *Xho*I restriction sites were also engineered at the 5' and 3' ends of *proC*, respectively, and subsequently exploited to insert SCO3337 into the expression vector pGEX-4T-1. The resulting construct, termed pGEXproCSc, was used for complementation studies of an *E. coli* proline auxotroph deficient in P5C-R activity (see below Section 6.2.2).

The putative *proC* gene was also cloned including its native promoter. The first attempts included the use of the cloning vector pNEB193 from New England Biolabs and a *Bam*HI / *Bam*HI *proC* fragment. However, no positive clones were obtained using this approach after several attempts varying conditions such as the concentration of salt. This factor was modified because it is known that proline acts as an osmoprotectant in other bacteria (Kempf and Bremer, 1998). These experiments were done at an early

stage of this project suggesting that this negative result may have been due to the lack of cloning experience of the author rather than to a real biological reason. Eventually, the *proC* gene was cloned in *E. coli* although making use of cloning vectors from the pMTL family and different restriction sites, such that SCO3337 was cloned into the *Kpn*I (5') and *Pst*I (3') sites of the multi-cloning site of pMTL21. This resulted in the construct pMTL21proCSc, which was confirmed by sequencing. The putative *proC* gene was cleaved from pMTL21proCSc using the restriction enzymes *Pst*I and *Sph*I and recovered for further sub-cloning. Although there are two *Pst*I sites within the multi-cloning site of pMTL21 only the 5' end site relative to *proC* would be efficiently cleaved because *a priori* cleavage with *Sph*I removed enough nucleotides upstream the 3' end-*Pst*I site to hamper the efficiency of the digestion at this site (see Section 7.1.2 in Methods). The *Pst*I-*proC*-*Sph*I product was recovered and sub-cloned into the cloning vector pMTL23 using the same restriction sites leading to pMTL23proCSc. The sub-cloning steps were done in order to place the *proC* gene in the correct orientation within suitable restriction sites in order to be inserted into the unique *Pst*I and *Sph*I sites of pDHP14, which bears the *proBXA* genes (Hood *et al.*, 1992). This was successfully achieved after transformation of *S. lividans* TK24 protoplasts with the ligation mixtures, and a plasmid bearing the complete set of *pro* genes termed pFBP140 was isolated (see map in Figure 6.4). The correct orientation of *proC* relative to the *proBXA* genes within pFBP140 was confirmed by PCR amplification using primers specific to the backbone of pDHP14 and *proC* (see Section 7.1.2 in Methods). This construct was also used to assess the effect of proline synthesis on RED production (see Section 6.2.5)

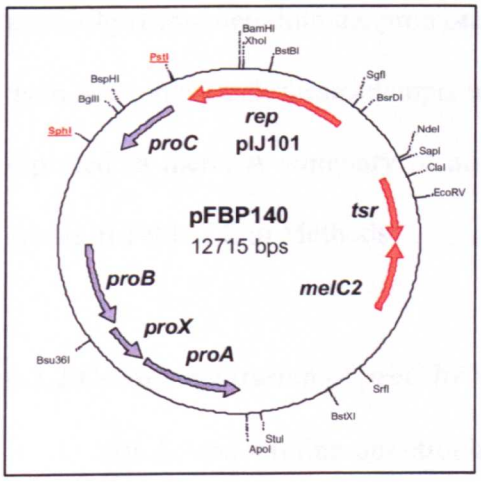


Figure 6.4 Map of pFBP140. The genes that are believed to be involved on the biosynthesis of proline are shown in light purple. Other genes from the backbone of the pIJ702 vector are also shown (red). The unique *Pst*I and *Sph*I sites present on pDHP14 that were exploited to insert the *proC* gene are shown in red and underlined. This construct was used in the experiments reported in Section 6.2.4.

The putative *proC* gene was also recovered from pMTL21proCSc after digestion with the restriction enzymes *Xho*I and *Hind*III. This *Xho*I-*proC*-*Hind*III fragment was sub-cloned into pMTL24 using the same restriction sites leading to pMTL24proCSc. Subsequently, a *Bam*HI-*proC*-*Bam*HI fragment was cleaved from pMTL24proCSc, which was used for sub-cloning SCO3337 into the *Streptomyces* cloning vector pIJ702 (Kieser *et al.*, 2000) making use of the *Bam*HI-compatible unique *Bgl*II site of this vector. Insertion into this site allowed screening of transformants by means of the pigmented phenotype conferred by the *melC* reporter gene, which is disrupted when a DNA fragment is inserted into the *Bgl*II site. The plasmid pFBP-2c (i.e. pIJ702 bearing SCO3337) was successfully obtained after transformation of *S. lividans* TK24 protoplasts with the ligation mixture and screening of colonies lacking black pigmentation. A final form of the putative *proC* gene suitable for promoter probe experiments including exclusively its native promoter plus the 5' end of SCO3337 was also obtained from most of the pMTLproCSc constructs by exploiting profitably a *Bgl*II site that lays just in the middle of *proC*. After restriction digestion of pMTL24proC with *Hind*III and *Bgl*II the 5'end of *proC* was obtained and sub-cloning of this *Hind*III-5' end

*proC-Bgl*II fragment into the promoter probe pIJ2842 was attempted. Unfortunately, just as in Section 5.2.2.2, these attempts were unsuccessful probably due to the same reasons reported in there. A summary of all the *proC* constructs reported in this Section can found in Table 7.2 in Methods.

6.2.2 Complementation of *proC* by SCO3337 in *E. coli*

An *E. coli* proline auxotroph that has been mapped to *proC* and therefore is expected to be deficient in P5C-R activity was obtained from the *E. coli* Genetic Stock Centre, University of Yale. This strain is *E. coli* χ342 (*proC29 metB1*; Hall, 1988; Serebrijski *et al.*, 1995). Only the genetic markers related to growth requirements of this strain are given here. This strain has been previously used in similar experiments including the *proC* gene cloned from the actinomycete *C. glutamicum* (Ankri *et al.*, 1996), suggesting that this bacterium's high GC-content genes can be transcribed and translated properly by *E. coli*. Transformants of χ342 bearing the construct pGEXproCSc were obtained after transformation of calcium-competent cells and selection on LB agar plus ampicillin. The ability of SCO3337 to complement the *proC* mutation of χ342 was tested in MA medium supplemented with the appropriate growth requirements (see Section 7.2 in Methods). It was found that χ342 proline auxotrophy was fully complemented by pGEXproCSc, suggesting that SCO3337 has P5C-R activity *in vivo*. However, in accordance with the data discussed in Section 1.1.5.1, complementation of *proC* by SCO3337 in *E. coli* could also be explained by a promiscuous 'secondary' activity of the product of the putative *proC* gene. Therefore,

complementation of *E. coli* χ342 by SCO3337 may be devoid of biological significance within the context of *S. coelicolor*.

6.2.3 Mutation of SCO3337 does not generates a proline auxotroph

To further characterise the function of SCO3337 *in vivo* this gene was inactivated using the technology REDIRECT® previously described (see Section 7.1.3.2 in Methods and Appendix 2). An *aac(3)IV* plus *oriT* marked mutant of SCO3337 was constructed in cosmid ScE7 and subsequently transferred by conjugation from *E. coli* to *S. coelicolor* M145. Two exconjugants that underwent double crossovers as judged from their kanamycin sensitivity were selected. The presence of the mutated allele and absence of the wild-type SCO3337 allele from the chromosome of these clones was confirmed by PCR amplification (Figure 6.5A). One of these strains was selected for further characterisation. Unexpectedly, when this mutant strain, termed WP101, was tested for its growth requirements in minimal medium it turned out to be prototrophic. It was rationalised that this result could be due to proline contamination of the plates. To rule out this possibility this experiment was repeated but incorporating the genetically marked strain *S. coelicolor* M124 (*argA1 proA1 cysD18*; Hopwood *et al.*, 1985 and references therein), which is a proline auxotroph. As it can be seen in Figure 6.5B the lack of growth of M124 in minimal medium plates supplemented with arginine and cysteine shows that no proline was present in the medium used. In contrast, the WP101 mutant strain can grow without any apparent effect due to inactivation of SCO3337, as compared to the parental strain M145.

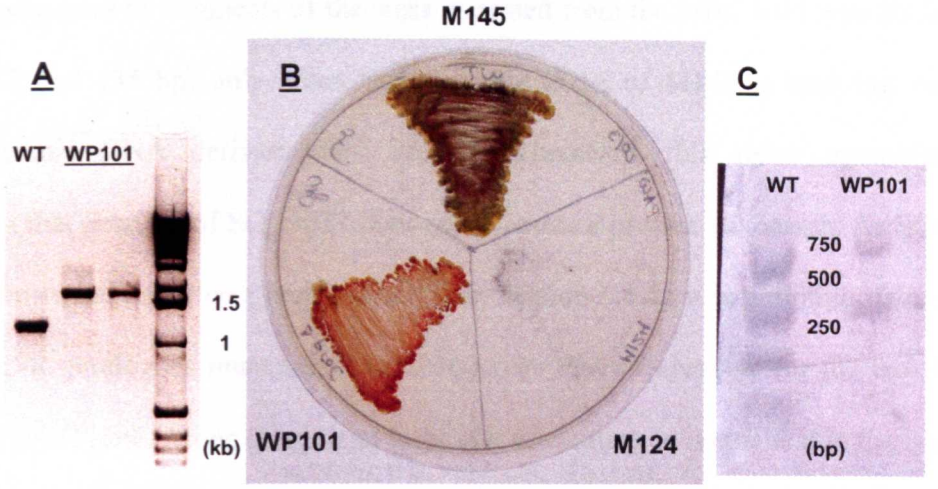


Figure 6.5 The mutant strain WP101 is not a proline auxotroph. **A** PCR amplification of the mutated SCO3337 allele (1.8 kb) and WT allele (1.3 kb) from chromosomal DNA of M145 (first lane) and two clones of WP101 (next two lanes) as templates using the same primers. **B** The growth requirements of *S. coelicolor* WP101 were tested on minimal medium plus arginine and cysteine. The proline auxotroph M124 cannot grow in this medium demonstrating that there is no proline on the plates used for this experiment. **C** Southern hybridisation of chromosal DNA from M145 and WP101 digested with *Sau*3A1 using as probe an internal fragment of SCO3337. The expected fragments for the WT allele (first lane: 499, 243, 162 and 135 bp) can be seen as four bands. These bands are absent from chromosomal DNA obtained from WP101, although two other bands of similar size are present (second lane). The approximate position of the size markers is given between the two lanes (750, 500, 250 bp). Refer to text for further explanation.

Although genome duplication on some *S. coelicolor* strains has been recently reported, as revealed by microarray experiments, there is no evidence suggesting that M145 bears duplications within its chromosome (C. Kao, personal communication). Nevertheless, the risk of DNA amplification is always present (see also Section 1.1.2), which can only be ruled out by Southern hybridisation experiments. It is possible that either duplication of a genomic region on M145 or DNA amplification while constructing the SCO3337 mutant strain may account for the prototrophic phenotype of WP101. The Southern hybridisation experiment shown in Figure 6.5C was done using a digoxigenin-labelled 570 bp probe generated by PCR amplification of an internal fragment of SCO3337 using the primers GCGCGGACGCCGTGGTT and CCGCCAAG GCCGCCGACAC as described in Section 7.1.4 in Methods. As it can be seen this

probe hybridises to fragments of the sizes expected from the *proC* wild type allele (499, 243, 162 and 135 bp) only when chromosomal DNA of M145 is used but not with chromosomal DNA derived from WP101. Therefore, this result unambiguously confirms that deletion of SCO3337 does not generate a proline auxotroph. As SCO3337 can complement an *E. coli proC* mutant (see Section 6.2.2) it is tempting to speculate that P5C-R paralogues must exist in *S. coelicolor* that compensate for the lack of this putative *proC* gene. Along this line of logic it is interesting to notice that in the Southern hybridisation experiment other bands that do not correspond to SCO3337 are present when DNA from the mutant strain WP101 is used (Figure 6.5C). This result would explain why no proline auxotrophs deficient in P5C-R have been obtained so far using traditional 'undirected' mutagenesis methods. Therefore, further experimentation is required if the function of SCO3337 is to be unambiguously established as a P5C-R involved in proline biosynthesis.

6.2.4 Cross-feeding experiments of *S. coelicolor* M124

In addition to proline auxotrophy one of the typical phenotypes expected from a *proC* mutant is the ability to excrete P5C into the medium. This situation has been previously reported for various *proC* minus mutants from different organisms including *E. coli* (Serebrijski *et al.*, 1995) and the actinomycete *C. glutamicum* (Ankri *et al.*, 1996). It is generally accepted that the basis of this observation is that P5C is toxic to the cell (Arst *et al.*, 1981; Ankri *et al.*, 1996). A way of removing the excess P5C is by secreting it into the medium. This phenotype can be tested indirectly by using mutants deficient in early steps on the biosynthesis of proline, i.e. mutants of either *proB* or *proA*

(Serebrijski *et al.*, 1995; Ankri *et al.*, 1996). In this system the *proBA* mutants are grown next to the *proC* mutants in minimal medium lacking any source of proline. Growth of the *proBA* minus mutant is then recorded, a situation that is expected to occur only if the *proC* minus strain is supporting the growth of the *proBA* mutant *via* P5C excreted into the medium.

This experiment was done using as potential donors of P5C the *S. coelicolor* strains WP101 and M145, in addition to their transformants with pDHP14 and pIJ702. The use of transformants bearing the pDHP14 plasmid was aimed at increasing the amount of P5C synthesised whereas the pIJ702 transformants served as control. The P5C recipient strain deficient for the early steps on the biosynthesis of proline was *S. coelicolor* M124, the genetically marked strain used in Section 6.2.3. The minimal medium plates contained the appropriate growth requirements for M124 other than proline. No cross feeding could be detected between any of the strains tested as judged from the complete lack of growth of M124 (data not shown). Only very slight levels of promotion could be seen similar to those shown at the centre of the plate of Figure 6.5B, although they were not particularly associated with any of the strains. It was thought then that the lack of selection for the plasmids might result in their lost and therefore P5C could not be produce in enough quantities. Therefore, a M124 transformant bearing pIJ702 was obtained, which was used in similar experiments but with thiostrepton added into the medium. Again, no cross feeding could be detected (data not shown).

This result may be explained on the grounds of the observation that *S. coelicolor* produces RED, a secondary metabolite that uses proline as precursor for synthesis (see also Sections 1.3.2 and 6.1.3). It has been previously suggested that in this organism

proline can be shunted into secondary metabolism when present in excess (Hood *et al.*, 1992; see also Sections 6.1.3 and 6.2.5). Although there is no evidence for the *direct* incorporation of P5C, or any other proline metabolite, into the structure of RED it is likely that the product of *redM*, the adenylation domain with substrate specificity for proline of the non-ribosomal peptide synthetase responsible for the condensation of the activated form of this amino acid (i.e. pyrrole-2-carboxylate; Thomas *et al.*, 2002), may show specificity for P5C when present in excess. Therefore, the cross-feeding phenotype characteristic of P5C overproduction was tested within an absolute RED minus background, i.e. on the *redD* minus regulatory mutant *S. coelicolor* JF1 (Kieser *et al.*, 2000). The regulatory nature of the *red* mutation of JF1 makes this strain incapable of producing RED or any of its intermediary metabolites. JF1 protoplasts were transformed with the plasmids pDHP14 and pFBP140, which should lead to overproduction of P5C and proline, respectively (see also Section 6.2.5). The pFBP140 construct was used as a 'positive' control. Once again, it was found that no cross feeding of M124, this time from JF1 (pDHP14) and JF1 (pFBP140), could be detected. This negative result can be explained in three different ways. It could be that *S. coelicolor* lacks the means for excreting P5C and proline into the medium or that these metabolites are diverted into another hitherto unidentified secondary metabolic pathway that uses these compounds as precursors. This latter possibility seems unlikely as judged from the results reported in the following Section. Alternatively, it may be that an excess of P5C and proline are inducing the catabolic enzymes P5C-DH and P-DH that in turn convert any excess of these metabolites into glutamate before they can be accumulated and excreted. On the basis of the regulatory results reported by Smith *et al.*, (1995) this latter

possibility seems to be the most likely explanation for these results. If this is the case cross-feeding experiments as a means for characterising *S. coelicolor pro* and *put* mutants are futile.

6.2.5 Effect of proline synthesis upon RED biosynthesis

The experiments reported in this Section were motivated by the results of Hood *et al.* (1992) on the overproduction of RED by the *put* minus mutant *S. coelicolor* Pum7r (see Section 6.1.3). These co-workers proposed that endogenous proline in excess, as a consequence of constitutive expression of proline biosynthesis and lack of proline catabolism and transport, was shunted into the secondary metabolic pathway of RED (see Figure 6.2). An alternative to the lack of proline degradation for generating an excess of proline within the cell is by increasing the rate of proline biosynthesis, e.g. *via* artificial expression of the whole set of *pro* genes from a high-copy number plasmid. With the construct pFBP140 on hand (see Section 6.2.1 and Figure 6.4), which contains the *proCXBA* genes, this alternative became possible. Furthermore, study of the effect of the different *pro* genes upon RED production was also possible by using the plasmids pDHP14 (*proBXA*), pFBP-1c and pFBP-2c (*proC*). All these constructs are pIJ101 derivatives, which is a high-copy number plasmid, ensuring a high gene dosage.

Previous data from Hodgson's laboratory suggested that pDHP14 had a significant, although transitory effect upon RED production (D.W. Hood and D.A. Hodgson, unpublished results). It was found that pDHP14 and pIJ364 *S. coelicolor* J802 transformants obtained simultaneously on the same R2YE-containing plate, could be screened by eye. Although pDHP14 is not a pIJ364 derivative (Kieser *et al.*, 1982), this

latter plasmid was used as control for this experiment because production of melanin by the products of the *melC* genes encoded by pIJ702 interfered with the screening. The J802 (pDHP14) transformants appeared to form bright red colonies whereas the pIJ364 derivatives did not. However, the bright-red phenotype disappeared with time as the transformants were transferred into plates containing fresh R2YE medium, a phenomenon that was attributed to a hitherto unidentified mechanism of "physiological adaptation" (D.A. Hodgson, personal communication). The J802 strain was originally used for these experiments because it is a SCP1^{NF} strain that was isolated in Hodgson's laboratory. This means that J802 has the SCP1 linear plasmid inserted into the chromosome. Since the *mmy* genes, which direct the synthesis of methylenomycin, another secondary metabolite produced by A3(2) (see also Section 1.3.2), are encoded by SCP1, the pool of precursors for synthesis of secondary metabolites is expected to be compromised when SCP1 is present. This in turn has been previously suggested to reduce the wild type titres of RED and ACT, which was an advantage when determination of RED had to be done using a spectrophotometric assay (D.A. Hodgson, unpublished observations).

A similar experiment (see above) but including pFBP140 was done. The same J802 genetic background was used to allow comparison between this study and previous results. A triple co-transformation of J802 protoplasts was done and in accordance with previous results the transformants were distinguishable between them by eye (Figure 6.6A). Screening of the transformants by eye and subsequent analysis of the isolated strains by recovering and analysing the plasmids by restriction digestion agreed well overall (data not shown). Although both pDHP14 and pFBP140 elicited a burst of RED

overproduction just after transformation, it was found that only transformants bearing pFBP140 retained this phenotype after prolonged periods of incubation or transfer into plates containing fresh R2YE medium. The typical phenotype associated with RED overproduction, i.e. lack of sporulation and development of a green-metallic sheen after red pigmentation was recorded consistently for *S. coelicolor* J802 (pFBP140) (see Figure 6.6B). Therefore, it was concluded that an excess of proline (assuming that SCO3337 is in fact a *proC* gene), and not P5C, is required in order to elicit a sustained burst of RED overproduction. This result is in accordance with the fact that RedM has been shown to have specificity for L-proline (Thomas *et al.*, 2002; see also Section 6.2.4), although the specificity of RedM for P5C was not tested.

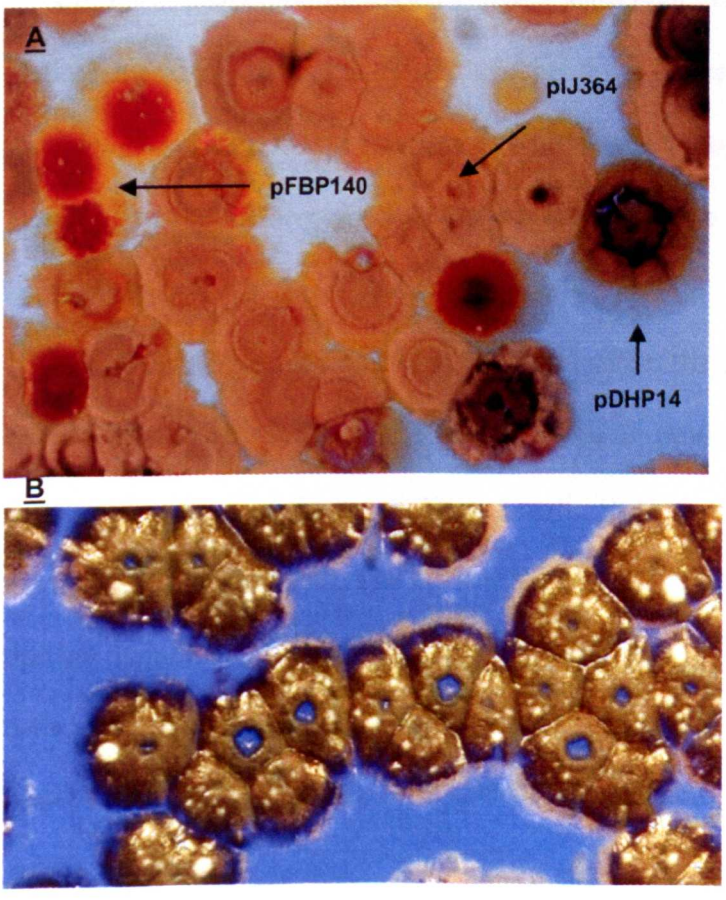


Figure 6.6 Phenotype of *pro* genes overexpression. **A** The picture shows the results of a triple co-transformation experiment with pIJ364, pDHP14 and pFBP140. The three different strains can be screened by eye. **B** *S. coelicolor* J802 (pFBP140) was found to overproduce RED, even after transfer into fresh medium and prolonged periods of incubation. The characteristic phenotype of RED overproduction, i.e. lack of sporulation and appearance of a green metallic sheen after red pigmentation, can be seen in the picture.

Transformants of J802 bearing pFBP-1c and pFBP-2c were also obtained. Interestingly, as judged by eye, a burst of RED overproduction was only detectable after transformation of J802 with pFBP-1c (data not shown), which bears a version of *proC* under the control of the thiostrepton-inducible promoter. Although this phenotype was not retained after transfer of J802 (pFBP-1c) into plates containing fresh medium, this observation is circumstantial evidence suggesting that *proC* may be regulated at the transcription level by an unknown effector (see also Section 6.1.1). While on pFBP-2c the *proC* gene is under the control of its native promoter, and probably transcription is repressed as suggested by the lack of RED overproduction, expression of *proC* from pFBP-1c is induced by thiostrepton and therefore is autonomous. The disappearance of RED overproduction after transfer into fresh medium may be related to the physiological adaptation discussed above. Please note that the comparison of RED titres elicited by pFBP-1c and pFBP-2c can be done because both constructs are pIJ101 derivatives and therefore are expected to have similar copy-numbers. However, it could also be that the differences detected are related to the strength of the two promoters compared.

As an attempt to quantitatively characterise the phenotypes associated with the overexpression of *pro* genes in terms of RED production, an HPLC method was developed with the help of Dr Greg Challis from the Department of Chemistry (see Section 7.3.4 in Methods). The cells were grown by duplicates on R2YE medium over dialysis membranes as described previously (Kieser *et al.*, 2000), which allowed the cells to be recovered from the agar surface. The strains studied were J802 (pIJ702) as control and J802 (pFBP-2c), J802 (pDHP14) and J802 (pFBP140), such that the effect

of the *proC*, *proBXA* and *proCBXA* genes independently transcribed from their native promoters was investigated. However, despite the fact that the characteristic phenotype for RED overproduction was recorded for J802 (pFBP140) as before, for some unknown reason this observation was not reflected by the quantitative HPLC data. Although the *proCBXA* and *proBXA* treatments appeared to produce approximately 20% more RED than the control after five days, this situation is reverted in the following days (Figure 6.7). Therefore, it was concluded that all four strains produced similar titres of RED, which interestingly decreased on time. It was thought then that perhaps the plasmids had undergone genetic rearrangements or were lost. This possibility was ruled out after recovering the plasmids and analysis by restriction digestion (data not shown). However, minor mutations that can escape detection by this method cannot be ruled out. Again, this observation may have to do with some sort of physiological adaptation of which nothing is known. Nevertheless, the general correlation between overproduction of RED, as judged by eye, and expression of the putative *proC* gene from pFBP140 and pFBP-1c, can be interpreted as circumstantial evidence supporting the view that SCO3337 encodes a P5C-R enzyme.

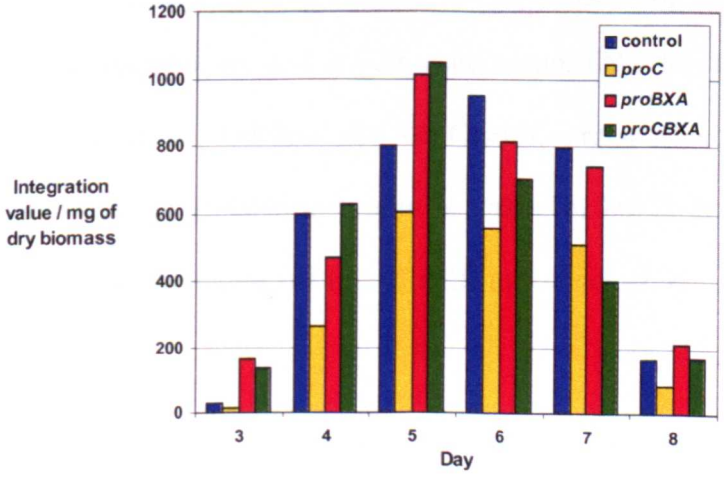


Figure 6.7 Effect of *pro* genes overexpression upon RED biosynthesis. The graph shows the relative production of RED by the four different strains, expressed as the integration values of the peaks detected by HPLC per mg of dry biomass. Each data point was obtained as the mean of duplicate experiments. The production of RED was followed on time and data from days three to eight is shown.

Another interesting observation that was immediately noticed from the HPLC data was that RED, which is known to be a complex of at least two different metabolites (Wassermann *et al.*, 1966; see also Section 6.1.3 and Figure 6.2), showed an interesting biosynthetic profile. It was found that the peak eluting just before eight minutes prevailed from day five of incubation whilst the second peak, eluting just after eight minutes, was the major metabolite after prolonged periods of incubation (see Figure 6.8). Moreover, this profile suggested that the kinetics of the biosynthesis of these metabolites were interrelated, i.e. as one of the peaks decreased the other peak increased. This observation was found to be independent of the *pro* genes overexpressed. The biosynthetic relationship between these two metabolites is shown in Figure 6.8 as a ratio of the integration values of the first peak (A) with respect to the second peak (B). Furthermore, on the grounds of the UV spectra obtained from these two peaks (data not shown) and previous reports (Wassermann *et al.*, 1966) it was possible to figure out that peak A corresponded to the cyclic metabolite *meta*-cycloheptylprodiginine (compound **4** in Figure 6.2) whilst the second peak was the linear tripyrrole undecylprodiginine (compound **3** in Figure 6.2). Further investigation by G.L. Challis and co-workers has unambiguously confirmed the chemical identity of these metabolites and a *red* gene responsible for the cyclation reaction has been identified (O. Odulate and G. Challis, unpublished results). The meaning of this biosynthetic metabolite profile is unknown but it is tempting to speculate that it is related to the biological role of RED in *S. coelicolor*.

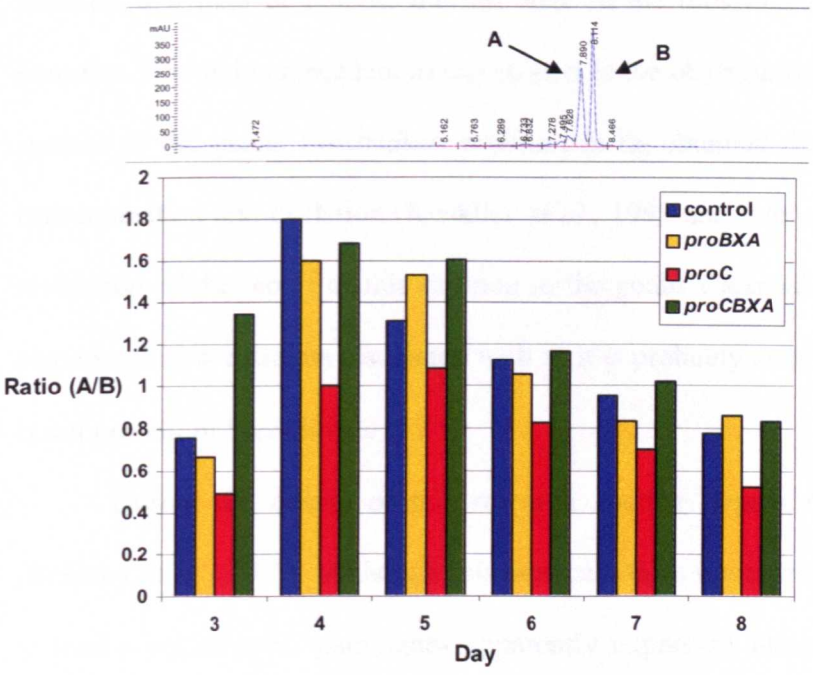


Figure 6.8 Biosynthetic profile of the linear and cyclic forms of RED. A typical HPLC chromatogram of RED is shown at the top of the Figure. The graph shows the biosynthetic relationship between the cyclic (peak A) and linear (peak B) forms of RED. The A to B ratio was calculated using the integration values of the peaks. When the main peak is A (i.e. the cyclic form) the A/B ratio is larger than the unit (e.g. day 5). In contrast, when the main peak is B (i.e. the linear form) the A/B ratio is smaller than the unit (e.g. day 8).

6.2.6 Identification of potential *proC* paralogues compensating for the deficiency of the P5C reductase activity of SCO3337

One possibility for the phenotype of WP101 is the existence of alternative pathways that bypass the common proline biosynthetic pathway from glutamate (see Figure 6.1). A bypass of the early steps of the biosynthesis of proline has been previously seen in other organisms. It was noticed very early that GK and GS-DH could be bypassed through synthesis of P5C from the arginine catabolite ornithine *via* the enzyme activity ornithine aminotransferase (EC 2.6.1.13; Cunin *et al.*, 1986; Kenklies *et al.*, 1999 and references therein). However, it is known that in streptomycetes arginine is degraded *via* γ-guanidinobutyrate and therefore no P5C is produced (Hodgson, 2000 and references therein). Recently, another possibility to this bypass has been reported on which complementation of *proA* is achieved by the *asd* gene of *E. coli* involved in lysine biosynthesis, although only under certain conditions (Serebrijski *et al.*, 1995). In

contrast, a bypass of P5C-R, the last step on the biosynthesis of proline, is not that common. The only antecedent in this respect is the observation that in certain anaerobic species of the genus *Clostridium* proline can be obtained directly from ornithine *via* transamination and cyclation (Kenklies *et al.*, 1999 and references therein). Due to the restriction of the report of this reaction to the genus *Clostridium*, and to the anaerobic physiological constraints associated with it, it is probably safe to state that this pathway is not present in *S. coelicolor*.

During the course of this research another bypass for P5C-R was reported (Belitsky *et al.*, 2001). Although this new case does not represent a novel pathway but instead a set of *proC* paralogues apparently expressed under different conditions, it represents an interesting possibility. These researchers reported that after analysis of the genome sequences of *Bacillus subtilis* and relatives that have been fully sequenced they identified at least four genes that could potentially encode P5C-R enzymes. This observation acquires its relevance when it is appreciated that no *proC* minus mutants from these organisms had been isolated up to that report, a situation that is shared with *S. coelicolor* and *S. lividans*. After multiple gene deletions from the chromosome of *B. subtillis* targeted at the identified putative *proC* genes Belitsky and co-workers (2001) came to the conclusion that deletion of three of the *proC* paralogues led to a proline auxotrophic phenotype. On the basis of this observation another explanation for the failure of detecting the typical phenotypes associated with a *proC* minus genotype in WP101 could be that there is at least one other *proC* paralogue on the chromosome of *S. coelicolor*. This possibility is supported by the observation that a probe generated from an internal fragment of SCO3337 hybridises to other genomic regions not associated

with SCO3337 (see Section 6.2.3 and Figure 6.5C), suggesting that the chromosome of *S. coelicolor* includes other DNA sequences homologous to SCO3337. It is tempting therefore to speculate that the extra bands detected correspond to putative *proC* paralogue(s).

BLASTP searches were done using as probes the *B. subtilis proC* products, as well as the product of SCO3337, with the hope of detecting the putative ProC paralogues of *S. coelicolor*. Only the gene product of SCO3337 was identified. Although it is predicted that the putative *proC* paralogues share some level of similarity with SCO3337, these similarities may be related to specific domains or signature motifs not detectable by standard BLASTP searches. It was therefore proposed that PSI-BLAST searches might unveil the putative ProC paralogues. After performing these searches it was found that the product of SCO3337 is homologous to the pair of proteins encoded by SCO5514 and SCO7154, which are annotated as *ilvC1* and *ilvC2*, respectively.

The *ilvC* gene in other organisms is involved in the biosynthesis of the branched-chain amino acids valine and isoleucine/leucine (reviewed by Dumas *et al.*, 2001). The product of *ilvC* is a bifunctional enzyme called acetohydroxy acid isomeroreductase (AHIR, EC1.1.1.86; also known as ketol-acid reductoisomerase, KARI). Since AHIR is involved in the biosynthesis of both valine and isoleucine/leucine, the substrates of this enzyme can be either 2-acetolactate (AL) or 2-aceto-2-hydroxybutyrate (AHB), respectively. The broad-substrate specificity of AHIR for AL and AHB, which differ by one methyl group, has been shown to be possible due to a hydrophobic pocket that can accommodate variable length non-polar chains (Biou *et al.*, 1997), suggesting that AHIR

is naturally prone to have promiscuous activities. Despite this observation it should be noted that the reaction accomplished by AHIR is a very complex two-step process. It involves an alkyl migration that is strictly dependent on Mg^{2+} ions accounting for the isomerisation of the substrates, followed by a reduction of the intermediaries in a NADPH-dependent fashion aided by a divalent cation (Dumas *et al.*, 1995; Biou *et al.*, 1997; Dumas *et al.*, 2001). Although the bifunctional AHIR is unique, it has been previously reported that the N-terminal domain of this enzyme, from both plants (class II; Biou *et al.*, 1997) and bacteria (class I; Jun-Ahn *et al.*, 2003), includes a reductase domain member of the superfamily of oxidoreductases that show the typical Rossman fold capable of binding a nucleotide cofactor NADPH (Todd *et al.*, 2001). The characteristic signature motif for binding of the diphosphate bridge of this nucleotide, i.e. GXGXX(G/A)XXX(G/A), can be detected in the N-terminus of AHIR within the so-called region I (Dumas *et al.*, 1995; see also Figure 6.9).

In order to learn more about the putative similarities between P5C-R and AHIR, and the products of SCO3337, SCO5514 and SCO7154, a sequence alignment was constructed (Figure 6.9). In addition to the *S. coelicolor* proteins the sequences of several other bacterial P5C-Rs, as well as the sequence of AHIR from *P. aeruginosa* (class I; Jun-Ahn *et al.*, 2003), were included. Exclusion of the plant-like AHIR from the alignment was done because this AHIR shows at least two extra domains with respect to the bacterial enzyme (Jun-Ahn *et al.*, 2003). Nevertheless, the important catalytic residues and the overall structure between these enzymes are highly conserved making possible the extrapolation of data from the well-characterised AHIR from spinach (Dumas *et al.*, 2001; Bious *et al.*, 1997; Dumas *et al.*, 1995) to the *Pseudomonas*

enzyme (Jun-Ahn *et al.*, 2003). As can be seen from the alignment the similarities between P5C-R and AHIR are limited to residues within region I (Figure 6.9). Dumas *et al.* (1995) discovered five conserved regions on the sequences of AHIR to which an extra region has very recently been added (region I′; Jun-Ahn *et al.*, 2003). Region I corresponds to the NADP-binding motif characteristic of this family of reductases (see above). It seems therefore that the PSI-BLAST searches identified the products of the two putative *ilvC* genes on the basis of the conserved region I. Unfortunately, nothing is known about the enzymology accomplished by P5C-R in mechanistic terms excluding the possibility of comparing the catalytic residues between both sets of sequences. Nevertheless, based on the experience gained from Chapters II and III it is possible that analogous enzymes that share very low similarity at the sequence level could be functionally and structurally related.

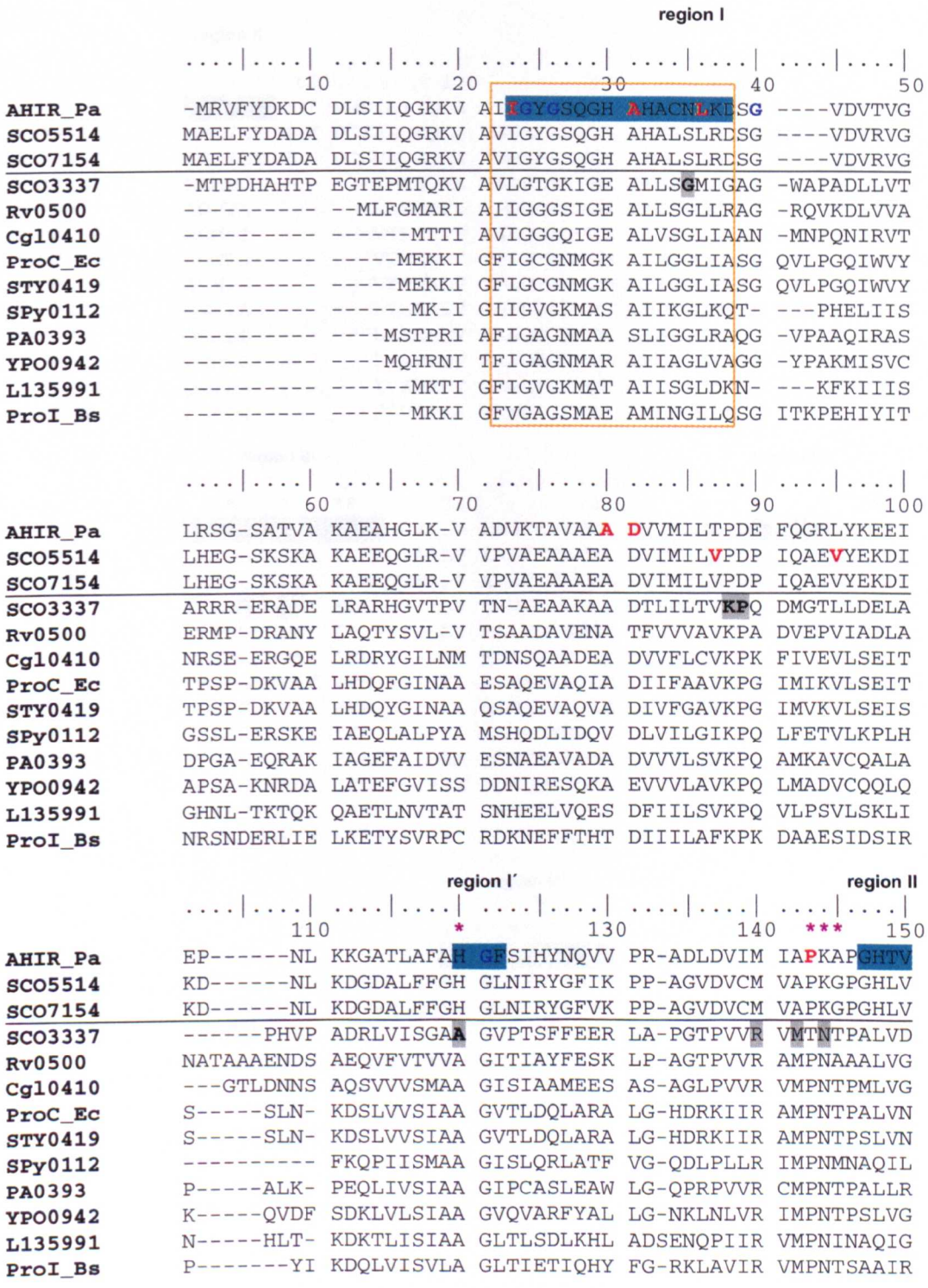


Figure 6.9 Sequence alignment of P5C-Rs and AHIRs. The top three sequences are AHIR enzymes including AHIR_Pa (*P. aeruginosa*). The sequences below the line are P5C-Rs: SCO3337, Rv0500 (*M. tuberculosis*), Cgl0410, (*C. glutamicum*), ProC_Ec (*E. coli*), STY0419 (*Salmonella*), SPy0112 (*Streptococcus*), PA0393 (*Pseudomonas*), YP00942 (*Yersinia*), L135991 (*Lactococcus*) and ProI_Bs (*B.subtilis*). The conserved regions of AHIR are shown in blue and the NADP-binding motif is highlighted with an orange box. *(Continued on next page)*

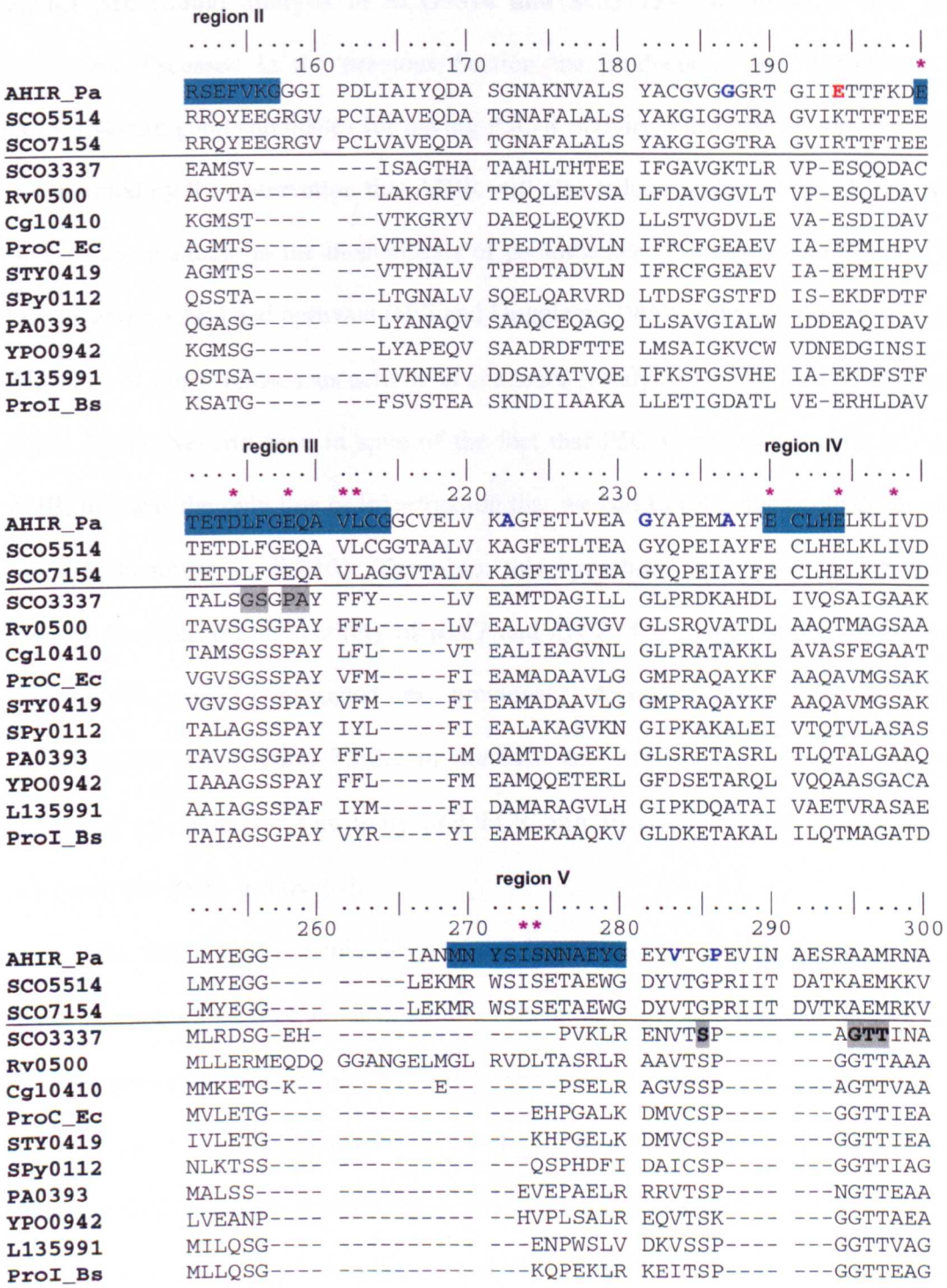


Figure 6.9 *(continues from previous page)*. The invariant residues across all the sequences are shown in bold blue in AHIR_Pa and the highly conserved residues in bold red. Residues that are invariant in P5C-R are shown in SCO3337 in gray and bold. The residues that are known to be important for the activity of AHIR are shown with a magenta asterisk. The last 60-70 residues of the alignment were removed since no conserved residues could be found. Refer to text for further explanation and discussion.

6.2.6.1 Mutational analysis of SCO5514 and SCO7154, the putative *ilvC* genes

As discussed in the previous Section the products of *ilvC* (SCO5514 and SCO7154) are good candidates for having P5C-R activity. Furthermore, this possibility is supported by the observation that AHIR will also reduce analogues of AL and AHB, such as ketopantoate in the biosynthesis of pantothenate (Primerano and Burns, 1983; Dumas *et al.*, 1995) and pyruvate (Hill and Duggleby, 1999), although the physiological relevance of this promiscuous activity is unknown (Frodyma and Downs, 1998; Dumas *et al.*, 2001). Nevertheless, in spite of the fact that P5C is not an analogue of AL or AHB, this was the only line of investigation that we had to solve the conundrum of the lack of phenotype of WP101. Therefore, this hypothesis was tested by mutational analysis by replacing the entirety of *ilvC1* and *ilvC2*. The mutant strains discussed in this section were constructed as previously described using the technology REDIRECT® (see Section 7.1.3.2 in Methods and Appendix 2). The mutants were confirmed by antibiotic sensitivity and PCR amplification targeted at the flanking regions of the genes inactivated.

The *ilvC2* gene (cosmid Sc9A4) was inactivated by replacement with the scar sequence as reported for the construction of the mutant strain WH101 in Section 2.2.3. This unmarked strain, called WILV101, showed no phenotype and although the construction was only confirmed by PCR (data not shown) further experiments involving double mutants that are reported below suggest that the construct is correct. In a similar way a *proC* scar unmarked mutant, termed WP102, was isolated which had the same phenotype as WP101. On the other hand, the *ilvC1* gene (cosmid Sc8D9), which is part of the cluster of genes shown in Figure 6.3, was inactivated after its replacement

with the *aac(3)IV* plus *oriT* cassette of pIJ773, resulting in an apramycin-marked strain. This strain was called WILV102 and in contrast to WILV101, WP101 and WP102 its growth-rate turned out to be slightly diminished on minimal medium. Addition of isoleucine and valine into the medium promoted the growth of WILV102 (data not shown). Therefore, this strain is a *bradytroph* for the branched-chain amino acids. This observation is not unexpected since it is likely that *ilvC1* is mainly involved on the biosynthesis of branched-chain amino acids as judged from its genomic position within *ilv* and *leu* genes (see Figure 6.3). Furthermore, this genomic organisation is conserved in *S. avermitilis* (De Rossi *et al.*, 1995; Ikeda *et al.*, 2003). In contrast, *ilvC2* lies in the right-hand side arm of the chromosome of *S. coelicolor* and therefore it is likely that its function may not be essential (see Section 1.3.2.1).

Double mutants of some of the possible combinations were prepared. A double *ilvC1* and *ilvC2* mutant, termed WILV103, was constructed by making use of the apramycin-marked Sc8D9 cosmid (*ΔilvC1::aac(3)IV*) and the unmarked *ilvC2* mutant WILV101. The resulting mutant, termed WILV103, was checked by PCR and its growth requirements were scored on minimal medium. WILV103 turned out to be auxotrophic for the branched-chain amino acids, although two observations that cannot be explained solely on the basis of growth requirements were recorded. First, it was observed that although WILV103 underwent normal development this double mutant showed a slower growth-rate not only on minimal medium but also on SFM and R2YE, media that are expected to be rich in these amino acids. This growth defect could not be completely restored by supplementation of the medium with a mixture of the branched-chain amino acids although it clearly alleviated to a certain degree the growth of the WILV103. This

could be due to a polar effect of the marked *ilvC1* mutation on the downstream *serA* gene (see gene organisation in Figure 6.3). The *serA* gene is predicted to encode for the enzyme 3-phosphoglycerate dehydrogenase (EC 1.1.1.95), which is involved in the biosynthesis of serine. However, addition of serine to the medium did not alleviate the slow growth-rate phenotype of WILV103. Moreover, the absence of this phenotype from the strain WILV102 suggests that in WILV103 a polar effect is not responsible for the slow-growing phenotype. Another possibility is that accumulation of AL and AHB due to the lack of AHIR activity is toxic to the cell. These metabolites would only build up if both AHIR activities were abolished. This suggestion has been previously made in *E. coli* and the putative toxic effect explained on the grounds of intracellular acidification due to accumulation of pyruvate and α-ketobutyrate (LaRossa *et al.*, 1987; LaRossa and Van Dyke, 1987; Van Dyke *et al.*, 1998). Secondly, when WILV103 was placed on minimal medium for prolonged periods very sick colonies started to appear. This colonies could be the result of *proC* compensating albeit deficiently for the lack of AHIR activity. It would be interesting to follow-up and characterise one of these slow-growing suppressants.

A double mutant of *proC* and *ilvC1* was constructed. This mutant strain was constructed using as background WP102, i.e. the scar mutant of SCO3337, to which the apramycin-marked Sc8D9 cosmid (*ΔilvC1::aac(3)IV*) was transferred by conjugation. The resulting mutant, termed WILVP101, was checked by PCR and its growth requirements were scored on minimal medium. It was found that the phenotype of this mutant was identical to the phenotype of the previously isolated mutant WILV102, which is a bradytroph for isoleucine and valine. This result is circumstantial evidence

that no complementation occurs between the putative AHIR and P5C-R enzymes of *S. coelicolor*, although the product of *ilvC2* may play a role in this phenotype. Unfortunately, a double *ilvC2* and *proC* double mutant has not been isolated.

A triple mutant was constructed making use of the *vph* plus *oriT* cassette of pIJ780, which confers resistance to the antibiotic viomycin (see Section 7.1.3.2 and Appendix 2). The use of a second marker different to apramycin avoided the need for a double scar mutant and allowed selection while an apramycin marker is already present on the chromosome. This triple mutant was recently isolated in the laboratory of Dr Greg Challis, Department of Chemistry, where the author of this thesis is currently conducting postdoctoral research. The mutant strain was termed WILVP102 and turned out to be a tight auxotroph. At the time of submission of this thesis it was known that growth of WILVP102 was not supported by minimal medium supplemented with valine/isoleucine, proline and serine. However, further experimentation revealed that WILVP102 grows on supplemented minimal medium (Kieser *et al.*, 2000), which contains Bacto Difco casaminoacids. Therefore, the growth requirements of WILVP102 are limited to amino acids implying that it should not be a major challenge to discover the amino acids required by this strain for growth on minimal medium. Nevertheless, irrespective of the identity of the amino acids required by WILVP102 the phenotype of this strain suggests that after all there is complementation between the products of *proC*, *ilvC1* and *ilvC2* implying extensive overlapping activities and gene duplication in *S. coelicolor*. This interesting result deserves further investigation.

6.2.7 Preliminary mutational analysis of *putB* (SCO5520) and *putA* (SCO5519)

The main motivation for the mutational analysis of the *put* genes was to confirm whether the results obtained by Hood *et al.*, (1992) were related to the *put* genes. For this purpose it was optimal to construct the *put* mutants on the same genetic background as Pum7r, i.e. *S. coelicolor* 802. This in addition would have allowed comparison of RED titres produced by the two types of putative proline accumulating strains, i.e. catabolic and biosynthetic. Apramycin-marked mutants for *putA* (SCO5519) and *putB* (SCO5520) were independently constructed in cosmid Sc8D9 as previously described (see Section 7.1.3.2 and Appendix 2). In addition, the Sc8D9 cosmids on which *putB* and *putA* were replaced were used to generate unmarked-scar mutants. This led to different combinations of single and double *put* mutant cosmids (see Table 7.4 in Methods). The Sc8D9 mutant cosmids bearing marked replacements of the *putA* and *putB* genes were transferred to *S. coelicolor* J802 as previously described. Colonies that were apramycin resistant and kanamycin sensitive were selected and analysed by PCR amplification. Unexpectedly, it was found that two PCR products were obtained, whose sizes agreed well with the mutated *and* the wild type alleles (data not shown). This negative result was confirmed for the putative *S. coelicolor* J802 *ΔputB::aac(3)IV* mutant by Southern hybridisation analysis (data not shown; see Section 7.1.4 in Methods). This result is at odds with the observation that the exconjugants isolated were apramycin resistant and kanamycin sensitive.

Different lines of logic can be followed in order to try to explain this conundrum. It is likely that genome duplication events of which we are not aware of took place within *S. coelicolor* 802. As it was discussed in Sections 1.1.2 and 6.2.3 genome

duplications on *S. coelicolor* are not rare and working with an uncharacterised strain in this respect always poses a risk. Another explanation to this observation could be that the *put* functions are essential and therefore amplifications when attempting to isolate the mutants occurred. However, whether the *put* functions are essential a double crossover recombination event, resulting in kanamycin sensitivity, would not be expected to occur. Through my eyes it appears more likely that a double crossover would be selected against rather than selection for an amplification event during the crossover event, which seems mechanistically more complicate. However, the observation from previous workers that *S. coelicolor* Pum7r was a very sick strain that only formed non-sporulating tiny colonies may have a bearing on this result. It is interesting to note here that when similar experiments to those reported in Section 2.2.3 aimed at identifying *priA* mutants from a transposon mutagenesis library (see also Section 7.1.3.1) were done for *putB*, only one putative correct PCR product was obtained. Sequence analysis of this DNA fragment revealed that the transposon was inserted just a few base pairs downstream *putB*, implying that mutation of *putB* may be selected against.

As a final attempt to try to correlate the phenotype of *S. coelicolor* Pum7r with the *put* genes, a genomic region of 2.95 kb from this strain putatively including the *putA* and *putB* genes was amplified by PCR using the primers TAAGATCTGTGCCTG CGGAACCTCTGGAAT and CACAATTACTCGGACTTCCTAC. The PCR product obtained was identical in size to a region amplified with the same primers from M145, ruling out the possibility that major mutations occurred at this locus on Pum7r. Moreover, this PCR product was sequenced almost entirely (only closing of minor gaps

were missing) and analysis of the sequence confirmed that this region included the *putA* and *putB* genes, but also that no mutations were present. Although further experimentation is required to unambiguously prove that the *put* genes are not associated with the phenotype of Pum7r, this result raises some doubts on the link between proline catabolism and RED overproduction. It is interesting to note that very recent a phenotype similar to the Pum7r RED overproducing phenotype was found on a *S. coelicolor* mutant affected in a two-component system integrally involved in osmotic tolerance (A. Bishop, P. Herron and P. Dyson, personal communication). It would be interesting to characterise this mutant strain in terms of proline degradation and proline uptake from the medium.

6.3 Conclusions

In vivo characterisation of the function of the *proC* homologue SCO3337 of *S. coelicolor* was reported. The evidence obtained strongly suggests that this gene in fact encodes a P5C-R responsible for the last step of proline biosynthesis in *S. coelicolor*. Direct evidence includes the ability of SCO3337 to restore prototrophy of a *proC* mutant of *E. coli* by complementation. Furthermore, it was shown that overexpression of this gene, particularly as part of a *pro* artificial operon, had an effect on synthesis of the secondary metabolite RED, offering circumstantial evidence that SCO3337 encodes a P5C-R enzyme. However, this view could not be unambiguously confirmed by mutational analysis since disruption of SCO3337 did not result in a proline auxotroph.

Two putative *proC* paralogues (SCO5514 and SCO7154), which are homologues of *ilvC* from the biosynthesis of the branched-chain amino acids, were identified *in silico*. It was shown *in vivo* that the products of these three genes compensate each other, as judged from the auxotrophic phenotype of a triple mutant. Although confirmation of the link between this phenotype with the putative P5C-R activity of SCO3337 was not reported, this result putatively represents the first example of a common physiology for the branched-chain amino acids and proline. Moreover, and evolutionary relationship between the proteins P5C-R and AHIR is implicit, a link that has not been suggested before.

Chapter VII

Chapter VII

Materials and Methods

7.1 Molecular genetics methods

This thesis heavily rests in molecular genetics techniques such as molecular cloning and gene disruption. The standard methods for molecular cloning in *E. coli* were as in the laboratory manual of molecular cloning by Sambrook *et al.* (1989 and 2000; see below Section 7.1.2). The gene disruptions were done by PCR-targeting mutagenesis using the technology REDIRECT® developed by Dr Bertolt Gust and co-workers at the John Innes Centre (Gust *et al.*, 2003) and currently being commercialised by Plant Biosciences Ltd (PBL, Norwich, UK; see below Section 7.1.3.2 and Appendix 2).

7.1.1 Extraction and purification of DNA

The manipulation of DNA was following standard methods as described previously (Sambrook *et al.*, 1989 and 2000). In addition, the Qiagen plasmid mini prep kit was used for purification of small amounts of plasmid DNA from *E. coli*. For large amounts of DNA the caesium chloride method was used (Sambrook *et al.*, 1989 and 2000). The Qiagen DNA purification kit was used to purify the PCR products. The *S. coelicolor* DNA was manipulated as described by the protocols of the *Streptomyces* genetics laboratory manuals by Hopwood *et al.* (1985) and Kieser *et al.* (2000). Additionally, chromosomal DNA from *S. coelicolor* for screening purposes (e.g.

screening of transformants and exconjugants) was obtained using the Dynal Beads kit following the manufacturer's instructions (Dynal). The chromosomal DNA of *M. tuberculosis* was a kind gift from David Roper, University of Warwick.

7.1.2 Molecular cloning by PCR amplification in *E. coli*

All the DNA fragments were prepared by PCR amplification using primers with restriction sites engineered at their 5' ends. Table 7.1 shows the primers and other cloning details of all DNA fragments that were successfully cloned. Although the cloning of *trpE2* (SCO3214), *putA* (SCO5519) and *putB* (SCO5520) are not reported within Chapters V and VI, since further experimentation using these constructs was not done, different versions of these genes were cloned (see details in Table 7.1). All the PCR products intended for cloning were obtained using the 100% proofreading DNA polymerase *Pwo* (Roche). Nothing was modified from the conditions recommended by the manufacturer other than addition to the reaction mix of 5% – 10% DMSO final concentration, which was added to aid amplification of high (G + C) content DNA. As a general rule it was found that *Streptomyces* DNA was optimally amplified using as annealing temperature 55 – 60 °C. In addition to the restriction sites engineered, the PCR products obtained had one or more base pairs at their ends, which were inserted while designing the PCR primers to allow efficient restriction digestion by the endonucleases of choice (see New England Biolabs catalogue, 2000-01, pp. 210-211). Digestion was done overnight with endonuclease restriction enzymes purchased from GIBCO or MBI Fermentas other than *Nde*I, which was purchased from New England Biolabs (NEB), since there is some laboratory folklore that the NEB enzyme gives

better results. The buffers used for all digestions were from New England Biolabs since this buffer system is suitable for double digestions. The digested PCR products were analysed by agarose gel electrophoresis as described previously (Sambrook *et al.* 2000) and recovered by excision with a blade for gel extraction using the Qiagen DNA purification kit following the manufacturer's instructions.

The digested and clean PCR products were ligated to the vector of choice with T4 DNA ligase (Gibco) following the manufacturer's instructions. The vectors were digested and treated in a similar way as the PCR products that served as inserts (see above). Prior to the ligation reaction the blunt-ends of the digested and purified DNA vectors were dephosphorylated using alkaline phosphatase from shrimp (Roche). As recommended by the manufacturer the phosphatase was inactivated by heat treatment for 15 min at 65 °C. The pMTL plasmids (Chambers *et al.*, 1988) were used for general cloning purposes whereas the pET vectors (Novagen), pGEX-4T-1 (Amersham Pharmacia) and pProEX (Life Technologies) were used for protein overexpression (see Tables 7.1 and 7.2). The constructs obtained after the ligation reaction were used to transform calcium-competent *E. coli* MC1061 cells (Sambrook *et al.*, 1989), which were prepared fresh for every cloning experiment. Plasmids recovered from colonies obtained after selection with ampicillin (100 μg/ml) on LB plates were screened by restriction digest and PCR amplification.

The primers used for screening of clones targeted the backbone of the vectors just at the flanking regions of the multi-cloning site, such that positive clones were identified by an increase on size of the PCR products. The PCR primers used for this purpose were GGCAGTGAGCGCAACGCAAT and TCAGGCTGCGCAACTGTTGG

(pMTL); TAATACGACTCACTATAGGG and GCTAGTTATTGCTCAGCGG (pET); and GGGCTGGCAAGCCACGTTTGGTG and CCGGGAGCTGCATGTGTCAGAGG (pGEX). DNA sequencing using these vector-specific primers and others designed accordingly was done by the Department of Biological Sciences in-house sequencing service. All the constructs obtained were fully sequenced and only those that proved to be free of mutations were used for sub-cloning into *Streptomyces* protoplasts as described previously (Kieser *et al.*, 2000). The orientation of *proC* when inserted into pDHP14 was screened by PCR amplification using the *proC* primers from Table 7.1 and the pDHP14-specific primer GCACCCCGCCGCATTCTCG. The details of the constructs generated by PCR can be figured out from Table 7.1. The constructs obtained after sub-cloning are listed in Table 7.2.

Table 7.1 Molecular cloning of *S. coelicolor*, *M. tuberculosis* and *E. coli* genes by PCR.

Gene(s)	Region cloned	Size (kb)	PCR template	Primers (restriction sites)		Vector
SCO2050	*priA* plus putative promoter	1.1	Sc4G6	CCCAAGCTTGGACGCCCGCTTCTACTT	(*Hin*dIII)	pMTL22
				CGGGATCCGGGCGAAACCGAAGGACTC	(*Bam*HI)	
SCO2050	*priA* minus putative promoter	0.85	Sc4G6	GGAATTCCATATGAGCAAGCTCGAACTC	(*Nde*I)	pET-3a
				CGGGATCCGGGCGAAACCGAAGGACTC	(*Bam*HI)	
SCO2050	*priA* minus putative promoter	0.85	Sc4G6	CCGGAATTCATGAGCAAGCTCGAACTC	(*Eco*RI)	pGEX-4T-1 and pProEX
				CCGCTCGAGGGCGAAACCGAAGGACTC	(*Xho*I)	
Rv1603	*priA* minus putative promoter	0.8	*M. tuberculosis* chromosomal DNA	GGAATTCGTGATGCCGCTGATACTTTT	(*Eco*RI)	pGEX-4T-1
				CCCTCGAGTTCAACCAGCGGGGCAACCAAC	(*Xho*I)	
His4_Ecoli (AAC75085)	*hisA* minus putative promoter	0.8	*E. coli* chromosomal DNA	GGAATTCATGATTATTCCGGCATTAGA	(*Eco*RI)	pGEX-4T-1
				CCGCTCGAGTGGAAGCGGTGATATCGTAGAACA	(*Xho*I)	
His4_Ecoli (AAC75085)	*hisA* minus putative promoter	0.8	*E. coli* chromosomal DNA	GGAATTCCATATGATTATTCCGGCATTAGA	(*Nde*I)	pET-3a
				CGGGATCCTGGAAGCGGTGATATCGTAGAACA	(*Bam*HI)	

(Continues in following page)

Table 7.1 *(Continued from previous page)*

Gene(s)	Region cloned	Size (kb)	PCR template	Primers (restriction sites)	Vector
TrpC_Ecoli (AAC74344)	3' end of *trpC* (*trpF*)	0.6	*E. coli* chromosomal DNA	GGAATTCATGGCCCATGACGATTTG (*Eco*RI) CCGCTCGAGTGGAAGCGGTGATATCGTAGAACA (*Xho*I)	pGEX-4T-1
SCO2117	*trpE1* plus putative attenuator	2.3	Sc6E10	ACTAAGCTTCGAGCGCCGCACCACGAT (*Hind*III) GAAGATCTTCCCGAACGTCCTGGTCA (*Bgl*II)	pMTL21
SCO2117	*trpE1* minus putative attenuator	2.06	Sc6E10	GTCAAGCTTCGGTGTTTTCGAGCCCG (*Hind*III) GAAGATCTTCCCGAACGTCCTGGTCA (*Bgl*II)	pMTL21
SCO3214	5' end of *trpE2* plus putative promoter	0.37	ScE8	TAGAAGCTTCACCATTCCGCGGCGTGCC (*Hind*III) GAGGATCCGGTCGGTGACCGACGTAC (*Bam*HI)	pMTL21
SCO3214	*trpE2* plus putative promoter region	1.7	ScE8	TAGAAGCTTCACCATTCCGCGGCGTGCC (*Hind*III) GAGGATCCGGCGACGTAGTGGACGAGG (*Bam*HI)	pMTL21
SCO3213 SCO3214	*trpE2* and *trpG* plus its putative promoter	2.3	ScE8	TAGAAGCTTCACCATTCCGCGGCGTGCC (*Hind*III) GAGGATCCGCGCATGATCACGCGCAT (*Bam*HI)	pMTL21
SCO3214 SCO3215	5' end of *trpE2* plus the gene upstream with its putative promoter	1.6	ScE8	TAGAAGCTTCCGGTCTCGGCTGCGTCC (*Hind*III) GAGGATCCGGTCGGTGACCGACGTAC (*Bam*HI)	pMTL21
SCO3337	*proC* plus putative promoter	1.2	ScE7	CGGGGTACCGGCGCGGCAACCCTGAAAC (*Kpn*I) AAAACTGCAGGGTTCTGGCCGGCTCCTGAC (*Pst*I)	pMTL21
SCO3337	*proC* minus putative promoter	0.96	ScE7	GGGAATTCCATATGACCCCGGACCACGCGCACA (*Nde*I) CGCGGATCCCGCCCGGGTGGTGACGA (*Bam*HI)	pET-3a
SCO3337	*proC* minus putative promoter	0.96	ScE7	GGAATTCATGACCCCGGACCACGCGCACA (*Eco*RI) CCGCTCGAGCGCCCGGGTGGTGACGA (*Xho*I)	pGEX-4T-1
SCO5520	*putB* plus putative promoter	2.3	Sc8D9	ATGGATCCGCTGATGGAGGGCGAGGGCTACC (*Bam*HI) TGTAAGCTTGACGTTCGGCTCCTACAAAGTTA (*Hind*III)	pMTL21
SCO5520	5' end of *putB* plus putative promoter	1.1	Sc8D9	ATGGATCCGCTGATGGAGGGCGAGGGCTACC (*Bam*HI) TGAAAGCTTCGATCGCGCTGAAGTTGA (*Hind*III)	pMTL21
SCO5519 SCO5520 SCO5521	*putA* with its putative promoter plus *putB* and the gene downstream	3.4	Sc8D9	TAAGATCTGTGCCTGCGGAACCTCTGGAATC (*Bgl*II) TGTAAGCTTGCGAGATTGAGGCTGCGAGACAT (*Hind*III)	pMTL21
SCO5519 SCO5520	*putA* with its putative promoter and the 5' end of *putB*	1.9	Sc8D9	TAAGATCTGTGCCTGCGGAACCTCTGGAATC (*Bgl*II) TGAAAGCTTCGATCGCGCTGAAGTTGA (*Hind*III)	pMTL21

Table 7.2 Plasmids generated in this study by sub-cloning

Insert (sites)	Cloning vector (sites)	New plasmid
priA (*Bam*HI/*Bam*HI) from pMTL22priASc	pIJ702 (*Bgl*II)	pIJ702priASc
priA (*Nde*I/*Bam*HI) from pET3a-priASc	pET-15b (*Nde*I/*Bam*HI)	pETpriASc
trpF (*Nco*I/*Bam*HI) from pMS401[†]	pET-22a (*Nco*I/*Bam*HI)	pET22a-trpFEc
trpF (*Nde*I/*Bam*HI) from pET22a-trpFEc	pIJ4123 (*Nde*I/*Bam*HI)	pIJ4123trpFEc
hisA (*Nde*I/*Bam*HI) from pET3a-hisAEc	pIJ4123 (*Nde*I/*Bam*HI)	pIJ4123hisAEc
proC (*Nde*I/*Bam*HI) from pET3a-proCSc	pIJ4123 (*Nde*I/*Bam*HI)	pFBP-1c
proC (*Pst*I/*Sph*I) from pMTL21proCSc	pMTL23 (*Pst*I/*Sph*I)	pMTL23proCSc
proC (*Sph*I/*Pst*I) from pMTL23proCSc	pDHP14 (*Sph*I/*Pst*I)	pFBP140
proC (*Xho*I/HindIII) from pMTL21proCSc	pMTL24 (*Xho*I/HindIII)	pMTL24proCSc
proC (*Bam*HI/*Bam*HI) from pMTL24proCSc	pIJ702 (*Bam*HI/*Bam*HI)	pFBP-2c

[†] pMS401 was a gift from Dr Jonathan Blackburn, University of Cambridge, and it is a pJB122 derivative (Altamirano *et al.*, 2000), which comes from pTrcHis (Invitrogen).

7.1.3 Disruption of *S. coelicolor* genes

7.1.3.1 Transposon mutagenesis

The transposon Tn*4560* mutagenesis system for isolation of *S. coelicolor* mutants has been previously described (Fowler, 2002; Gust *et al.*, 2003). The primers used for screening of putative *priA* (SCO2050) mutants were GGGCGAAACCGAAGG ACTC and TCGTGGCCGCCGTGGAGAACG for the first round of PCR amplification and CGCACGGCTTCGGATGTCATG and AGATGAGCCCGGTAGAGATGAGC for the second round of nested-PCR amplification. The primers used for screening of putative *putB* (SCO5520) mutants were CGCCCGCTCGATGGTCAG and GACGTTC GGCTCCTACAAAG for the first round of PCR amplification and CACCCGCTCGCA

AGACCGCTCAG and CACAATTACTCGGACTTCCTAC for the second round of nested-PCR amplification.

7.1.3.2 PCR-targeting mutagenesis: REDIRECT® technology

Details on PCR-mutagenesis in prokaryotes can be found in Datsenko and Wanner (2000) and Gust *et al.* (2003). Due to the relevance of this technology to this thesis the manual provided by PBL is included as Appendix 2 (with permission from Dr Bertolt Gust and PBL; see also Figure 2.5). The mutagenesis oligonucleotides used and the mutants obtained are listed in Table 7.3. The disruptions were first done within *E. coli* in cosmids kindly provided by Dr Helen Kieser from the John Innes Centre (Redenbach *et al.*, 1996), therefore a mutated cosmid for each disruption is available. For the case of the unmarked scar mutants, where reengineering of the SuperCos1 backbone was required to transfer the mutant cosmid by conjugation, the *neo* gene was replaced with the *aac(3)IV* plus *oriT* cassette of pIJ773 using primers kindly provided by Dr Bertolt Gust of the John Innes Centre (Table 7.4). The cosmids generated for disrupting *putA* and *putB* are also included in Table 7.4. When possible the PCR primers reported in Table 7.1 were also used for screening of the putative mutants. In some cases new primers were designed and they are as follows: GGGCGAAACCGAAGGACTC and TCGTGGCCGCCGTGGAGAACG (*priA*); CCGCCCACGTCTTCGTCCTCTTCT and CCGGCCTCGTTCCGCTGGTCAAGT (*ilvC2*); TGAAGGTGCGCTCCGACAAC GAGA and GTCCTGCGCCTCTGCGGAAGCATC (*ilvC1*).

Table 7.3 *Streptomyces coelicolor* mutants generated in this study.

Strain	Gene(s) (cosmid)	Comments	Genotype	Disruption primers (sequence homology)
WH101	SCO2050 (Sc4G6)	Internal region of *priA*	M145 *ΔpriA::scar*	TGGGTCGCCAAGGTCATCGCCGAGCACGGCGACAAGA TCATTCCGGGGATCCGTCGACC CTTCCCGACGATGGCCCCCTCGACACCGGCCGGGACG AGTGTAGGCTGGAGCTGCTTC
WH102	SCO2050 SCO2049 (Sc4G6)	*priA* plus the 5' end of the *yabJ* homolog	M145 *ΔpriA:: aac(3)IV*	CAACTGGATCGAAACCCTCTAGAGATGAGCCCGGTAGA GATTCCGGGGATCCGTCGACC GCGCGGGCGAAACCGAAGGACTCTTCCCAGGGACTCC CGTGTAGGCTGGAGCTGCTC
WPZ101	SCO2117 (Sc6E10)	*trpE1*	M145 *ΔtrpE1:: aac(3)IV*	ATGACACCGCTCGACAGCCTGCTGACCGGCACCCGCC CCATTCCGGGGATCCGTCGACC CTACCGGGAGGGCCGCCGCTCGGAGAACGTGCTCGTA CCTGTAGGCTGGAGCTGCTTC
WH103	SCO2043 (Sc4G6)	*trpE3*	M145 *ΔtrpE3:: aac(3)IV*	ATGGACGTCACGCATGACATGGACCTCGACACCTTCCG CATTCCGGGGATCCGTCGACC CTATTGCGCCAGCCGGTTGGCCGTGTGTACGGCCCTC AGTGTAGGCTGGAGCTGCTTC
WH104	SCO2117 (Sc6E10)	*trpE1* on a *trpE3* minus background	WH103 *ΔtrpE1:: vio*	As in WPZ101
WP101 WP102	SCO3337 (ScE7)	*proC*	M145 *ΔproC::aac(3)IV* M145 *ΔproC::scar*	ATGACCCCGGACCACGCGCACACCCCGGAGGGCACC GAAATTCCGGGGATCCGTCGACC TCAGTCCTTGGTGCCGGTGGCCAGTTCGCGGCTGCGG TCTGTAGGCTGGAGCTGCTTC
WILV101 WILV102	SCO7154 (Sc9A4)	*ilvC2*	M145 *ΔilvC2::aac(3)IV* M145 *ΔilvC2::scar*	ATCCCCGCTCGGGGAGTTCTCCGAAAGGTGTGTCTCAT GATTCCGGGGATCCGTCGACC GCGCCGCCGTGGCGCGCCCACCGGCCTGCCCGGATG TCATGTAGGCTGGAGCTGCTTC
WILV103	SCO5514 (Sc8D9)	*ilvC1*	M145 *ΔilvC1::aac(3)IV*	GTGGGACGCAACACCTGCACACAAGGAGAGAACCCAG TGATTCCGGGGATCCGTCGACC CGCCCGGCGCGCCGGGCGCCCCCGCGGGGCAAAGAC CTATGTAGGCTGGAGCTGCTTC
WILV104	SCO5514 (Sc8D9)	*ilvC1* on a *ilvC2* minus background	WILV102 *ΔilvC1::aac(3)IV*	As in WILV103
WILV103	SCO5514 (Sc8D9)	*ilvC1* on a *proC* minus background	WP102 *ΔilvC1::aac(3)IV*	As in WILV103
WILVP101	SCO3337 (ScE7)	*proC* on a double *ilvC* minus background	WILV104 *ΔproC::vio*	As in WP101 and WP102

Table 7.4 Mutant cosmids not included in Table 7.3.

Gene(s) (cosmid)	Comments	Genotype	Disruption primers (sequence homology)
neo (Any Sc cosmids bearing scar replacements)	*neo* marker on a scar minus background	Any Sc cosmids bearing scar replacements *Δneo::aac(3)IV*	ATGATTGAACAAGATGGATTGCACGCAGGTTCTCCGGC CATTCCGGGGATCCGTCGACC TCAGAAGAACTCGTCAAGAAGGCGATAGAAGGCGATG CGTGTAGGCTGGAGCTGCTTC
SCO5519 (Sc8D9) ‡	*putA*	Sc8D9 *ΔputA:: aac(3)IV* Sc8D9 *ΔputB::scar*	GCGGCCCCTGTGTTGGTTGGTAAGTGGAGGTGCCCCG TGATTCCGGGGATCCGTCGACC ACTCCTTTAACTGAGCGGTCTTGCGAGCGGGTGGGCT CATGTAGGCTGGAGCTGCTTC
SCO5520 (Sc8D9) ‡	*putB*	Sc8D9 *ΔputB:: aac(3)IV* Sc8D9 *ΔputB::scar*	ACCCGCTCGCAAGACCGCTCAGTTAAAGGAGTTACGG AAATTCCGGGGATCCGTCGACC GGAGCGTCTGTCTCCACAATTACTCGGACTTCCTACTA TTGTAGGCTGGAGCTGCTTC

‡ Sc8D9 cosmids with double *putA* and *putB* replacements were also constructed.

7.1.4 Southern hybridisations

The general Southern hybridisation protocols have been previously described (Sambrook *et al.*, 1989 and 2000). The method for purification of the chromosomal DNA from *S. coelicolor* was as in Kieser *et al.* (2000). Chromosomal DNA was digested with *Sau*3A1 (Gibco). The probes used were generated by PCR amplification. The DIG High Prime DNA Labelling and Detection Starter kit from Roche was used. The manufacturer's instructions were followed other than the modifications described. The amounts of hexanucleotides, dNTPs and Klenow on the labelling reaction were doubled. The labelled DNA was blotted into a nylon membrane Hybond NX from Amersham Pharmacia. Prehybridisation was achieved in 60 ml of standard pre-hybridisation solution for at least two hours at 68 °C in a roller bottle. The pre-hybridisation solution was poured off and 20 ml of hybridisation solution containing the heat-denaturated *S. coelicolor* probes at concentrations of 25 ng/ml approximately were added and allowed to hybridise overnight at 68 °C. Post hybridisation washes with pre-heated solutions (68 °C) were carried out as follows: two 15 minutes washes with 0.5X

SSC, 0.1% SDS; and three 20 minutes washes with 0.1X SSC, 0.1% SDS under constant agitation. The chemiluminescent detection procedure was done as described in the user's guide (Roche) except that no Tween was added to the washing buffer and the membrane was blocked with blocking solution for 120 – 180 minutes rather that 30 – 60 minutes.

7.2 Strains and growth requirements

Strains Hfr G6, W3110 *trpC*(*F*del), and χ342 were provided by the *E. coli* Genetic Stock Center, New Haven. Minimal A medium (Miller, 1972) supplemented with the appropriate growth factors was used to test strains Hfr G6 and χ342 and derivatives while modified Vogel-Bonner medium (Darimont *et al.*, 1998) was used to test W3110 *trpC*(*F*del) and derivatives. When testing the amino acid requirements of *E. coli* strains the compounds were added at a final concentration of 100 µg/ml. Plasmids were selected with 100 µg/ml of ampicillin or 25 µg/ml of chloramphenicol (Sigma) as needed. Induction of *lacI*-dependent promoters was achieved with IPTG. The *Streptomyces* minimal medium (Kieser *et al.*, 2000) was supplemented with amino acids at a concentration of 37.5 µg/ml other than histidine, which was at 50 µg/ml. Thiostrepton (Sigma) was used to select for plasmids and induction of the thiostrepton-inducible promoter in *Streptomyces* at 50 µg/ml. Apramycin (Sigma) selection was at 50 µg/ml and viomycin (a kind gift from T. Kieser) at 30 µg/ml. See Appendix 2 for details on strains used for PCR-targeting mutagenesis.

7.3 Biochemical methods

The methods for the purification and characterisation of PriA are described below (Sections 7.3.1 to 7.3.3). See also Barona-Gómez and Hodgson (2003) and Wright *et al.* (2003). The HPLC conditions used for the isolation of RED produced by *S. coelicolor* are also reported within this Section (see Section 7.3.4).

7.3.1 Assay for PRAI activity

N'-(5'-phosphoribosyl)anthranilate (PRA) was synthesised *in situ* (Kirschner *et al.*, 1987) and PRA isomerase activity was measured by detecting production of 1-(o-carboxyphenylamino)-1-deoxy-ribulose-5-phosphate (CDRP) at 350 nm immediately after the enzyme preparation was added to the reaction mix. The reaction mix contained 80 m*M* Tris-HCl, pH 8.6, 0.4 mM EDTA, 0.1 mM anthranilic acid and 0.1 mM ribose-5-phosphate (Sigma). The readings were recorded after allowing the reaction to undergo for two minutes.

7.3.2 Overexpression and purification of PriA, eHisA and eTrpF

Overexpressed PriA was purified as a 6x his-tagged fusion from the expression plasmid pETpriASc in *E. coli* strain C41(DE3) (Miroux & Walker, 1996) containing the auxiliary plasmid pRIL (Stratagene) grown on LB broth. Soluble protein was obtained after induction with 1 mM IPTG at an OD_{600} of between 0.6-0.8 and overnight expression at 20 °C. The cells from a 0.5 litre culture were harvested and resuspended in 25 ml of ice-cold 50 mM Tris-HCl (pH 8.0), 0.5 mM EDTA, 50 mM NaCl, 5%

glycerol, 0.1 mM DTT and 100 μg/ml lysozyme. The cells were sonicated at 70% power in 20 seconds intervals for 4 minutes. The sonicate was centrifuged at 16,000 rpm for 20 minutes at 4 °C and the soluble extract was bound to three millilitres of chelating sepharose fast flow (Amersham Pharmacia) activated with $NiCl_2$. The sepharose was washed twice with 50mM HEPES (pH 8.0), 1M NaCl, 10mM imidazole, then washed once with 50 mM HEPES (pH 8.0), 1M NaCl and 20 mM imidazole. PriA was eluted with 50 mM HEPES (pH 8.0), 100 mM NaCl and 100 mM imidazole. The elute was concentrated by ultrafiltration and loaded onto a Superdex 75 HR26/60 column (Amersham Pharmacia) at 4 °C pre-equilibrated with 50 mM Tris-HCl pH 8, 200 mM NaCl, 1 mM EDTA and 0.5 mM DTT. The column flow rate was 2 ml/min and 5 ml fractions were collected. Fractions containing PriA were identified using the enzyme assay described in Section 7.3.1. Fractions were pooled and concentrated by centrifugation (Centriprep, Millipore) to approximately 10 mg/ml for crystallisation trials. The degree of purity of PriA as determined by SDS-PAGE was greater than 95%. The eHisA and eTrpF GST fusions were produced in soluble form from pGEXhisAEc and pGEXtrpFEc, respectively. The hosts used were Hfr G6 and W3110 *trpC*(*F*del) (see Section 7.2) and the GST fusions were purified using the MicroSpin GST Purification Module (Amersham Pharmacia) following the manufacturer's instructions. The buffers sparse matrix used to screen for the solubility of the GST-PriA fusion during sonication was as in Lindwall *et al.* (2000)

7.3.3 Crystallisation of PriA

Initial crystallisation trials were performed with screens from Molecular Dimensions Limited (MDL) and Emerald Biostructures Incorporated using the hanging-drop vapour-diffusion technique. Needles were obtained with condition '30' from the MDL Structure Screen No. 1 using 1µl of protein at ~10 mg/ml mixed with and equal volume of mother liquor. After optimisation, crystals grew after 4 to 5 days at 25 °C in mother liquor consisting of 1.5 M ammonium sulphate and 100 mM sodium citrate (pH 4.8). The largest crystals obtained were 0.3 x 0.05 x 0.05 mm in size.

7.3.4 HPLC analysis of the RED complex

Cell pellets from duplicate experiments were harvested by centrifugation at 3000 *g* for 10 minutes. The pellets were resuspended in four volumes of a 1:1 mixture of CH_3OH : CH_3CN and acidify to pH 1-2 with 6N HCl. The biomass was span down by centrifugation as above and saved for determination of dried biomass. The supernatants were collected and prior to injection the samples were filtered (Millipore) to remove any impurities that may block the column. The chromatographic separation was done on a Phenomenex Kingsorb C-18 column, particle size 150 x 4.6 mm, at 40 °C. The metabolites were monitored at a wavelength of 364 and 533 nm, but only the readings at the latter wavelength were used for obtaining the integration values. The addition of these values was used as the total RED production. A gradient profile with solvent A CF_3COOH : CH_3OH : CH_3CN : H_2O (2.5 : 400 : 100 : 497.5) and solvent B CH_3OH from 20% to 100% of solvent B in 15 minutes was used (G. Challis, personal communication).

7.4 Biocomputational tools

All the biocomputational programs and databases used in this thesis were accessed from the Internet. The *S. coelicolor* databases were accessed from the Wellcome Trust Sanger Institute homepage (http://www.sanger.ac.uk/Projects/S_coelicolor/). The BLAST and PSI-BLAST (Altschul *et al.*, 1997) searches were done from the homepage of the National Centre for Biotechnology Information (http://www.ncbi.nlm.nih.gov/BLAST/). The hidden Markov models (Karplus *et al.*, 1998) were done using the University of California at Santa Cruz SAM-T99 HMM web servers (http:// www.cse.ucsc.edu/research/compbio/sam.html). The sequence alignments were done using the program ClustalX 1.81 or ClustalW, the latter being loaded from within the program BioEdit developed by the North Carolina State University (http://www.mbio.ncsu.edu/BioEdit/ bioedit.html). The phylogenetic tree of PriA and homologues was done using protdist and the neighbour-joining method with 100 bootstrap replications. The phylogenetic co-occurrence of *trpF* was done using the web server STRING (http://dag.embl-heidelberg.de/newstring_cgi/show_input_page.pl).

Bibliography

Bibliography

Adams, E. and Frank, L. (1980). "Metabolism of proline and the hydroxyprolines." *Annu Rev Biochem* **49**: 1005-61.

Ahn, H. J., Eom, S. J., Yoon, H. J., Lee, B. I., Cho, H. and Suh, S. W. (2003). "Crystal structure of class I acetohydroxy acid isomeroreductase from *Pseudomonas aeruginosa*." *J Mol Biol* **328**(2): 505-15.

Alifano, P., Fani, R., Lio, P., Lazcano, A., Bazzicalupo, M., Carlomagno, M. S. and Bruni, C. B. (1996). "Histidine biosynthetic pathway and genes: Structure, regulation, and evolution." *Microbiol Rev* **60**(1): 44-69.

Altamirano, M. M., Blackburn, J. M., Aguayo, C. and Fersht, A. R. (2000). "Directed evolution of new catalytic activity using the alpha/beta-barrel scaffold." *Nature* **403**(6770): 617-22.

Altamirano, M. M., Blackburn, J. M., Aguayo, C. and Fersht, A. R. (2002). "Retraction. Directed evolution of new catalytic activity using the alpha/beta-barrel scaffold." *Nature* **417**(6887): 468.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997). "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs." *Nucleic Acids Res* **25**(17): 3389-402.

Alves, R., Chaleil, R. A. and Sternberg, M. J. (2002). "Evolution of enzymes in metabolism: A network perspective." *J Mol Biol* **320**(4): 751-70.

Anantharaman, V., Aravind, L. and Koonin, E. V. (2003). "Emergence of diverse biochemical activities in evolutionarily conserved structural scaffolds of proteins." *Curr Opin Chem Biol* **7**(1): 12-20.

Anderson, A. S. and Wellington, E. M. (2001). "The taxonomy of *Streptomyces* and related genera." *Int J Syst Evol Microbiol* **51**(Pt 3): 797-814.

Anderson, R. P. and Roth, J. R. (1977). "Tandem genetic duplications in phage and bacteria." *Annu Rev Microbiol* **31**: 473-505.

Ankri, S., Serebrijski, I., Reyes, O. and Leblon, G. (1996). "Mutations in the *Corynebacterium glutamicum* proline biosynthetic pathway: A natural bypass of th *proA* step." *J Bacteriol* **178**(15): 4412-9.

Aravind, L., Mazumder, R., Vasudevan, S. and Koonin, E. V. (2002). "Trends in protein evolution inferred from sequence and structure analysis." *Curr Opin Struct Biol* **12**(3): 392-9.

Arhin, F. F. and Vining, L. C. (1993). "Organization of the genes encoding p-aminobenzoic acid synthetase from *Streptomyces lividans* 1326." *Gene* **126**(1): 129-33.

Arst, H. N., Jr., Jones, S. A. and Bailey, C. R. (1981). "A method for the selection of deletion mutations in the l-proline catabolism gene cluster of *Aspergillus nidulans*." *Genet Res* **38**(2): 171-95.

Banner, D. W., Bloomer, A. C., Petsko, G. A., Phillips, D. C., Pogson, C. I., Wilson, I. A., Corran, P. H., Furth, A. J., Milman, J. D., Offord, R. E., Priddle, J. D. and Waley, S. G. (1975). "Structure of chicken muscle triose phosphate isomerase determined crystallographically at 2.5 Angstrom resolution using amino acid sequence data." *Nature* **255**(5510): 609-14.

Bao, K. and Cohen, S. N. (2001). "Terminal proteins essential for the replication of linear plasmids and chromosomes in *Streptomyces*." *Genes Dev* **15**(12): 1518-27.

Bao, K. and Cohen, S. N. (2003). "Recruitment of terminal protein to the ends of *Streptomyces* linear plasmids and chromosomes by a novel telomere-binding protein essential for linear DNA replication." *Genes Dev* **17**(6): 774-85.

Barona-Gomez, F. and Hodgson, D. A. (2003). "Occurrence of a putative ancient-like isomerase involved in histidine and tryptophan biosynthesis." *EMBO Rep* **4**(3): 296-300.

Bartlett, G. J., Porter, C. T., Borkakoti, N. and Thornton, J. M. (2002). "Analysis of catalytic residues in enzyme active sites." *J Mol Biol* **324**(1): 105-21.

Bathe, B., Kalinowski, J. and Puhler, A. (1996). "A physical and genetic map of the *Corynebacterium glutamicum* ATCC 13032 chromosome." *Mol Gen Genet* **252**(3): 255-65.

Beadle, G. and Tatum, E. L. (1941). "Genetic control of biochemical reactions in *Neurospora*." *Proc Natl Acad Sci U S A* **27**: 499-506.

Behal, V. (2000). "Bioactive products from *Streptomyces*." *Adv Appl Microbiol* **47**: 113-56.

Belitsky, B. R., Brill, J., Bremer, E. and Sonenshein, A. L. (2001). "Multiple genes for the last step of proline biosynthesis in *Bacillus subtilis*." *J Bacteriol* **183**(14): 4389-92.

Bentley, S. D., Chater, K. F., Cerdeno-Tarraga, A. M., Challis, G. L., Thomson, N. R., James, K. D., Harris, D. E., Quail, M. A., Kieser, H., Harper, D., Bateman, A., Brown, S., Chandra, G., Chen, C. W., Collins, M., Cronin, A., Fraser, A., Goble, A., Hidalgo, J., Hornsby, T., Howarth, S., Huang, C. H., Kieser, T., Larke, L., Murphy, L., Oliver, K., O'Neil, S., Rabbinowitsch, E., Rajandream, M. A., Rutherford, K., Rutter, S., Seeger, K., Saunders, D., Sharp, S., Squares, R., Squares, S., Taylor, K., Warren, T., Wietzorrek, A., Woodward, J., Barrell, B. G., Parkhill, J. and Hopwood, D. A. (2002). "Complete genome sequence of the model actinomycete *Streptomyces coelicolor* a3(2)." *Nature* **417**(6885): 141-7.

Bentley, S. D., Maiwald, M., Murphy, L. D., Pallen, M. J., Yeats, C. A., Dover, L. G., Norbertczak, H. T., Besra, G. S., Quail, M. A., Harris, D. E., von Herbay, A., Goble, A., Rutter, S., Squares, R., Squares, S., Barrell, B. G., Parkhill, J. and Relman, D. A. (2003). "Sequencing and analysis of the genome of the whipple's disease bacterium *Tropheryma whipplei*." *Lancet* **361**(9358): 637-44.

Berka, R. M., Cui, X. and Yanofsky, C. (2003). "Genomewide transcriptional changes associated with genetic alterations and nutritional supplementation affecting tryptophan metabolism in *Bacillus subtilis*." *Proc Natl Acad Sci U S A* **100**(10): 5682-7.

Bertolaet, B. L. and Knowles, J. R. (1995). "Complementation of fragments of triosephosphate isomerase defined by exon boundaries." *Biochemistry* **34**(17): 5736-43.

Bessette, P. H., Cotto, J. J., Gilbert, H. F. and Georgiou, G. (1999). "In vivo and in vitro function of the *Escherichia coli* periplasmic cysteine oxidoreductase DsbG." *J Biol Chem* **274**(12): 7784-92.

Betran, E. and Long, M. (2002). "Expansion of genome coding regions by acquisition of new genes." *Genetica* **115**(1): 65-80.

Biou, V., Dumas, R., Cohen-Addad, C., Douce, R., Job, D. and Pebay-Peyroula, E. (1997). "The crystal structure of plant acetohydroxy acid isomeroreductase complexed with NADPH, two magnesium ions and a herbicidal transition state analog determined at 1.65 A resolution." *Embo J* **16**(12): 3405-15.

Birch, A., Hausler, A. and Hutter, R. (1990). "Genome rearrangement and genetic instability in *Streptomyces* spp." *J Bacteriol* **172**(8): 4138-42.

Birch, A., Hausler, A., Ruttener, C. and Hutter, R. (1991). "Chromosomal deletion and rearrangement in *Streptomyces glaucescens*." *J Bacteriol* **173**(11): 3531-8.

Blanc, V., Gil, P., Bamas-Jacques, N., Lorenzon, S., Zagorec, M., Schleuniger, J., Bisch, D., Blanche, F., Debussche, L., Crouzet, J. and Thibaut, D. (1997). "Identification and analysis of genes from *Streptomyces pristinaespiralis* encoding enzymes involved in the biosynthesis of the 4-dimethylamino-l-phenylalanine precursor of pristinamycin I." *Mol Microbiol* **23**(2): 191-202.

Bork, P., Gellerich, J., Groth, H., Hooft, R. and Martin, F. (1995). "Divergent evolution of a beta/alpha-barrel subclass: Detection of numerous phosphate-binding sites by motif search." *Protein Sci* **4**(2): 268-74.

Branden, C. I. (1990). "Protein chemistry - founding fathers and families." *Nature* **346**(6285): 607-608.

Brickman, T. J., Ozenberger, B. A. and McIntosh, M. A. (1990). "Regulation of divergent transcription from the iron-responsive *fepB-entC* promoter-operator regions in *Escherichia coli*." *J Mol Biol* **212**(4): 669-82.

Brown, M. P., Aidoo, K. A. and Vining, L. C. (1996). "A role for *pabAB*, a p-aminobenzoate synthase gene of *Streptomyces venezuelae* ISP5230, in chloramphenicol biosynthesis." *Microbiology* **142**: 1345-55.

Bryson, V. and Vogel, H.J. (1965). *Evolving Genes and Proteins*. New York, Academic Press.

Buss, K., Muller, R., Dahm, C., Gaitatzis, N., Skrzypczak-Pietraszek, E., Lohmann, S., Gassen, M. and Leistner, E. (2001). "Clustering of isochorismate synthase genes *menF* and *entC* and channeling of isochorismate in *Escherichia coli*." *Biochim Biophys Acta* **1522**(3): 151-7.

Cerdeno, A. M., Bibb, M. J. and Challis, G. L. (2001). "Analysis of the prodiginine biosynthesis gene cluster of *Streptomyces coelicolor* A3(2): New mechanisms for chain initiation and termination in modular multienzymes." *Chem Biol* **8**(8): 817-29.

Chambers, S. P., Prior, S. E., Barstow, D. A. and Minton, N. P. (1988). "The pMTL *nic*- cloning vectors. I. Improved puc polylinker regions to facilitate the use of sonicated DNA for nucleotide sequencing." *Gene* **68**(1): 139-49.

Chang, P. C. and Cohen, S. N. (1994). "Bidirectional replication from an internal origin in a linear *Streptomyces* plasmid." *Science* **265**(5174): 952-4.

Chang, Z., Sun, Y., He, J. and Vining, L. C. (2001). "P-aminobenzoic acid and chloramphenicol biosynthesis in *Streptomyces venezuelae*: Gene sets for a key enzyme, 4-amino-4-deoxychorismate synthase." *Microbiology* **147**(Pt 8): 2113-26.

Chater, K. F. (1993). "Genetics of differentiation in *Streptomyces*." *Annu Rev Microbiol* **47**: 685-713.

Chater, K. F. (2001). "Regulation of sporulation in *Streptomyces coelicolor* A3(2): A checkpoint multiplex?" *Curr Opin Microbiol* **4**(6): 667-73.

Chater, K. F., Bucca, G., Dyson, P., Fowler, K., Gust, B., Herron, P., Hesketh, A., Hotchkiss, G., Kieser, T., Mersinias, V. and Smith, C. P. (2002). *Streptomyces coelicolor* A3(2): From genome sequence to function. In *Functional Microbial Genomics*. B. Wren and N. Dorrell. Oxford, Academic Press. Methods in Microbiology sereies: **33**.

Chater, K. F. and Horinouchi, S. (2003). "Signalling early developmental events in two highly diverged *Streptomyces* species." *Mol Microbiol* **48**(1): 9-15.

Chaudhuri, B. N., Lange, S. C., Myers, R. S., Chittur, S. V., Davisson, V. J. and Smith, J. L. (2001). "Crystal structure of imidazole glycerol phosphate synthase: A tunnel through a $(beta/alpha)_8$ barrel joins two active sites." *Structure (Camb)* **9**(10): 987-97.

Chen, C. W., Huang, C. H., Lee, H. H., Tsai, H. H. and Kirby, R. (2002). "Once the circle has been broken: Dynamics and evolution of *Streptomyces* chromosomes." *Trends Genet* **18**(10): 522-9.

Chong, P. P., Podmore, S. M., Kieser, H. M., Redenbach, M., Turgay, K., Marahiel, M., Hopwood, D. A. and Smith, C. P. (1998). "Physical identification of a chromosomal locus encoding biosynthetic genes for the lipopeptide calcium-dependent antibiotic (CDA) of *Streptomyces coelicolor* A3(2)." *Microbiology* **144 (Pt 1)**: 193-9.

Chothia, C. (1988). "Protein structure: The 14th barrel rolls out." *Nature* **333**(6174): 598-9.

Clayton, T. M. and Bibb, M. J. (1990). "*Streptomyces* promoter-probe plasmids that utilise the *xylE* gene of *Pseudomonas putida*." *Nucleic Acids Res* **18**(4): 1077.

Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry, C. E., 3rd, Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Barrell, B. G. and et al. (1998). "Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence." *Nature* **393**(6685): 537-44.

Cole, S. T., Eiglmeier, K., Parkhill, J., James, K. D., Thomson, N. R., Wheeler, P. R., Honore, N., Garnier, T., Churcher, C., Harris, D., Mungall, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R. M., Devlin, K., Duthoy, S., Feltwell, T., Fraser, A., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Lacroix, C., Maclean, J., Moule, S., Murphy, L., Oliver, K., Quail, M. A., Rajandream, M. A., Rutherford, K. M., Rutter, S., Seeger, K., Simon, S., Simmonds, M., Skelton, J., Squares, R., Squares, S., Stevens, K., Taylor, K., Whitehead, S., Woodward, J. R. and Barrell, B. G. (2001). "Massive gene decay in the leprosy bacillus." *Nature* **409**(6823): 1007-11.

Copley, R. R. and Bork, P. (2000). "Homology among (beta/alpha)$_8$ barrels: Implications for the evolution of metabolic pathways." *J Mol Biol* **303**(4): 627-41.

Copley, S. D. (2003). "Enzymes with extra talents: Moonlighting functions and catalytic promiscuity." *Curr Opin Chem Biol* **7**(2): 265-72.

Craster, H. L., Potter, C. A. and Baumberg, S. (1999). "End-product control of expression of branched-chain amino acid biosynthesis genes in *Streptomyces coelicolor* A3(2): Paradoxical relationships between DNA sequence and regulatory phenotype." *Microbiology* **145 (Pt 9)**: 2375-84.

Crawford, I. P. (1989). "Evolution of a biosynthetic pathway: The tryptophan paradigm." *Annu Rev Microbiol* **43**: 567-600.

Criado, L. M., Martin, J. F. and Gil, J. A. (1993). "The *pab* gene of *Streptomyces griseus*, encoding p-aminobenzoic acid synthase, is located between genes possibly involved in candicidin biosynthesis." *Gene* **126**(1): 135-9.

Crosa, J. H. and Walsh, C. T. (2002). "Genetics and assembly line enzymology of siderophore biosynthesis in bacteria." *Microbiol Mol Biol Rev* **66**(2): 223-49.

Cunin, R., Glansdorff, N., Pierard, A. and Stalon, V. (1986). "Biosynthesis and metabolism of arginine in bacteria." *Microbiol Rev* **50**(3): 314-52.

da Silva, A. C., Ferro, J. A., Reinach, F. C., Farah, C. S., Furlan, L. R., Quaggio, R. B., Monteiro-Vitorello, C. B., Van Sluys, M. A., Almeida, N. F., Alves, L. M., do Amaral, A. M., Bertolini, M. C., Camargo, L. E., Camarotte, G., Cannavan, F., Cardozo, J., Chambergo, F., Ciapina, L. P., Cicarelli, R. M., Coutinho, L. L., Cursino-Santos, J. R., El-Dorry, H., Faria, J. B., Ferreira, A. J., Ferreira, R. C., Ferro, M. I., Formighieri, E. F., Franco, M. C., Greggio, C. C., Gruber, A., Katsuyama, A. M., Kishi, L. T., Leite, R. P., Lemos, E. G., Lemos, M. V., Locali, E. C., Machado, M. A., Madeira, A. M., Martinez-Rossi, N. M., Martins, E. C., Meidanis, J., Menck, C. F., Miyaki, C. Y., Moon, D. H., Moreira, L. M., Novo, M. T., Okura, V. K., Oliveira, M. C., Oliveira, V. R., Pereira, H. A., Rossi, A., Sena, J. A., Silva, C., de Souza, R. F., Spinola, L. A., Takita, M. A., Tamura, R. E., Teixeira, E. C., Tezza, R. I., Trindade dos Santos, M., Truffi, D., Tsai, S. M., White, F. F., Setubal, J. C. and Kitajima, J. P. (2002). "Comparison of the genomes of two *Xanthomonas* pathogens with differing host specificities." *Nature* **417**(6887): 459-63.

Dabrowski, S. and Kur, J. (1998). "Cloning and expression in *Escherichia coli* of the recombinant his-tagged DNA polymerases from *Pyrococcus furiosus* and *Pyrococcus woesei*." *Protein Expr Purif* **14**(1): 131-8.

Dahm, C., Muller, R., Schulte, G., Schmidt, K. and Leistner, E. (1998). "The role of isochorismate hydroxymutase genes *entC* and *menF* in enterobactin and menaquinone biosynthesis in *Escherichia coli*." *Biochim Biophys Acta* **1425**(2): 377-86.

Dale, J. W. (1998). *Molecular Genetics of Bacteria.* New York., John Wiley & Sons.

Dandekar, T., Snel, B., Huynen, M. and Bork, P. (1998). "Conservation of gene order: A fingerprint of proteins that physically interact." *Trends Biochem Sci* **23**(9): 324-8.

Darimont, B., Stehlin, C., Szadkowski, H. and Kirschner, K. (1998). "Mutational analysis of the active site of indoleglycerol phosphate synthase from *Escherichia coli*." *Protein Sci* **7**(5): 1221-32.

Davisson, V. J., Deras, I. L., Hamilton, S. E. and Moore, L. L. (1994). "A plasmid-based approach for the synthesis of a histidine biosynthetic intermediate." *Journal of Organic Chemistry* **59**(1): 137-143.

de Daruvar, A., Collado-Vides, J. and Valencia, A. (2002). "Analysis of the cellular functions of *Escherichia coli* operons and their conservation in *Bacillus subtilis*." *J Mol Evol* **55**(2): 211-21.

De Rossi, E., Leva, R., Gusberti, L., Manachini, P. L. and Riccardi, G. (1995). "Cloning, sequencing and expression of the *ilvBNC* gene cluster from *Streptomyces avermitilis*." *Gene* **166**(1): 127-32.

Demain, A. L. (1999). "Pharmaceutically active secondary metabolites of microorganisms." *Appl Microbiol Biotechnol* **52**(4): 455-63.

Demain, A. L. (2002). "Prescription for an ailing pharmaceutical industry." *Nat Biotechnol* **20**(4): 331.

Dittrich, W., Betzler, M. and Schrempf, H. (1991). "An amplifiable and deletable chloramphenicol-resistance determinant of *Streptomyces lividans* 1326 encodes a putative transmembrane protein." *Mol Microbiol* **5**(11): 2789-97.

Donadio, S., Sosio, M. and Lancini, G. (2002). "Impact of the first *Streptomyces* genome sequence on the discovery and production of bioactive substances." *Appl Microbiol Biotechnol* **60**(4): 377-80.

Doolittle, R. F. (1994). "Convergent evolution: The need to be explicit." *Trends Biochem Sci* **19**(1): 15-8.

Doolittle, R. F. (1995). "The multiplicity of domains in proteins." *Annu Rev Biochem* **64**: 287-314.

Doolittle, W. F. (1992). Evolution and basic features of gene and genome structure. In *Prokaryotic Structure and Function, a New Perspective.* S. Mohan, C. Dow and J. A. Cole. Cambridge, Cambridge University Press. 47th Symposium of the Society for General Microbiology: 23-41.

Doolittle, W. F. (1999). "Phylogenetic classification and the universal tree." *Science* **284**(5423): 2124-9.

Dosselaere, F. and Vanderleyden, J. (2001). "A metabolic node in action: Chorismate-utilizing enzymes in microorganisms." *Crit Rev Microbiol* **27**(2): 75-131.

Douangamath, A., Walker, M., Beismann-Driemeyer, S., Vega-Fernandez, M. C., Sterner, R. and Wilmanns, M. (2002). "Structural evidence for ammonia tunneling across the (beta alpha)$_8$ barrel of the imidazole glycerol phosphate synthase bienzyme complex." *Structure (Camb)* **10**(2): 185-93.

Dumas, R., Biou, V., Halgand, F., Douce, R. and Duggleby, R. G. (2001). "Enzymology, structure, and dynamics of acetohydroxy acid isomeroreductase." *Acc Chem Res* **34**(5): 399-408.

Dumas, R., Butikofer, M. C., Job, D. and Douce, R. (1995). "Evidence for two catalytically different magnesium-binding sites in acetohydroxy acid isomeroreductase by site-directed mutagenesis." *Biochemistry* **34**(18): 6026-36.

Eberhard, M., Tsai-Pflugfelder, M., Bolewska, K., Hommel, U. and Kirschner, K. (1995). "Indoleglycerol phosphate synthase-phosphoribosyl anthranilate isomerase: Comparison of the bifunctional enzyme from *Escherichia coli* with engineered monofunctional domains." *Biochemistry* **34**(16): 5419-28.

Eder, J. and Kirschner, K. (1992). "Stable substructures of eightfold beta alpha-barrel proteins: Fragment complementation of phosphoribosylanthranilate isomerase." *Biochemistry* **31**(14): 3617-25.

Egan, S., Wiener, P., Kallifidas, D. and Wellington, E. M. (1998). "Transfer of streptomycin biosynthesis gene clusters within streptomycetes isolated from soil." *Appl Environ Microbiol* **64**(12): 5061-3.

Enos-Berlage, J. L., Langendorf, M. J. and Downs, D. M. (1998). "Complex metabolic phenotypes caused by a mutation in *yjgf*, encoding a member of the highly conserved YER057C/*yjgf* family of proteins." *J Bacteriol* **180**(24): 6519-28.

Essar, D. W., Eberly, L., Hadero, A. and Crawford, I. P. (1990). "Identification and characterization of genes for a second anthranilate synthase in *Pseudomonas aeruginosa*: Interchangeability of the two anthranilate synthases and evolutionary implications." *J Bacteriol* **172**(2): 884-900.

Fani, R., Lio, P., Chiarelli, I. and Bazzicalupo, M. (1994). "The evolution of the histidine biosynthetic genes in prokaryotes: A common ancestor for the *hisA* and *hisF* genes." *J Mol Evol* **38**(5): 489-95.

Fani, R., Lio, P. and Lazcano, A. (1995). "Molecular evolution of the histidine biosynthetic pathway." *J Mol Evol* **41**(6): 760-74.

Fani, R., Mori, E., Tamburini, E. and Lazcano, A. (1998). "Evolution of the structure and chromosomal distribution of histidine biosynthetic genes." *Orig Life Evol Biosph* **28**(4-6): 555-70.

Fani, R., Tamburini, E., Mori, E., Lazcano, A., Lio, P., Barberio, C., Casalone, E., Cavalieri, D., Perito, B. and Polsinelli, M. (1997). "Paralogous histidine biosynthetic genes: Evolutionary analysis of the *Saccharomyces cerevisiae* HIS6 and HIS7 genes." *Gene* **197**(1-2): 9-17.

Farber, G. K. (1993). "An alpha/beta-barrel full of evolutionary trouble." *Current Opinion in Structural Biology* **3**(3): 409-412.

Farber, G. K. and Petsko, G. A. (1990). "The evolution of alpha/beta barrel enzymes." *Trends Biochem Sci* **15**(6): 228-34.

Fleischmann, R. D., Alland, D., Eisen, J. A., Carpenter, L., White, O., Peterson, J., DeBoy, R., Dodson, R., Gwinn, M., Haft, D., Hickey, E., Kolonay, J. F., Nelson, W. C., Umayam, L. A., Ermolaeva, M., Salzberg, S. L., Delcher, A., Utterback, T., Weidman, J., Khouri, H., Gill, J., Mikula, A., Bishai, W., Jacobs Jr, W. R., Jr., Venter, J. C. and Fraser, C. M. (2002). "Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains." *J Bacteriol* **184**(19): 5479-90.

Floss, H. G. (1997). "Natural products derived from unusual variants of the shikimate pathway." *Nat Prod Rep* **14**(5): 433-52.

Forsyth, W. R. and Matthews, C. R. (2002). "Folding mechanism of indole-3-glycerol phosphate synthase from *Sulfolobus solfataricus*: A test of the conservation of folding mechanisms hypothesis in $(beta/alpha)_8$ barrels." *J Mol Biol* **320**(5): 1119-33.

Forterre, P. (2002). "The origin of DNA genomes and DNA replication proteins." *Curr Opin Microbiol* **5**(5): 525-32.

Fowler, K. (2002) PhD Thesis. University of East Anglia.

Frodyma, M. E. and Downs, D. (1998). "The *panE* gene, encoding ketopantoate reductase, maps at 10 minutes and is allelic to *apbA* in *Salmonella typhimurium*." *J Bacteriol* **180**(17): 4757-9.

Fudou, R., Jojima, Y., Seto, A., Yamada, K., Kimura, E., Nakamatsu, T., Hiraishi, A. and Yamanaka, S. (2002). "*Corynebacterium efficiens* sp. Nov., a glutamic-acid-producing species from soil and vegetables." *Int J Syst Evol Microbiol* **52**(Pt 4): 1127-31.

Furstner, A. (2003). "Chemistry and biology of roseophilin and the prodigiosin alkaloids: A survey of the last 2500 years." *Angew Chem Int Ed Engl* **42**(31): 3582-603.

Gan, K., Gupta, S. D., Sankaran, K., Schmid, M. B. and Wu, H. C. (1993). "Isolation and characterization of a temperature-sensitive mutant of *Salmonella typhimurium* defective in prolipoprotein modification." *J Biol Chem* **268**(22): 16544-50.

Gebhardt, K., Schimana, J., Krastel, P., Dettner, K., Rheinheimer, J., Zeeck, A. and Fiedler, H. P. (2002). "Endophenazines A-D, new phenazine antibiotics from the arthropod associated endosymbiont *Streptomyces anulatus*. I. Taxonomy, fermentation, isolation and biological activities." *J Antibiot (Tokyo)* **55**(9): 794-800.

Geiger, A., Keller-Schierlein, W., Brandl, M. and Zahner, H. (1988). "Metabolites of microorganisms. 247. Phenazines from *Streptomyces antibioticus*, strain tu 2706." *J Antibiot (Tokyo)* **41**(11): 1542-51.

Gerlt, J. A. (2000). "New wine from old barrels." *Nat Struct Biol* **7**(3): 171-3.

Gerlt, J. A. and Babbitt, P. C. (1998). "Mechanistically diverse enzyme superfamilies: The importance of chemistry in the evolution of catalysis." *Curr Opin Chem Biol* **2**(5): 607-12.

Gerlt, J. A. and Babbitt, P. C. (2001). "Barrels in pieces?" *Nat Struct Biol* **8**(1): 5-7.

Gerlt, J. A. and Babbitt, P. C. (2001). "Divergent evolution of enzymatic function: Mechanistically diverse superfamilies and functionally distinct suprafamilies." *Annu Rev Biochem* **70**: 209-46.

Gerlt, J. A. and Raushel, F. M. (2003). "Evolution of function in $(beta/alpha)_8$-barrel enzymes." *Curr Opin Chem Biol* **7**(2): 252-64.

Gerstein, M. (1998). "Patterns of protein-fold usage in eight microbial genomes: A comprehensive structural census." *Proteins* **33**(4): 518-34.

Giddens, S. R., Feng, Y. and Mahanty, H. K. (2002). "Characterization of a novel phenazine antibiotic gene cluster in *Erwinia herbicola* EH1087." *Mol Microbiol* **45**(3): 769-83.

Gil, J. A. and Campelo-Diez, A. B. (2003). "Candicidin biosynthesis in *Streptomyces griseus*." *Appl Microbiol Biotechnol* **60**(6): 633-42.

Gil, J. A. and Hopwood, D. A. (1983). "Cloning and expression of a p-aminobenzoic acid synthetase gene of the candicidin-producing *Streptomyces griseus*." *Gene* **25**(1): 119-32.

Gilpin, M. L., Fulston, M., Payne, D., Cramp, R. and Hood, I. (1995). "Isolation and structure determination of two novel phenazines from a *Streptomyces* with inhibitory activity against metallo-enzymes, including metallo-beta-lactamase." *J Antibiot (Tokyo)* **48**(10): 1081-5.

Glansdorff, N. (2000). "About the last common ancestor, the universal life-tree and lateral gene transfer: A reappraisal." *Mol Microbiol* **38**(2): 177-85.

Gogarten, J. P., Doolittle, W. F. and Lawrence, J. G. (2002). "Prokaryotic evolution in light of gene transfer." *Mol Biol Evol* **19**(12): 2226-38.

Goncharoff, P. and Nichols, B. P. (1984). "Nucleotide sequence of *Escherichia coli pabB* indicates a common evolutionary origin of p-aminobenzoate synthetase and anthranilate synthetase." *J Bacteriol* **159**(1): 57-62.

Gopalakrishnan, A.S., Chen, Y.-C., Temkin, M. and Dowhan, W. "Structure and expression of the gene locus encoding the phosphatidylglycerolphosphate synthase of *Escherichia coli*." *J Biol Chem* **261**(3): 1329-1338.

Gorisch, H. (1987). "Chorismate mutase from *Streptomyces aureofaciens*." *Methods Enzymol* **142**: 463-72.

Goupil-Feuillerat, N., Cocaign-Bousquet, M., Godon, J. J., Ehrlich, S. D. and Renault, P. (1997). "Dual role of alpha-acetolactate decarboxylase in *Lactococcus lactis* subsp. Lactis." *J Bacteriol* **179**(20): 6285-93.

Grauer, D. and W.-H., L. (2000). *Fundamentals of Molecular Evolution*. Sunderland MA, Sinauer Associates, Inc.

Gu, W., Williams, D. S., Aldrich, H. C., Xie, G., Gabriel, D. W. and Jensen, R. A. (1997). "The AroQ and PheA domains of the bifunctional P-protein from *Xanthomonas campestris* in a context of genomic comparison." *Microb Comp Genomics* **2**(2): 141-58.

Gu, W., Zhao, G., Eddy, C. and Jensen, R. A. (1995). "Imidazole acetol phosphate aminotransferase in *Zymomonas mobilis*: Molecular genetic, biochemical, and evolutionary analyses." *J Bacteriol* **177**(6): 1576-84.

Gu, Z., Steinmetz, L. M., Gu, X., Scharfe, C., Davis, R. W. and Li, W. H. (2003). "Role of duplicate genes in genetic robustness against null mutations." *Nature* **421**(6918): 63-6.

Gunasekaran, K., Eyles, S. J., Hagler, A. T. and Gierasch, L. M. (2001). "Keeping it in the family: Folding studies of related proteins." *Curr Opin Struct Biol* **11**(1): 83-93.

Gupta, R. S. (2000). "The natural evolutionary relationships among prokaryotes." *Crit Rev Microbiol* **26**(2): 111-31.

Gust, B., Challis, G. L., Fowler, K., Kieser, T. and Chater, K. F. (2003). "PCR-targeted *Streptomyces* gene replacement identifies a protein domain needed for biosynthesis of the sesquiterpene soil odor geosmin." *Proc Natl Acad Sci U S A* **100**(4): 1541-6.

Hasson, M. S., Schlichting, I., Moulai, J., Taylor, K., Barrett, W., Kenyon, G. L., Babbitt, P. C., Gerlt, J. A., Petsko, G. A. and Ringe, D. (1998). "Evolution of an enzyme active site: The structure of a new crystal form of muconate lactonizing enzyme compared with mandelate racemase and enolase." *Proc Natl Acad Sci U S A* **95**(18): 10396-401.

He, J., Magarvey, N., Piraee, M. and Vining, L. C. (2001). "The gene cluster for chloramphenicol biosynthesis in *Streptomyces venezuelae* ISP5230 includes novel shikimate pathway homologues and a monomodular non-ribosomal peptide synthetase gene." *Microbiology* **147**(Pt 10): 2817-29.

Hegyi, H. and Gerstein, M. (1999). "The relationship between protein structure and function: A comprehensive survey with application to the yeast genome." *J Mol Biol* **288**(1): 147-64.

Heinz, D. W., Essen, L. O. and Williams, R. L. (1998). "Structural and mechanistic comparison of prokaryotic and eukaryotic phosphoinositide-specific phospholipases C." *J Mol Biol* **275**(4): 635-50.

Henkin, T. M. and Yanofsky, C. (2002). "Regulation by transcription attenuation in bacteria: How RNA provides instructions for transcription termination/antitermination decisions." *Bioessays* **24**(8): 700-7.

Henner, D. J., Band, L., Flaggs, G. and Chen, E. (1986). "The organization and nucleotide sequence of the *Bacillus subtilis hisH, tyrA* and *aroE* genes." *Gene* **49**(1): 147-52.

Hennig, M., Schlesier, B., Dauter, Z., Pfeffer, S., Betzel, C., Hohne, W. E. and Wilson, K. S. (1992). "A TIM barrel protein without enzymatic activity? Crystal-structure of narbonin at 1.8 A resolution." *FEBS Lett* **306**(1): 80-4.

Henn-Sax, M., Hocker, B., Wilmanns, M. and Sterner, R. (2001). "Divergent evolution of (beta/alpha)$_8$-barrel enzymes." *Biol Chem* **382**(9): 1315-20.

Henn-Sax, M., Thoma, R., Schmidt, S., Hennig, M., Kirschner, K. and Sterner, R. (2002). "Two (beta/alpha)$_8$-barrel enzymes of histidine and tryptophan biosynthesis have similar reaction mechanisms and common strategies for protecting their labile substrates." *Biochemistry* **41**(40): 12032-42.

Hesketh, A. R., Chandra, G., Shaw, A. D., Rowland, J. J., Kell, D. B., Bibb, M. J. and Chater, K. F. (2002). "Primary and secondary metabolism, and post-translational protein modifications, as portrayed by proteomic analysis of *Streptomyces coelicolor*." *Mol Microbiol* **46**(4): 917-32.

Hill, C. M. and Duggleby, R. G. (1999). "Purified recombinant *Escherichia coli* ketol-acid reductoisomerase is unsuitable for use in a coupled assay of acetohydroxyacid synthase activity due to an unexpected side reaction." *Protein Expr Purif* **15**(1): 57-61.

Himmelreich, R., Plagens, H., Hilbert, H., Reiner, B. and Herrmann, R. (1997). "Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium*." *Nucleic Acids Res* **25**(4): 701-12.

Hocker, B., Beismann-Driemeyer, S., Hettwer, S., Lustig, A. and Sterner, R. (2001). "Dissection of a (beta/alpha)$_8$-barrel enzyme into two folded halves." *Nat Struct Biol* **8**(1): 32-6.

Hocker, B., Jurgens, C., Wilmanns, M. and Sterner, R. (2001). "Stability, catalytic versatility and evolution of the (beta/alpha)$_8$-barrel fold." *Curr Opin Biotechnol* **12**(4): 376-81.

Hodgson, D. A. (1992). Differentation in actinomycetes. In *Prokaryotic Structure and Function, a New Perspective*. S. Mohan, C. Dow and J. A. Cole. Cambridge, Cambridge Univeristy Press. 47th Symposium of the Society for General Microbiology: 407-440.

Hodgson, D. A. (2000). "Primary metabolism and its control in streptomycetes: A most unusual group of bacteria." *Adv Microb Physiol* **42**: 47-238.

Hojati, Z., Milne, C., Harvey, B., Gordon, L., Borg, M., Flett, F., Wilkinson, B., Sidebottom, P. J., Rudd, B. A., Hayes, M. A., Smith, C. P. and Micklefield, J. (2002). "Structure, biosynthetic origin, and engineered biosynthesis of calcium-dependent antibiotics from *Streptomyces coelicolor*." *Chem Biol* **9**(11): 1175-87.

Holmes, D. J., Caso, J. L. and Thompson, C. J. (1993). "Autogenous transcriptional activation of a thiostrepton-induced gene in *Streptomyces lividans*." *Embo J* **12**(8): 3183-91.

Hommel, U., Eberhard, M. and Kirschner, K. (1995). "Phosphoribosyl anthranilate isomerase catalyzes a reversible Amadori reaction." *Biochemistry* **34**(16): 5429-39.

Hood, D. W., Heidstra, R., Swoboda, U. K. and Hodgson, D. A. (1992). "Molecular genetic analysis of proline and tryptophan biosynthesis in *Streptomyces coelicolor* A3(2): Interaction between primary and secondary metabolism - a review." *Gene* **115**(1-2): 5-12.

Hopwood, D. A. (1997). "Genetic contributions to understanding polyketide synthases." *Chem Rev* **97**(7): 2465-2498.

Hopwood, D. A. (1999). "Forty years of genetics with *Streptomyces*: From *in vivo* through *in vitro* to *in silico*." *Microbiology* **145**: 2183-202.

Hopwood, D. A. (2003). "The *Streptomyces* genome-be prepared!" *Nat Biotechnol* **21**(5): 505-6.

Hopwood, D. A., Bibb, M. J., Chater, K. F., Kieser, T., Bruton, C. J., Kieser, H. M., Lydiate, D. J., Smith, C. P., Ward, J. M. and Schremp, H. (1985). *Genetic Manipulation of* Streptomyces*: A Laboratory Manual*. Norwich, John Innes Foundation.

Hopwood, D. A., Chater, K. F., Dowding, J. E. and Vivian, A. (1973). "Advances in *Streptomyces coelicolor* genetics." *Bacteriol Rev* **37**(3): 371-405.

Horinouchi, S. (2002). "A microbial hormone, A-factor, as a master switch for morphological differentiation and secondary metabolism in *Streptomyces griseus*." *Front Biosci* **7**: d2045-57.

Horowitz, N. H. (1945). "On the evolution of biochemical syntheses." *Proc Natl Acad Sci U S A* **31**: 153-7.

Horowitz, N. H. (1965). The evolution of biochemical syntheses - retrospect and prospect. In *Evolving Genes and Proteins*. V. Bryson and H. J. Vogel. New York, Academic Press.

Hu, D.S. (1995) PhD Thesis. University of Warwick.

Hu, D. S., Hood, D. W., Heidstra, R. and Hodgson, D. A. (1999). "The expression of the *trpD*, *trpC* and *trpBA* genes of *Streptomyces coelicolor* A3(2) is regulated by growth rate and growth phase but not by feedback repression." *Mol Microbiol* **32**(4): 869-80.

Huang, J., Lih, C. J., Pan, K. H. and Cohen, S. N. (2001). "Global analysis of growth phase responsive gene expression and regulation of antibiotic biosynthetic pathways in *Streptomyces coelicolor* using DNA microarrays." *Genes Dev* **15**(23): 3183-92.

Huddleston, A. S., Cresswell, N., Neves, M. C., Beringer, J. E., Baumberg, S., Thomas, D. I. and Wellington, E. M. (1997). "Molecular detection of streptomycin-producing streptomycetes in brazilian soils." *Appl Environ Microbiol* **63**(4): 1288-97.

Hughes, A. L. (1994). "The evolution of functionally novel proteins after gene duplication." *Proc R Soc Lond B Biol Sci* **256**(1346): 119-24.

Huynen, M., Snel, B., Lathe, W., 3rd and Bork, P. (2000). "Predicting protein function by genomic context: Quantitative evaluation and qualitative inferences." *Genome Res* **10**(8): 1204-10.

Huynen, M., Snel, B., Lathe, W. and Bork, P. (2000). "Exploitation of gene context." *Curr Opin Struct Biol* **10**(3): 366-70.

Ikeda, H., Ishikawa, J., Hanamoto, A., Shinose, M., Kikuchi, H., Shiba, T., Sakaki, Y., Hattori, M. and Omura, S. (2003). "Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*." *Nat Biotechnol* **21**(5): 526-31.

Itoh, T., Takemoto, K., Mori, H. and Gojobori, T. (1999). "Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes." *Mol Biol Evol* **16**(3): 332-46.

Jackson, M., Crick, D. C. and Brennan, P. J. (2000). "Phosphatidylinositol is an essential phospholipid of mycobacteria." *J Biol Chem* **275**(39): 30092-9.

Jacob, F., Perrin, D., Sánchez, C. and Monod, J. (1960). "L'opéron: Groupe de gènes à expression coordonné par un opérateur." *C.R. Acad. Sci.* **250**: 1727-9.

Jain, R., Rivera, M. C., Moore, J. E. and Lake, J. A. (2002). "Horizontal gene transfer in microbial genome evolution." *Theor Popul Biol* **61**(4): 489-95.

James, T. Y., Boulianne, R. P., Bottoli, A. P., Granado, J. D., Aebi, M. and Kues, U. (2002). "The *pab1* gene of *Coprinus cinereus* encodes a bifunctional protein for para-aminobenzoic acid (PABA) synthesis: Implications for the evolution of fused PabA synthases." *J Basic Microbiol* **42**(2): 91-103.

Jasanoff, A., Davis, B. and Fersht, A. R. (1994). "Detection of an intermediate in the folding of the (beta/alpha)$_8$-barrel n-(5'-phosphoribosyl)anthranilate isomerase from *Escherichia coli*." *Biochemistry* **33**(20): 6350-5.

Jeffery, C. J. (1999). "Moonlighting proteins." *Trends Biochem Sci* **24**(1): 8-11.

Jensen, R. A. (1976). "Enzyme recruitment in evolution of new function." *Annu Rev Microbiol* **30**: 409-25.

Jensen, R. A. and Gu, W. (1996). "Evolutionary recruitment of biochemically specialized subdivisions of family I within the protein superfamily of aminotransferases." *J Bacteriol* **178**(8): 2161-71.

Jeong, W., Cha, M. K. and Kim, I. H. (2000). "Thioredoxin-dependent hydroperoxide peroxidase activity of bacterioferritin comigratory protein (BCP) as a new member of the thiol-specific antioxidant protein (TSA)/alkyl hydroperoxide peroxidase c (AHPC) family." *J Biol Chem* **275**(4): 2924-30.

Juers, D. H., Huber, R. E. and Matthews, B. W. (1999). "Structural comparisons of TIM barrel proteins suggest functional and evolutionary relationships between beta-galactosidase and other glycohydrolases." *Protein Sci* **8**(1): 122-36.

Jung, S. I., Han, M. S., Kwon, J. H., Cheon, C. I., Min, K. H. and Lee, M. S. (1998). "Cloning of the histidine biosynthetic genes of *Corynebacterium glutamicum*: Organization and sequencing analysis of the *hisA*, *impA*, and *hisF* gene cluster." *Biochem Biophys Res Commun* **247**(3): 741-5.

Jurgens, C., Strom, A., Wegener, D., Hettwer, S., Wilmanns, M. and Sterner, R. (2000). "Directed evolution of a (beta/alpha)$_8$-barrel enzyme to catalyze related reactions in two different metabolic pathways." *Proc Natl Acad Sci U S A* **97**(18): 9925-30.

Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., Naruo, K., Okumura, S., Shimpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M. and Tabata, S. (1996). "Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. Strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions." *DNA Res* **3**(3): 109-36.

Kaplan, J. B., Goncharoff, P., Seibold, A. M. and Nichols, B. P. (1984). "Nucleotide sequence of the *Acinetobacter calcoaceticus trpGDC* gene cluster." *Mol Biol Evol* **1**(6): 456-72.

Kempf, B. and Bremer, E. (1998). "Uptake and synthesis of compatible solutes as microbial stress responses to high-osmolality environments." *Arch Microbiol* **170**(5): 319-30.

Kempter, C., Kaiser, D., Haag, S., Nicholson, G., Gnau, V., Walk, T., Gierling, G. H., Decker, H., Zahner, H. and Jung, G. (1997). "CDA: Calcium-dependent peptide antibiotics from *Streptomyces coelicolor* A3(2) containing unusual residues." *Angew. Chem.Int. Ed. Engl.* **36**: 498-501.

Kenklies, J., Ziehn, R., Fritsche, K., Pich, A. and Andreesen, J. R. (1999). "Proline biosynthesis from l-ornithine in *Clostridium sticklandii*: Purification of delta1-pyrroline-5-carboxylate reductase, and sequence and expression of the encoding gene, *proC*." *Microbiology* **145**: 819-26.

Kessler, A., Dittrich, W., Betzler, M. and Schrempf, H. (1989). "Cloning and analysis of a deletable tetracycline-resistance determinant of *Streptomyces lividans* 1326." *Mol Microbiol* **3**(8): 1103-9.

Khokhlov, A. S., Tovarova, II, Borisova, L. N., Pliner, S. A., Shevchenko, L. N., Kornitskaia, E., Ivkina, N. S. and Rapoport, I. A. (1967). "[the A-factor, responsible for streptomycin biosynthesis by mutant strains of actinomyces streptomycini]." *Dokl Akad Nauk SSSR* **177**(1): 232-5.

Khosla, C. (2000). "Natural product biosynthesis: A new interface between enzymology and medicine." *J Org Chem* **65**(24): 8127-33.

Kieser, T., Bibb, M. J., Buttner, M. J., Chater, K. J. and Hopwood, D. A. (2000). *Practical* Streptomyces *Genetics*. Norwich, John Innes Foundation.

Kieser, T., Hopwood, D. A., Wright, H. M. and Thompson, C. J. (1982). "pIJ101, a multi-copy broad host-range *Streptomyces* plasmid: Functional analysis and development of DNA cloning vectors." *Mol Gen Genet* **185**(2): 223-8.

Kim, W. G., Ryoo, I. J., Yun, B. S., Shin-ya, K., Seto, H. and Yoo, I. D. (1997). "New diphenazines with neuronal cell protecting activity, phenazostatins A and B, produced by *Streptomyces* sp." *J Antibiot (Tokyo)* **50**(9): 715-21.

Kirschner, K., Szadkowski, H., Jardetzky, T. S. and Hager, V. (1987). "Phosphoribosylanthranilate isomerase-indoleglycerol-phosphate synthase from *Escherichia coli*." *Methods Enzymol* **142**: 386-97.

Klem, T. J. and Davisson, V. J. (1993). "Imidazole glycerol phosphate synthase: The glutamine amidotransferase in histidine biosynthesis." *Biochemistry* **32**(19): 5177-86.

Knaggs, A. R. (2001). "The biosynthesis of shikimate metabolites." *Nat Prod Rep* **18**(3): 334-55.

Knaggs, A. R. (2003). "The biosynthesis of shikimate metabolites." *Nat Prod Rep* **20**(1): 119-36.

Knochel, T., Ivens, A., Hester, G., Gonzalez, A., Bauerle, R., Wilmanns, M., Kirschner, K. and Jansonius, J. N. (1999). "The crystal structure of anthranilate synthase from *Sulfolobus solfataricus*: Functional implications." *Proc Natl Acad Sci U S A* **96**(17): 9479-84.

Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I. and Koonin, E. V. (2002). "Selection in the evolution of gene duplications." *Genome Biol* **3**(2): RESEARCH0008.

Koonin, E. V. and Galperin, M. Y. (1997). "Prokaryotic genomes: The emerging paradigm of genome-based microbiology." *Curr Opin Genet Dev* **7**(6): 757-63.

Koonin, E. V., Mushegian, A. R. and Bork, P. (1996). "Non-orthologous gene displacement." *Trends Genet* **12**(9): 334-6.

Krastel, P., Zeeck, A., Gebhardt, K., Fiedler, H. P. and Rheinheimer, J. (2002). "Endophenazines A-D, new phenazine antibiotics from the athropod associated endosymbiont *Streptomyces anulatus* II. Structure elucidation." *J Antibiot (Tokyo)* **55**(9): 801-6.

Lang, D., Thoma, R., Henn-Sax, M., Sterner, R. and Wilmanns, M. (2000). "Structural evidence for evolution of the beta/alpha barrel scaffold by gene duplication and fusion." *Science* **289**(5484): 1546-50.

LaRossa, R. A., Van Dyk, T. K. and Smulski, D. R. (1987). "Toxic accumulation of alpha-ketobutyrate caused by inhibition of the branched-chain amino acid biosynthetic enzyme acetolactate synthase in *Salmonella typhimurium*." *J Bacteriol* **169**(4): 1372-8.

Lathe, W. C., 3rd, Snel, B. and Bork, P. (2000). "Gene context conservation of a higher order than operons." *Trends Biochem Sci* **25**(10): 474-9.

Lautru, S. and Challis, G. L. (2003). "Understanding mechanisms of substrate recognition by non-ribosomal peptide synthetase multienzymes: The keys to exploiting microbial biosynthesis?" *Microbiology*. **In press.**

Lawrence, J. (1999). "Selfish operons: The evolutionary impact of gene clustering in prokaryotes and eukaryotes." *Curr Opin Genet Dev* **9**(6): 642-8.

Lawrence, J. G. and Ochman, H. (1998). "Molecular archaeology of the *Escherichia coli* genome." *Proc Natl Acad Sci U S A* **95**(16): 9413-7.

Lawrence, J. G. and Ochman, H. (2002). "Reconciling the many faces of lateral gene transfer." *Trends Microbiol* **10**(1): 1-4.

Lawrence, J. G. and Roth, J. R. (1996). "Selfish operons: Horizontal transfer may drive the evolution of gene clusters." *Genetics* **143**(4): 1843-60.

Lazcano, A. and Miller, S. L. (1999). "On the origin of metabolic pathways." *J Mol Evol* **49**(4): 424-31.

Leblond, P., Demuyter, P., Simonet, J. M. and Decaris, B. (1990). "Genetic instability and hypervariability in *Streptomyces ambofaciens*: Towards an understanding of a mechanism of genome plasticity." *Mol Microbiol* **4**(5): 707-14.

Lesk, A. M., Branden, C. I. and Chothia, C. (1989). "Structural principles of alpha/beta barrel proteins: The packing of the interior of the sheet." *Proteins* **5**(2): 139-48.

Leskela, S., Wahlstrom, E., Kontinen, V. P. and Sarvas, M. (1999). "Lipid modification of prelipoproteins is dispensable for growth but essential for efficient protein secretion in *Bacillus subtilis*: Characterization of the *lgt* gene." *Mol Microbiol* **31**(4): 1075-85.

Lewis, E. B. (1951). *Cold Spring Harbor Symp. Quant. Biol.* **16**: 159-63.

Limauro, D., Avitabile, A., Cappellano, C., Puglia, A. M. and Bruni, C. B. (1990). "Cloning and characterization of the histidine biosynthetic gene cluster of *Streptomyces coelicolor* A3(2)." *Gene* **90**(1): 31-41.

Limauro, D., Avitabile, A., Puglia, A. M. and Bruni, C. B. (1992). "Further characterization of the histidine gene cluster of *Streptomyces coelicolor* A3(2): Nucleotide sequence and transcriptional analysis of *hisD*." *Res Microbiol* **143**(7): 683-93.

Lin, C., Paradkar, A. S. and Vining, L. C. (1998). "Regulation of an anthranilate synthase gene in *Streptomyces venezuelae* by a *trp* attenuator." *Microbiology* **144**: 1971-80.

Lin, Y. S., Kieser, H. M., Hopwood, D. A. and Chen, C. W. (1993). "The chromosomal DNA of *Streptomyces lividans* 66 is linear." *Mol Microbiol* **10**(5): 923-33.

Lindwall, G., Chau, M., Gardner, S. R. and Kohlstaedt, L. A. (2000). "A sparse matrix approach to the solubilization of overexpressed proteins." *Protein Eng* **13**(1): 67-71.

Ling, M., Allen, S. W. and Wood, J. M. (1994). "Sequence analysis identifies the proline dehydrogenase and delta 1-pyrroline-5-carboxylate dehydrogenase domains of the multifunctional *Escherichia coli* PutA protein." *J Mol Biol* **243**(5): 950-6.

Luger, K., Hommel, U., Herold, M., Hofsteenge, J. and Kirschner, K. (1989). "Correct folding of circularly permuted variants of a beta alpha barrel enzyme *in vivo*." *Science* **243**(4888): 206-10.

Luger, K., Szadkowski, H. and Kirschner, K. (1990). "An 8-fold beta alpha barrel protein with redundant folding possibilities." *Protein Eng* **3**(4): 249-58.

Lynch, M. and Conery, J. S. (2000). "The evolutionary fate and consequences of duplicate genes." *Science* **290**(5494): 1151-5.

Manderville, R. A. (2001). "Synthesis, proton-affinity and anti-cancer properties of the prodigiosin-group natural products." *Curr Med Chem Anti-Canc Agents* **1**(2): 195-218.

Margolies, M. N. and Goldberger, R. F. (1966). "Isolation of the fourth (isomerase) of histidine biosynthesis from *Salmonella typhimurium*." *J Biol Chem* **241**(14): 3262-9.

Margolies, M. N. and Goldberger, R. F. (1967). "Physical and chemical characterization of the isomerase of histidine biosynthesis in *Salmonella typhimurium*." *J Biol Chem* **242**(2): 256-64.

Martin, W. (1999). "Mosaic bacterial chromosomes: A challenge en route to a tree of genomes." *Bioessays* **21**(2): 99-104.

Matney, T. S., Goldschmidt, E. P., Erwin, N. S. and Scroggs, R. A. (1964). "A preliminary map of genomic sites for F-attachment in *Escherichia coli* K12." *Biochem Biophys Res Commun* **17**(3): 278-81.

Matsui, K., Sano, K. and Ohtsubo, E. (1986). "Complete nucleotide and deduced amino-acid-sequences of the *Brevibacterium lactofermentum* tryptophan operon." *Nucleic Acids Research* **14**(24): 10113-10114.

Mavrodi, D. V., Bonsall, R. F., Delaney, S. M., Soule, M. J., Phillips, G. and Thomashow, L. S. (2001). "Functional analysis of genes for biosynthesis of pyocyanin and phenazine-1-carboxamide from *Pseudomonas aeruginosa* PAO1." *J Bacteriol* **183**(21): 6454-65.

Mavrodi, D. V., Ksenzenko, V. N., Bonsall, R. F., Cook, R. J., Boronin, A. M. and Thomashow, L. S. (1998). "A seven-gene locus for synthesis of phenazine-1-carboxylic acid by *Pseudomonas fluorescens* 2-79." *J Bacteriol* **180**(9): 2541-8.

Mayans, O., Ivens, A., Nissen, L. J., Kirschner, K. and Wilmanns, M. (2002). "Structural analysis of two enzymes catalysing reverse metabolic reactions implies common ancestry." *Embo Journal* **21**(13): 3245-3254.

McDonald, M., Mavrodi, D. V., Thomashow, L. S. and Floss, H. G. (2001). "Phenazine biosynthesis in *Pseudomonas fluorescens*: Branchpoint from the primary shikimate biosynthetic pathway and role of phenazine-1,6-dicarboxylic acid." *J Am Chem Soc* **123**(38): 9459-60.

Mehta, P. K., Hale, T. I. and Christen, P. (1993). "Aminotransferases - demonstration of homology and division into evolutionary subgroups." *European Journal of Biochemistry* **214**(2): 549-561.

Meyer, A. (2003). "Molecular evolution: Duplication, duplication." *Nature* **421**(6918): 31-2.

Miles, E. W., Yutani, K. and Ogasahara, K. (1982). "Guanidine hydrochloride induced unfolding of the alpha subunit of tryptophan synthase and of the two alpha proteolytic fragments: Evidence for stepwise unfolding of the two alpha domains." *Biochemistry* **21**(11): 2586-92.

Mira, A., Klasson, L. and Andersson, S. G. (2002). "Microbial genome evolution: Sources of variability." *Curr Opin Microbiol* **5**(5): 506-12.

Miroux, B. and Walker, J. E. (1996). "Over-production of proteins in *Escherichia coli*: Mutant hosts that allow synthesis of some membrane proteins and globular proteins at high levels." *J Mol Biol* **260**(3): 289-98.

Miyadoh, S. (1997). *Atlas of Actinomycetes*. Japan, Asakura Publising Co.

Morollo, A. A. and Bauerle, R. (1993). "Characterization of composite aminodeoxyisochorismate synthase and aminodeoxyisochorismate lyase activities of anthranilate synthase." *Proc Natl Acad Sci U S A* **90**(21): 9983-7.

Morollo, A. A. and Eck, M. J. (2001). "Structure of the cooperative allosteric anthranilate synthase from *Salmonella typhimurium*." *Nat Struct Biol* **8**(3): 243-7.

Murzin, A. G. (1998). "How far divergent evolution goes in proteins." *Curr Opin Struct Biol* **8**(3): 380-7.

Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia, C. (1995). "SCOP: A structural classification of proteins database for the investigation of sequences and structures." *J Mol Biol* **247**(4): 536-40.

Mushegian, A. R. and Koonin, E. V. (1996). "Gene order is not conserved in bacterial evolution." *Trends Genet* **12**(8): 289-90.

Nagano, N., Hutchinson, E. G. and Thornton, J. M. (1999). "Barrel structures in proteins: Automatic identification and classification including a sequence analysis of TIM barrels." *Protein Sci* **8**(10): 2072-84.

Nagano, N., Orengo, C. A. and Thornton, J. M. (2002). "One fold with many functions: The evolutionary relationships between TIM barrel families based on their sequences, structures and functions." *J Mol Biol* **321**(5): 741-65.

Neidhart, D. J., Howell, P. L., Petsko, G. A., Powers, V. M., Li, R. S., Kenyon, G. L. and Gerlt, J. A. (1991). "Mechanism of the reaction catalyzed by mandelate racemase. 2. Crystal structure of mandelate racemase at 2.5-A resolution: Identification of the active site and possible catalytic residues." *Biochemistry* **30**(38): 9264-73.

Nester, E. W. and Montoya, A. L. (1976). "An enzyme common to histidine and aromatic amino acid biosynthesis in *Bacillus subtilis*." *J Bacteriol* **126**(2): 699-705.

Nichols, B. P. and Green, J. M. (1992). "Cloning and sequencing of *Escherichia coli ubiC* and purification of chorismate lyase." *J Bacteriol* **174**(16): 5309-16.

Nigou, J., Dover, L. G. and Besra, G. S. (2002). "Purification and biochemical characterization of *Mycobacterium tuberculosis suhB*, an inositol monophosphatase involved in inositol biosynthesis." *Biochemistry* **41**(13): 4392-8.

Noland, B. W., Dangott, L. J. and Baldwin, T. O. (1999). "Folding, stability, and physical properties of the alpha subunit of bacterial luciferase." *Biochemistry* **38**(49): 16136-45.

Novotna, J., Vohradsky, J., Berndt, P., Gramajo, H., Langen, H., Li, X. M., Minas, W., Orsaria, L., Roeder, D. and Thompson, C. J. (2003). "Proteomic studies of diauxic lag in the differentiating prokaryote *Streptomyces coelicolor* reveal a regulatory network of stress-induced proteins and central metabolic enzymes." *Mol Microbiol* **48**(5): 1289-303.

O'Brien, P. J. and Herschlag, D. (1999). "Catalytic promiscuity and the evolution of new enzymatic activities." *Chem Biol* **6**(4): R91-R105.

Ochman, H., Lawrence, J. G. and Groisman, E. A. (2000). "Lateral gene transfer and the nature of bacterial innovation." *Nature* **405**(6784): 299-304.

Ohno, S. (1970). *Evolution by Gene Duplication*. Heidelberg, Springer-Verlag.

Ohyama, T., Igarashi, K. and Kobayashi, H. (1994). "Physiological role of the *chaA* gene in sodium and calcium circulations at a high pH in *Escherichia coli*." *J Bacteriol* **176**(14): 4311-5.

Omura, S., Ikeda, H., Ishikawa, J., Hanamoto, A., Takahashi, C., Shinose, M., Takahashi, Y., Horikawa, H., Nakazawa, H., Osonoe, T., Kikuchi, H., Shiba, T., Sakaki, Y. and Hattori, M. (2001). "Genome sequence of an industrial microorganism *Streptomyces avermitilis*: Deducing the ability of producing secondary metabolites." *Proc Natl Acad Sci U S A* **98**(21): 12215-20.

Orengo, C. A., Jones, D. T. and Thornton, J. M. (1994). "Protein superfamilies and domain superfolds." *Nature* **372**(6507): 631-4.

Osterman, A. and Overbeek, R. (2003). "Missing genes in metabolic pathways: A comparative genomics approach." *Curr Opin Chem Biol* **7**(2): 238-51.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. and Maltsev, N. (1999). "The use of gene clusters to infer functional coupling." *Proc Natl Acad Sci U S A* **96**(6): 2896-901.

Ozenberger, B. A., Brickman, T. J. and McIntosh, M. A. (1989). "Nucleotide sequence of *Escherichia coli* isochorismate synthetase gene entc and evolutionary relationship of isochorismate synthetase and other chorismate-utilizing enzymes." *J Bacteriol* **171**(2): 775-83.

Panina, E. M., Vitreschak, A. G., Mironov, A. A. and Gelfand, M. S. (2003). "Regulation of biosynthesis and transport of aromatic amino acids in low-GC Gram-positive bacteria." *FEMS Microbiol Lett* **222**(2): 211-20.

Paradkar, A., Trefzer, A., Chakraburtty, R. and Stassi, D. (2003). "*Streptomyces* genetics: A genomic perspective." *Crit Rev Biotechnol* **23**(1): 1-27.

Paradkar, A. S., Stuttard, C. and Vining, L. C. (1991). "Molecular cloning of the genes for anthranilate synthetase from *Streptomyces venezuelae* ISP5230." *FEMS Microbiol Lett* **62**(2-3): 177-81.

Paradkar, A. S., Vining, L. C. and Stuttard, C. (1991). "Characterization of tryptophan-requiring auxotrophs of *Streptomyces venezuelae* ISP5230." *Canadian Journal of Microbiology* **37**(5): 333-338.

Parish, T., Liu, J., Nikaido, H. and Stoker, N. G. (1997). "A *Mycobacterium smegmatis* mutant with a defective inositol monophosphate phosphatase gene homolog has altered cell envelope permeability." *J Bacteriol* **179**(24): 7827-33.

Parsons, J. F., Jensen, P. Y., Pachikara, A. S., Howard, A. J., Eisenstein, E. and Ladner, J. E. (2002). "Structure of *Escherichia* coli aminodeoxychorismate synthase: Architectural conservation and diversity in chorismate-utilizing enzymes." *Biochemistry* **41**(7): 2198-208.

Parsot, C. (1987). "A common origin for enzymes involved in the terminal step of the threonine and tryptophan biosynthetic pathways." *Proc Natl Acad Sci U S A* **84**(15): 5207-10.

Perona, J. J. and Craik, C. S. (1997). "Evolutionary divergence of substrate specificity within the chymotrypsin-like serine protease fold." *J Biol Chem* **272**(48): 29987-90.

Petsko, G. A. (2000). "Enzyme evolution. Design by necessity." *Nature* **403**(6770): 606-7.

Piatigorsky, J. (2003). "Crystallin genes: Specialization by changes in gene regulation may precede gene duplication." *J Struct Funct Genomics* **3**(1-4): 131-7.

Piatigorsky, J. and Wistow, G. (1991). "The recruitment of crystallins: New functions precede gene duplication." *Science* **252**(5010): 1078-9.

Pierson, L. S., 3rd, Gaffney, T., Lam, S. and Gong, F. (1995). "Molecular analysis of genes encoding phenazine biosynthesis in the biological control bacterium *Pseudomonas aureofaciens* 30-84." *FEMS Microbiol Lett* **134**(2-3): 299-307.

Priestle, J. P., Grutter, M. G., White, J. L., Vincent, M. G., Kania, M., Wilson, E., Jardetzky, T. S., Kirschner, K. and Jansonius, J. N. (1987). "Three-dimensional structure of the bifunctional enzyme n-(5'-phosphoribosyl)anthranilate isomerase-indole-3-glycerol-phosphate synthase from *Escherichia coli*." *Proc Natl Acad Sci U S A* **84**(16): 5690-4.

Primerano, D. A. and Burns, R. O. (1983). "Role of acetohydroxy acid isomeroreductase in biosynthesis of pantothenic acid in *Salmonella typhimurium*." *J Bacteriol* **153**(1): 259-69.

Primrose, S. B. and Twyman, R. M. (2003). *Principles of Genome Analysis and Genomics*. Oxford, Blackwell Publishing.

Prince, V. E. and Pickett, F. B. (2002). "Splitting pairs: The diverging fates of duplicated genes." *Nat Rev Genet* **3**(11): 827-37.

Pujadas, G. and Palau, J. (1999). "TIM barrel fold: Structural, functional and evolutionary characteristics in natural and designed molecules." *Biologia* **54**(3): 231-253.

Pusecker, K., Laatsch, H., Helmke, E. and Weyland, H. (1997). "Dihydrophencomycin methyl ester, a new phenazine derivative from a marine streptomycete." *J Antibiot (Tokyo)* **50**(6): 479-83.

Raine, A. R., Scrutton, N. S. and Mathews, F. S. (1994). "On the evolution of alternate core packing in eightfold beta/alpha-barrels." *Protein Sci* **3**(10): 1889-92.

Rao, V., Guan, C. and Van Roey, P. (1995). "Crystal structure of endo-beta-n-acetylglucosaminidase H at 1.9 A resolution: Active-site geometry and substrate recognition." *Structure* **3**(5): 449-57.

Rappu, P., Shin, B. S., Zalkin, H. and Mantsala, P. (1999). "A role for a highly conserved protein of unknown function in regulation of *Bacillus subtilis purA* by the purine repressor." *J Bacteriol* **181**(12): 3810-5.

Reardon, D. and Farber, G. K. (1995). "The structure and evolution of alpha/beta barrel proteins." *Faseb J* **9**(7): 497-503.

Redenbach, M., Kieser, H. M., Denapaite, D., Eichner, A., Cullum, J., Kinashi, H. and Hopwood, D. A. (1996). "A set of ordered cosmids and a detailed genetic and physical map for the 8 Mb *Streptomyces coelicolor* A3(2) chromosome." *Mol Microbiol* **21**(1): 77-96.

Rison, S. C., Teichmann, S. A. and Thornton, J. M. (2002). "Homology, pathway distance and chromosomal localization of the small molecule metabolism enzymes in *Escherichia coli*." *J Mol Biol* **318**(3): 911-32.

Rison, S. C. and Thornton, J. M. (2002). "Pathway evolution, structurally speaking." *Curr Opin Struct Biol* **12**(3): 374-82.

Rogozin, I. B., Makarova, K. S., Murvai, J., Czabarka, E., Wolf, Y. I., Tatusov, R. L., Szekely, L. A. and Koonin, E. V. (2002). "Connected gene neighborhoods in prokaryotic genomes." *Nucleic Acids Res* **30**(10): 2212-23.

Rowland, B. M., Grossman, T. H., Osburne, M. S. and Taber, H. W. (1996). "Sequence and genetic organization of a *Bacillus subtilis* operon encoding 2,3-dihydroxybenzoate biosynthetic enzymes." *Gene* **178**(1-2): 119-23.

Roy, S. (1999). "Multifunctional enzymes and evolution of biosynthetic pathways: Retro-evolution by jumps." *Proteins* **37**(2): 303-9.

Ryding, N. J., Anderson, T. B. and Champness, W. C. (2002). "Regulation of the *Streptomyces coelicolor* calcium-dependent antibiotic by *absA*, encoding a cluster-linked two-component system." *J Bacteriol* **184**(3): 794-805.

Sambrook, J., Fritsche, K. and Maniatis, T. (1989). *Molecular Cloning: A Laboratory Manual*. New York, Cold Spring Harbor Press.

Sambrook, J. and Russell, D. W. (2000). *Molecular Cloning: A Laboratory Manual*. New York, Cold Spring Harbor Press.

Sanchez del Pino, M. M. and Fersht, A. R. (1997). "Nonsequential unfolding of the alpha/beta barrel protein indole-3-glycerol-phosphate synthase." *Biochemistry* **36**(18): 5560-5.

Sankaran, K. and Wu, H. C. (1994). "Lipid modification of bacterial prolipoprotein. Transfer of diacylglyceryl moiety from phosphatidylglycerol." *J Biol Chem* **269**(31): 19701-6.

Schell, M. A., Karmirantzou, M., Snel, B., Vilanova, D., Berger, B., Pessi, G., Zwahlen, M. C., Desiere, F., Bork, P., Delley, M., Pridmore, R. D. and Arigoni, F. (2002). "The genome sequence of *Bifidobacterium longum* reflects its adaptation to the human gastrointestinal tract." *Proc Natl Acad Sci U S A* **99**(22): 14422-7.

Serebrijski, I., Reyes, O. and Leblon, G. (1995). "Corrected gene assignments of *Escherichia coli pro* mutations." *J Bacteriol* **177**(24): 7261-4.

Serebrijski, I., Wojcik, F., Reyes, O. and Leblon, G. (1995). "Multicopy suppression by *asd* gene and osmotic stress-dependent complementation by heterologous *proA* in *proA* mutants." *J Bacteriol* **177**(24): 7255-60.

Silverman, J. A., Balakrishnan, R. and Harbury, P. B. (2001). "Reverse engineering the (beta/alpha)$_8$ barrel fold." *Proc Natl Acad Sci U S A* **98**(6): 3092-7.

Silverman, J. A. and Harbury, P. B. (2002). "The equilibrium unfolding pathway of a (beta/alpha)$_8$ barrel." *J Mol Biol* **324**(5): 1031-40.

Sinha, S., Rappu, P., Lange, S. C., Mantsala, P., Zalkin, H. and Smith, J. L. (1999). "Crystal structure of *Bacillus subtilis yabJ*, a purine regulatory protein and member of the highly conserved *yjgF* family." *Proc Natl Acad Sci U S A* **96**(23): 13074-9.

Sinha, S. C. and Smith, J. L. (2001). "The PRT protein family." *Curr Opin Struct Biol* **11**(6): 733-9.

Slock, J., Stahly, D. P., Han, C. Y., Six, E. W. and Crawford, I. P. (1990). "An apparent *Bacillus subtilis* folic acid biosynthetic operon containing *pab*, an amphibolic *trpG* gene, a third gene required for synthesis of para-aminobenzoic acid, and the dihydropteroate synthase gene." *J Bacteriol* **172**(12): 7211-26.

Smith, D. D., Wood, N. J. and Hodgson, D. A. (1995). "Interaction between primary and secondary metabolism in *Streptomyces coelicolor* A3(2): Role of pyrroline-5-carboxylate dehydrogenase." *Microbiology* **141 (Pt 7)**: 1739-44.

Smithers, C. M. and Engel, P. P. (1974). "Gene-enzyme relations of tryptophan mutants in *Streptomyces coelicolor* A3(2)." *Genetics* **78**(3): 799-808.

Snel, B., Bork, P. and Huynen, M. A. (2002). "The identification of functional modules from the genomic association of genes." *Proc Natl Acad Sci U S A* **99**(9): 5890-5.

Snel, B., Lehmann, G., Bork, P. and Huynen, M. A. (2000). "STRING: A web-server to retrieve and display the repeatedly occurring neighbourhood of a gene." *Nucleic Acids Res* **28**(18): 3442-4.

Sonnhammer, E. L. and Koonin, E. V. (2002). "Orthology, paralogy and proposed classification for paralog subtypes." *Trends Genet* **18**(12): 619-20.

Soto, M. J., Jimenez-Zurdo, J. I., van Dillewijn, P. and Toro, N. (2000). "*Sinorhizobium meliloti putA* gene regulation: A new model within the family rhizobiaceae." *J Bacteriol* **182**(7): 1935-41.

Spraggon, G., Kim, C., Nguyen-Huu, X., Yee, M. C., Yanofsky, C. and Mills, S. E. (2001). "The structures of anthranilate synthase of *Serratia marcescens* crystallized in the presence of (i) its substrates, chorismate and glutamine, and a product, glutamate, and (ii) its end-product inhibitor, l-tryptophan." *Proc Natl Acad Sci U S A* **98**(11): 6021-6.

Sterner, R., Merz, A., Thoma, R. and Kirschner, K. (2001). "Phosphoribosylanthranilate isomerase and indoleglycerol-phosphate synthase: Tryptophan biosynthetic enzymes from *Thermotoga maritima*." *Methods Enzymol* **331**: 270-80.

Subramaniam, P., Bhatnagar, R., Hooper, A. and Jensen, R. A. (1994). "The dynamic progression of evolved character states for aromatic amino acid biosynthesis in Gram-negative bacteria." *Microbiology* **140 (Pt 12)**: 3431-40.

Surber, M. W. and Maloy, S. (1998). "The *putA* protein of *Salmonella typhimurium* catalyzes the two steps of proline degradation via a leaky channel." *Arch Biochem Biophys* **354**(2): 281-7.

Takano, E., White, J., Thompson, C. J. and Bibb, M. J. (1995). "Construction of thiostrepton-inducible, high-copy-number expression vectors for use in *Streptomyces* spp." *Gene* **166**(1): 133-7.

Tamames, J., Casari, G., Ouzounis, C. and Valencia, A. (1997). "Conserved clusters of functionally related genes in two bacterial genomes." *J Mol Evol* **44**(1): 66-73.

Tatum, E. L. (1965). Opening address in *Evolving Genes and Proteins*. V. Bryson and H. J. Vogel. New York, Academic Press.

Teichmann, S. A., Rison, S. C., Thornton, J. M., Riley, M., Gough, J. and Chothia, C. (2001). "The evolution and structural anatomy of the small molecule metabolic pathways in *Escherichia coli*." *J Mol Biol* **311**(4): 693-708.

Teichmann, S. A., Rison, S. C., Thornton, J. M., Riley, M., Gough, J. and Chothia, C. (2001). "Small-molecule metabolism: An enzyme mosaic." *Trends Biotechnol* **19**(12): 482-6.

Thoma, R., Obmolova, G., Lang, D. A., Schwander, M., Jeno, P., Sterner, R. and Wilmanns, M. (1999). "Efficient expression, purification and crystallisation of two hyperthermostable enzymes of histidine biosynthesis." *FEBS Lett* **454**(1-2): 1-6.

Thoma, R., Schwander, M., Liebl, W., Kirschner, K. and Sterner, R. (1998). "A histidine gene cluster of the hyperthermophile *Thermotoga maritima*: Sequence analysis and evolutionary significance." *Extremophiles* **2**(4): 379-89.

Thomas, M. G., Burkart, M. D. and Walsh, C. T. (2002). "Conversion of l-proline to pyrrolyl-2-carboxyl-S-PCP during undecylprodigiosin and pyoluteorin biosynthesis." *Chem Biol* **9**(2): 171-84.

Thompson, C. J., Fink, D. and Nguyen, L. D. (2002). "Principles of microbial alchemy: Insights from the *Streptomyces coelicolor* genome sequence." *Genome Biol* **3**(7): REVIEWS1020.

Todd, A. E., Orengo, C. A. and Thornton, J. M. (2001). "Evolution of function in protein superfamilies, from a structural perspective." *J Mol Biol* **307**(4): 1113-43.

Todd, A. E., Orengo, C. A. and Thornton, J. M. (2002). "Plasticity of enzyme active sites." *Trends Biochem Sci* **27**(8): 419-26.

Ueda, K., Seki, T., Kudo, T., Yoshida, T. and Kataoka, M. (1999). "Two distinct mechanisms cause heterogeneity of 16s rRNA." *J Bacteriol* **181**(1): 78-82.

Umbarger, H. E. (1996). The branched-chain amino acids. In Escherichia coli *and* Salmonella: *Celullar and molecular biology*. F. C. Neidhart. Washington, ASM Press.

Urfer, R. and Kirschner, K. (1992). "The importance of surface loops for stabilizing an eightfold beta alpha barrel protein." *Protein Sci* **1**(1): 31-45.

Van Dyke, T. K., Ayers, B. L., Morgan, R. W. and Larossa, R. A. (1998). "Constricted flux through the branched-chain amino acid biosynthetic enzyme acetolactate synthase triggers elevated expression of genes regulated by RpoS and internal acidification." *J Bacteriol* **180**(4): 785-92.

Van Roey, P., Rao, V., Plummer, T. H., Jr. and Tarentino, A. L. (1994). "Crystal structure of endo-beta-n-acetylglucosaminidase F1, an alpha/beta-barrel enzyme adapted for a complex substrate." *Biochemistry* **33**(47): 13989-96.

Vilchez, S., Molina, L., Ramos, C. and Ramos, J. L. (2000). "Proline catabolism by *Pseudomonas putida*: Cloning, characterization, and expression of the *put* genes in the presence of root exudates." *J Bacteriol* **182**(1): 91-9.

Vinod, M. P., Bellur, P. and Becker, D. F. (2002). "Electrochemical and functional characterization of the proline dehydrogenase domain of the PutA flavoprotein from *Escherichia coli*." *Biochemistry* **41**(20): 6525-32.

Voet, D. and Voet, J. G. (1995). *Biochemistry*. New York, John Wiley & Sons.

Volff, J. N. and Altenbuchner, J. (1998). "Genetic instability of the *Streptomyces* chromosome." *Mol Microbiol* **27**(2): 239-46.

von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P. and Snel, B. (2003). "STRING: A database of predicted functional associations between proteins." *Nucleic Acids Res* **31**(1): 258-61.

Wagner, R. (2000). *Transcription Regulation in Prokaryotes*. Oxford, Oxford University Press.

Wasserman, H. H., Rodgers, G. C. and Keith, D. D. (1966). "The structure and synthesis of undecylprodigiosin. A prodigiosin analogue from *Streptomyces*." *Chem. Comm.* **22**: 825-6.

Wasserman, H. H., Sykes, R. J., Peverada, P., Shaw, C. K., Cushley, R. J. and Lipsky, C. R. (1973). "Biosynthesis of prodigiosin. Incorporation patterns of c-labeled alanine, proline, glycine and serine elucidated by fourier transform nuclear magnetic resonance." *J. Am. Chem. Soc.* **95**: 6874-5.

Watanabe, H., Mori, H., Itoh, T. and Gojobori, T. (1997). "Genome plasticity as a paradigm of eubacteria evolution." *J Mol Evol* **44 Suppl 1**: S57-64.

Watson, J. D. and Crick, F. H. (1953). "Molecular structure of nucleic acids: The structure for deoxyribose nucleic acid." *Nature*(737): 4356.

Watve, M. G., Tickoo, R., Jog, M. M. and Bhole, B. D. (2001). "How many antibiotics are produced by the genus *Streptomyces*?" *Arch Microbiol* **176**(5): 386-90.

White, O., Eisen, J. A., Heidelberg, J. F., Hickey, E. K., Peterson, J. D., Dodson, R. J., Haft, D. H., Gwinn, M. L., Nelson, W. C., Richardson, D. L., Moffat, K. S., Qin, H., Jiang, L., Pamphile, W., Crosby, M., Shen, M., Vamathevan, J. J., Lam, P., McDonald, L., Utterback, T., Zalewski, C., Makarova, K. S., Aravind, L., Daly, M. J., Fraser, C. M. and et al. (1999). "Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* r1." *Science* **286**(5444): 1571-7.

Wierenga, R. K. (2001). "The TIM-barrel fold: A versatile framework for efficient enzymes." *FEBS Lett* **492**(3): 193-8.

Wilmanns, M. and Eisenberg, D. (1993). "Three-dimensional profiles from residue-pair preferences: Identification of sequences with beta/alpha-barrel fold." *Proc Natl Acad Sci U S A* **90**(4): 1379-83.

Wilmanns, M., Hyde, C. C., Davies, D. R., Kirschner, K. and Jansonius, J. N. (1991). "Structural conservation in parallel beta/alpha-barrel enzymes that catalyze three sequential reactions in the pathway of tryptophan biosynthesis." *Biochemistry* **30**(38): 9161-9.

Wilmanns, M., Priestle, J. P., Niermann, T. and Jansonius, J. N. (1992). "Three-dimensional structure of the bifunctional enzyme phosphoribosylanthranilate isomerase: Indoleglycerolphosphate synthase from *Escherichia coli* refined at 2.0 A resolution." *J Mol Biol* **223**(2): 477-507.

Wise, E., Yew, W. S., Babbitt, P. C., Gerlt, J. A. and Rayment, I. (2002). "Homologous (beta/alpha)$_8$-barrel enzymes that catalyze unrelated reactions: Orotidine 5'-monophosphate decarboxylase and 3-keto-l-gulonate 6-phosphate decarboxylase." *Biochemistry* **41**(12): 3861-9.

Woese, C. R. (1987). "Bacterial evolution." *Microbiol Rev* **51**(2): 221-71.

Woese, C. R. (2000). "Interpreting the universal phylogenetic tree." *Proc Natl Acad Sci U S A* **97**(15): 8392-6.

Wolf, Y. I., Rogozin, I. B., Kondrashov, A. S. and Koonin, E. V. (2001). "Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context." *Genome Res* **11**(3): 356-72.

Wright, H., Barona-Gomez, F., Hodgson, D.A., Fulop, V. (2003) "Expression, purification and preliminary crystallographic analysis of a novel phosphoribosyl isomerase (PriA) from *Streptomyces coelicolor*." *Acta Crystallographica Section D.* **In press.**

Yanai, I., Mellor, J. C. and DeLisi, C. (2002). "Identifying functional links between genes using conserved chromosomal proximity." *Trends Genet* **18**(4): 176-9.

Yano, T., Oue, S. and Kagamiyama, H. (1998). "Directed evolution of an aspartate aminotransferase with new substrate specificities." *Proc Natl Acad Sci U S A* **95**(10): 5511-5.

Yanofsky, C. (2001). "Advancing our knowledge in biochemistry, genetics, and microbiology through studies on tryptophan metabolism." *Annu Rev Biochem* **70**: 1-37.

Yanofsky, C. (2003). "Using studies on tryptophan metabolism to answer basic biological questions." *J Biol Chem* **278**(13): 10859-78.

Zalkin, H. and Smith, J. L. (1998). "Enzymes utilizing glutamine as an amide donor." *Adv Enzymol Relat Areas Mol Biol* **72**: 87-144.

Zhang, Z., Wang, Y. and Ruan, J. (1998). "Reclassification of *Thermomonospora* and *Microtetraspora*." *Int J Syst Bacteriol* **48 Pt 2**: 411-22.

Zheng, Y., Roberts, R. J. and Kasif, S. (2002). "Genomic functional annotation using co-evolution profiles of gene clusters." *Genome Biol* **3**(11): RESEARCH0060.

Zitzewitz, J. A., Gualfetti, P. J., Perkons, I. A., Wasta, S. A. and Matthews, C. R. (1999). "Identifying the structural boundaries of independent folding domains in the alpha subunit of tryptophan synthase, a beta/alpha barrel protein." *Protein Sci* **8**(6): 1200-9.

Zuckerland, E. and Pauling, L. (1965). Evolutionary divergence and convergence in proteins. In *Evolving Genes and Proteins*. V. Bryson and H. J. Vogel. New York, Academic Press: 97-166.

Appendix 1

EMBO reports

scientific report

Occurrence of a putative ancient-like isomerase involved in histidine and tryptophan biosynthesis

Francisco Barona-Gómez[+] *& David A. Hodgson*

Department of Biological Sciences, University of Warwick, Coventry, UK

We report the occurrence of an isomerase with a putative $(\beta\alpha)_8$-barrel structure involved in both histidine and tryptophan biosynthesis in *Streptomyces coelicolor* A3(2) and *Mycobacterium tuberculosis* HR37Rv. Deletion of a *hisA* homologue (SCO2050) putatively encoding *N'*-[(5'-phosphoribosyl)-formimino]-5 amino-imidazole-4-carboxamide ribonucleotide isomerase from the chromosome of *S. coelicolor* A3(2) generated a double auxotrophic mutant for histidine and tryptophan. The bifunctional gene SCO2050 and its orthologue Rv1603 from *M. tuberculosis* complemented both *hisA* and *trpF* mutants of *Escherichia coli*. Expression of the *E. coli trpF* gene in the *S. coelicolor* mutant only complemented the tryptophan auxotrophy, and the *hisA* gene only complemented the histidine auxotrophy. The discovery of this enzyme, which has a broad substrate specificity, has implications for the evolution of metabolic pathways and may prove to be important for understanding the evolution of the $(\beta\alpha)_8$-barrels.

EMBO reports 4, 296–300 (2003)

doi:10.1038/sj.embor.embor771

INTRODUCTION

The evolution of enzymes and metabolic pathways are intrinsically related (Copley & Bork, 2000; Teichmann *et al.*, 2001). Many enzymes involved in central metabolic pathways have $(\beta\alpha)_8$-barrel scaffolds, to which are attached different catalytic and substrate-binding folds (Gerlt & Babbitt, 2001; Todd *et al.*, 2001). The enzymes *N'*-[(5'-phosphoribosyl)formimino]-5-aminoimidazole-4-carboxamide ribonucleotide (ProFAR) isomerase (HisA, EC 5.3.1.16), which is involved in histidine biosynthesis (Lang *et al.*, 2000), and *N'*-(5'-phosphoribosyl)-anthranilate (PRA) isomerase (TrpF, EC 5.3.1.24), which is involved in tryptophan biosynthesis (Priestle *et al.*, 1987), are members of this structural family. It is generally accepted that the $(\beta\alpha)_8$-barrel enzymes involved in central metabolic pathways arose by divergent evolution (Copley & Bork, 2000; Henn-Sax *et al.*, 2001; Gerlt & Babbitt, 2001; Nagano *et al.*, 2002). Functional evidence for a common ancestry includes the directed evolution of the activity of TrpF from the HisA protein using random mutagenesis and selection *in vivo* (Jürgens *et al.*, 2000). TrpF and HisA catalyse an Amadori rearrangement of their cognate aminoaldoses into the corresponding aminoketoses (Fig. 1; Henn-Sax *et al.*, 2002).

The evolution of enzymes by divergence has a bearing on how biosynthetic pathways may have evolved to their current architecture. Among the hypotheses put forward on how metabolic pathways evolved the 'patchwork' hypothesis (Jensen, 1976) offers the most generally accepted explanation. This hypothesis states that metabolic pathways arose by the recruitment of enzymes with similar activity, and by subsequent modification of their substrate-binding ability. The model implies the existence of ancestral enzymes with broad substrate specificity, catalysing related reactions in different pathways. The recruitment of enzyme function, as inferred from predictions of the evolutionary relatedness of enzymes from different metabolic pathways (Parsot, 1987; Copley & Bork, 2000; Teichmann *et al.*, 2001) and detection of promiscuous activities (O'Brien & Herschlag, 1999) is widely acknowledged.

In *Escherichia coli* and its relatives, the enzyme PRA isomerase (TrpF), and the next enzyme downstream in the tryptophan biosynthetic pathway, indole-3-glycerol-phosphate (IGP) synthase (TrpC; EC 4.1.1.48), are present on a single peptide chain encoded by *trpC* (Priestle *et al.*, 1987). During our investigations into the regulation of tryptophan biosynthesis in *Streptomyces coelicolor* A3(2) (Hu *et al.*, 1999), we failed to discover a *trpF* gene next to *trpC*. This paradox of a function without a gene was confirmed after completion of the sequencing of the complete genome of *S. coelicolor* (Bentley *et al.*, 2002). A similar situation was encountered in the genomes of *Mycobacterium tuberculosis* HR37Rv (Cole *et al.*, 1998) and *M. leprae* (Cole *et al.*, 2001). Here, we report the occurrence of an isomerase, with a putative $(\beta\alpha)_8$-barrel structure predicted from its sequence, with a dual function in both histidine and aromatic amino-acid biosynthesis in *S. coelicolor* and *M. tuberculosis*. This report expands on the suggested physiological link between tryptophan and histidine biosynthesis observed in other organisms (Nester & Montoya, 1976). The discovery of a putative $(\beta\alpha)_8$-barrel enzyme, with a predicted ancient-like broad substrate specificity, may be relevant for understanding the evolution of this important structural family.

RESULTS AND DISCUSSION

Functional genomics of *trpF in silico*

At present, the only actinomycete *trpF* sequence available in the database is that of *Corynebacterium glutamicum* (Matsui *et al.*, 1986). Sequence analysis using this sequence as a probe, either

Department of Biological Sciences, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK

[+]Corresponding author. Tel: +44 24 76 522 572; Fax: +44 24 76 523 701; E-mail: fbarona-gomez@bio.warwick.ac.uk

Received 23 October 2002; revised 13 January 2003; accepted 15 January 2003
Published online 21 February 2003

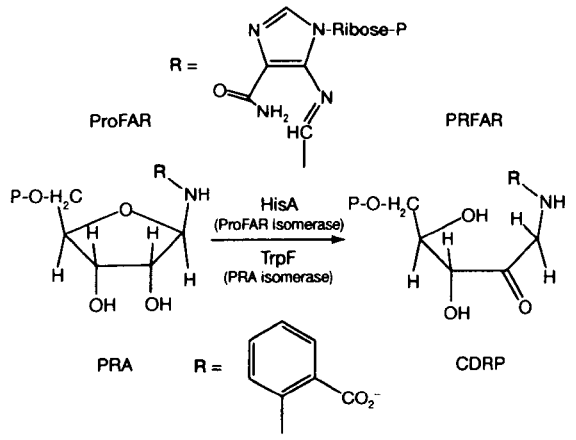


Fig. 1 | HisA and TrpF catalyse similar reactions. HisA and TrpF catalyse analogous Amadori rearrangements of *N*′-[(5′-phosphoribosyl)formimino]-5-aminoimidazole-4-carboxamide ribonucleotide (ProFAR) and *N*′-(5′-phosphoribosyl)anthranilate (PRA) into *N*′-[(5′-phosphoribulosyl) formimino]-5-aminoimidazole-4-carboxamide (PRFAR) and 1-[(2-carboxyphenyl)amino]-1-deoxyribulose 5-phosphate (CDRP), respectively. Red 'R's indicate the different side-chains in the two substrates.

using hidden Markov models (Karplus *et al.*, 1998) or by carrying out BLAST searches (Altschul *et al.*, 1997), failed to identify a homologue in *S. coelicolor* or the mycobacteria. It has been proposed in other organisms that TrpC and TrpF proteins share a common ancestry (Wilmanns *et al.*, 1991; Gerlt & Babbitt, 2001). We tested whether the sequence of TrpC (open-reading frame (ORF) number SCO2039) of *S. coelicolor* could reveal a TrpF-like sequence from an actinomycete by performing PSI-BLAST searches (Altschul *et al.*, 1997). Again, no TrpF-like sequence was identified. These observations indicate that the *trpF* genes of *S. coelicolor* and *Mycobacterium* species are of a different family to those found in other bacteria.

On the basis of the report by Jürgens *et al.* (2000) on the directed evolution of TrpF from HisA, we proposed that the HisA protein might have TrpF activity in *S. coelicolor*, which would account for the lack of *trpF* in this species. A *hisA* homologue was identified in *S. coelicolor* as part of a histidine biosynthetic cluster (Limauro *et al.*, 1990). The completion of the genome sequencing project revealed that this cluster is localized on the chromosome upstream of a cluster of *trp* genes (Fig. 2A; Bentley *et al.*, 2002). A similar bifunctionality of *hisA* is implied in *M. tuberculosis* by the synteny of the *his* and *trp* clusters in this species (Fig. 2A).

Tryptophan and histidine *S. coelicolor* auxotrophy

We disrupted the *hisA* homologue of *S. coelicolor* M145 (ORF number SCO2050, according to the annotation of Bentley *et al.* (2002)); this was done by deleting two of the three catalytically important amino-acid residues, Asp 130 and Thr 166, which were identified by aligning the sequences of SCO2050 and *hisA* from *Thermotoga maritima* (Fig. 2B; Jürgens *et al.*, 2000; Henn-Sax *et al.*, 2002). This in-frame deletion was constructed in a way that avoided possible polar effects. As hypothesized, the resulting mutant (WH101) could not grow on minimal medium unless supplemented with both histidine and tryptophan (Table 1). We propose that SCO2050 should be renamed *priA* (*phosphoribosyl isomerase A*) to reflect the common effects of TrpF and HisA on phosphoribosylated substrates.

Gene complementation studies

The *priA* gene was inserted into the plasmid pIJ702 to make pIJ702–PriASc, which was used to transform WH101 mutants. The resulting transformants were prototrophic, showing that the tryptophan and histidine auxotrophy of WH101 mutants is exclusively due to the loss of functional *priA* (Table 1). We tested whether the *trpF* and *hisA* genes of *E. coli* could complement the WH101 tryptophan and histidine auxotrophies. The *E. coli* genes were cloned into the *Streptomyces* expression vector pIJ4123 under the control of the thiostrepton-inducible promoter. Thiostrepton-dependent expression of *trpF* (using pIJ4123–TrpFEc) in WH101 mutants restored tryptophan independence only, whereas expression of *hisA* (using pIJ4123–HisAEc) restored histidine independence only (Table 1). We also investigated the ability of *priA* to complement independent *hisA* and *trpF* deletions in *E. coli*. For this purpose, *priA* was cloned into the expression vector pGEX-4T-1 to form pGEX–PriASc, which was used to transform *E. coli* auxotrophs with mutations in *trpF* (W3110 *trpC* (Fdel); see Darimont *et al.*, 1998), and *hisA* (Hfr G6; see Matney *et al.*, 1964). Complementation of both mutations was achieved by expression of *priA* (Table 2). The *hisA* and *trpF* genes of *E. coli* cloned into pGEX-4T-1, to create pGEX–HisAEc and pGEX–TrpFEc, respectively, were used as controls (Table 2).

The intergenic complementation of *trpF* and *hisA* by *priA*, and the partial complementation of *priA* by either *trpF* or *hisA*, confirms that the product of *priA* is involved in the biosynthesis of both histidine and tryptophan in *S. coelicolor*. This discovery places into a physiological context the remarkable observation that TrpF and HisA activities can co-exist in a single protein (Jürgens *et al.*, 2000) despite the low identity (10%) between the sequences of these enzymes. Preliminary characterization of the *in vitro* activity of PriA has shown that this HisA-like enzyme has PRA isomerase activity (data not shown).

Mycobacterium tuberculosis has a *priA* gene

To test the possibility that mycobacteria were similar to streptomycetes in having a single gene that encodes both TrpF and HisA activity, we cloned the putative *priA* orthologue (ORF number Rv1603, according to Cole *et al.* (1998)) of *M. tuberculosis* H37Rv into the expression vector pGEX-4T-1, to produce pGEX–PriAMt. This plasmid was shown to complement *hisA* and *trpF* mutations in *E. coli* (Table 2). As there is no *trpF* gene in *Mycobacterium* species, these results suggest that the protein product of Rv1603 in *M. tuberculosis* has the same metabolic role as that of PriA in *S. coelicolor*. This indicates that the presence of PriA in *S. coelicolor* and mycobacteria is due to the retention of ancient characteristics, rather than the modification of HisA to provide TrpF activity and a subsequent loss of TrpF. The failure to detect any *trpF*-like sequence in *S. coelicolor* supports the former interpretation.

Why did the streptomycetes and mycobacteria not develop independent *trpF* genes? In general, the streptomycetes do not regulate amino-acid biosynthesis by feedback repression of gene expression (Hodgson, 2000). Therefore, as enzyme expression is not co-ordinately regulated on a pathway-specific basis, enzymes with functions in multiple pathways are possible. The genome sequence of *C. glutamicum* shows that this actinobacterium has a *trp* operon

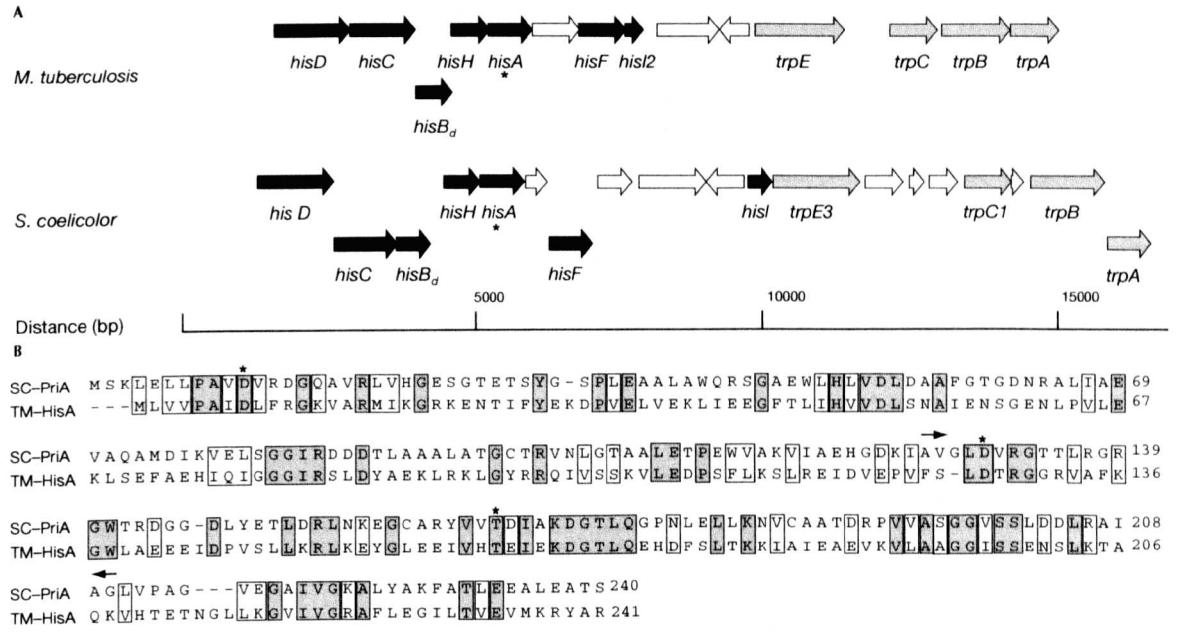


Fig. 2 | Synteny of the *his/trp* region and sequence similarity of PriA and HisA. (**A**) Organization of the *his/trp* cluster of *Streptomyces coelicolor* (Bentley *et al.*, 2002) and *Mycobacterium tuberculosis* (Cole *et al.*, 1998). The *hisA* homologues of *S. coelicolor* (SCO2050) and *M. tuberculosis* (Rv1603) are marked with asterisks. Genes involved in histidine (*his*; black) and tryptophan (*trp*; grey) biosynthesis are shown. (**B**) Sequence alignment of the HisA homologues of *S. coelicolor* (SC–PriA) and *Thermotoga maritima* (TM–HisA), which show 25% identity. Data from *T. maritima* (Jürgens *et al.*, 2000; Henn-Sax *et al.*, 2002) were used to identify the catalytically important residues in SCO2050 (Asp 11, Asp 130 and Thr 166 (asterisks)). The region of SCO2050 replaced in WH101 (this study) lies between Ala 126 and Gly 210 (indicated by arrows).

that contains a fused *trpCF* gene. This strain does regulate amino acid biosynthesis by feedback regulation (Sano & Matsui, 1987), and a simple explanation is that its ancestor acquired a complete, regulated *trp* operon after it separated from the streptomycete and mycobacteria evolutionary lines. Crawford (1989) has previously proposed this possibility. It would be interesting to test the putative *hisA* product of *C. glutamicum* for PRA isomerase activity.

An evolutionary interpretation of the nature of PriA

An obvious mechanistic explanation for the bifunctionality of PriA would be that it shows a broad specificity for the substrates PRA and ProFAR (Fig. 1). Enzymes with broad substrate specificity are not rare, and examples of these have accumulated since the postulation of the patchwork hypothesis by Jensen (1976). Nevertheless, the existence of broad substrate specificity, which stands at the core of the patchwork hypothesis, is usually inferred from the promiscuous activities of enzymes *in vitro* (for a review, see O'Brien & Herschlag, 1999) and/or overlapping specificities *in vivo*. Examples of the latter include members of the superfamily of aminotransferases (Jensen & Gu, 1996). This family includes another example of an enzyme that functions in both histidine and aromatic amino acid biosynthesis, which was first identified by Nester & Montoya (1976).

A number of enzymes involved in the biosynthesis of branched-chain amino acids show specificity for multiple substrates, such that isoleucine and valine are produced in the same metabolic pathway (Umbarger, 1996). However, this situation is different to that of PriA, because the substrates of PriA show marked difference in size and shape (Fig. 1). Therefore, if the broad substrate specificity of a given enzyme is in fact an ancient feature, as suggested by Jensen (1976), this implies that this feature has been retained in PriA throughout the course of its evolution. Thus, we conclude that PriA is a 'molecular fossil'.

Speculation

We believe that *priA* did not evolve as a consequence of loss of *trpF* and subsequent broadening of the specificity of the HisA protein. On the basis of the broad substrate specificity of PriA, one possibility is that *trpF* and *hisA* could have evolved from *priA* after gene duplication and specialization, as suggested by the patchwork hypothesis (Jensen, 1976). This would be an example of

Table 1 | Growth requirements of *Streptomyces coelicolor* WH101 and its transformants

Strain	Genotype	Growth requirements
WH101	*priA*::scar	Histidine and tryptophan
WH101(pIJ702)	*priA*::scar	Histidine and tryptophan
WH101(pIJ702–PriASc)	*priA*::scar (*priA*$^+$)	Prototrophic
WH101(pIJ4123–TrpFEc)	*priA*::scar (*trpF*$^+$)	Histidine
WH101(pIJ4123–HisAEc)	*priA*::scar (*hisA*$^+$)	Tryptophan

Table 2 | Complementation of *Escherichia coli* strains W3110 *trpC*(*F*del) and Hfr G6

Plasmid	Strain	
	W3110 *trpC*(*F*del)	Hfr G6
pGEX-4T-1	–	–
pGEX–PriASc	+	+
pGEX–PriAMt	+	+
pGEX–HisAEc	–	+
pGEX–TrpFEc	+	–

Complementation detected (+) or not detected (–) by expression of the insert from the *lacI*-dependent promoter induced with 10 μM isopropylthiogalactoside.

divergent evolution, and implies that the ancestor of *priA* encoded one of the older members of the $(\beta\alpha)_8$-barrel protein family (Fani *et al.*, 1994; Copley & Bork, 2000; Lang *et al.*, 2000), although it is also possible that the *priA* ancestor evolved relatively recently (Nagano *et al.*, 2002). An alternative possibility is that a TrpF function could have evolved from another enzyme, such as TrpC (Wilmanns *et al.*, 1991; Gerlt & Babbitt, 2001), allowing PriA to lose its TrpF activity, an example of convergent evolution. Therefore, convergent or divergent evolution may have accounted for the evolution of the extant $(\beta\alpha)_8$-barrel proteins.

METHODS

Computational sequence analysis and searches. The sequences used for the searches were TrpCF (accession number E24723), the genome of *C. glutamicum* (accession number NC_003450), TrpC (accession number SCO2039) and the genome of *S. coelicolor* (accession number NC_003888). The program ClustalW was used for sequence alignment.

Growth requirements of strains. Strains Hfr G6 and W3110 *trpC* (*F*del) were provided by the *E. coli* Genetic Stock Center. Minimal A medium (Miller, 1972) was used to test Hfr G6 and its derivatives, and modified Vogel–Bonner medium (Darimont *et al.*, 1998) was used to test W3110 *trpC* (*F*del) and its derivatives. Tryptophan and histidine were added to a final concentration of 100 μg ml^{-1}. Plasmids were selected for using ampicillin at 100 μg ml^{-1}; induction of *lacI*-dependent promoters was carried out using 10 μM isopropylthiogalactoside. The minimal medium for streptomycetes (Kieser *et al.*, 2000) was supplemented with tryptophan (37.5 μg ml^{-1}) and histidine (50 μg ml^{-1}) as appropriate. Thiostrepton (Sigma) was used at 50 μg ml^{-1} to select for plasmids and for induction of thiostrepton-inducible promoters. Apramycin (Sigma) selection was carried out using the antibiotic at 50 μg ml^{-1}.

Deletion of SCO2050 from the chromosome of *S. coelicolor*. The auxotrophic strain WH101 was constructed using REDIRECT® (Gust *et al.*, 2003). The protocol, plasmids and strains were provided by PBL Biomedical Laboratories. The oligonucleotides used for this were: 5′-TGGGTCGCCAAGGTCATCGCCGAGCACGGCGCAAGATCATTCCGGGGATCCGTCGACC-3′ and 5′-CTTCCCGACGATGGCCCCCTCGACACCGGCCGGACGAGTGTAGGCTGGAGCTGCTTC-3′ (the bases that are identical in the SCO2050 sequence are underlined). The disruption cassette was made by PCR using Expand high-fidelity DNA polymerase (Roche). SCO2050 was mutagenized in cosmid SC4G6 (Redenbach *et al.*, 1996) by homologous recombination (double crossover) replacing 255 bp from the 5′ end of the gene. The disruption cassette was removed by the FLP-recombinase system, leaving behind a 'scar' of 81 nucleotides with no stop codons. The newly mutagenized cosmid (carrying only the original marker) was re-engineered, inserting the selectable marker *aac(3)IV* and an RP4 *oriT*. The resulting construct was introduced into *S. coelicolor* M145 by RP4-based conjugation (Kieser *et al.*, 2000) and selected for using apramycin. Authentic double crossovers were obtained after two rounds of growth on fresh plates containing soya-flour mannitol medium (Kieser *et al.*, 2000) without selection. The replacement of SCO2050 was identified in colonies that were apramycin sensitive and auxotrophic for tryptophan and histidine. The presence of the wild type or of the mutated form of SCO2050 was detected by PCR using the primers 5′-GGGCGAAACCGAAGGACTC-3′ and 5′-TCGTGGCCGCCGTGGAGAACG-3′, and sequencing was used to confirm that the desired replacement event had taken place.

Cloning of the *hisA* and *trpF* genes of *E. coli*, and *priA* of *S. coelicolor* and *M. tuberculosis*. All DNA fragments were produced by PCR amplification, using primers with restriction sites engineered at their 5′ ends (see supplementary information online). PCR was performed using *Pwo* DNA polymerase (Roche). The PCR products were digested with the appropriate restriction enzymes and were ligated using T4 DNA ligase (Gibco) into the vectors pMTL22 (Chambers *et al.*, 1988), pET3a, pET22a (Novagen) or pGEX-4T-1 (Amersham Pharmacia). The resulting plasmids were used to transform *E. coli* MC1061 using the calcium method (Sambrook *et al.*, 1989). Fragments from restriction digests of the pMTL22 and pET constructs were ligated into plasmids pIJ702 and pIJ4123, and the resulting constructs were used to transform *Streptomyces lividans* TK24 protoplasts (Kieser *et al.*, 2000). The desired constructs were isolated from TK24 and used to transform *S. coelicolor* WH101 protoplasts. The *priA* gene (SCO2050) from *S. coelicolor* was amplified by PCR from cosmid 4G6. The *hisA* and *trpF* genes of *E. coli* were amplified from chromosomal DNA. The *trpF* gene was also cut out of pMS401 (a gift from M. Samaddar and J. Blackburn) using *Nco*I and *Bam*HI and sub-cloned into pET22a, from where *trpF* was subsequently removed using *Nde*I and *Bam*HI and subcloned into pIJ4123. The *priA* orthologue (Rv1603) of *M. tuberculosis* H37Rv was amplified from chromosomal DNA, provided by D. Roper. Sequencing was used to confirm all constructs.

Supplementary information is available at *EMBO reports* online (http://www.emboreports.org).

ACKNOWLEDGEMENTS

We thank B. Gust of the John Innes Centre for help in constructing the mutant strain. We are indebted to C. Orengo, D. Lee, A. Grant and S. Rison of University College, London, for help with the bioinformatic analysis and for useful discussions. We would also like to thank J. Blackburn and W. Patrick of Cambridge, UK, for providing us with pMS401 and for useful comments. We are grateful to D. Roper from our department for supplying us with mycobacterial DNA. F.B.-G is a recipient of a Conacyt scholarship, Mexico (number 111558).

REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bentley, S.D. *et al.* (2002) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature*, **417**, 141–147.

Chambers, S.P., Prior, S.E., Barstow, D.A. & Minton, N.P. (1988) The pMTL *nic*⁻ cloning vectors. I. Improved pUC polylinker regions to

facilitate the use of sonicated DNA for nucleotide sequencing. *Gene*, **68**, 139–149.

Cole, S.T. *et al.* (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, **393**, 537–544.

Cole, S.T. *et al.* (2001) Massive gene decay in the leprosy bacillus. *Nature*, **409**, 1007–1011.

Copley, R.R. & Bork, P. (2000) Homology among $(\beta\alpha)_8$ barrels: implications for the evolution of metabolic pathways. *J. Mol. Biol.*, **303**, 627–640.

Crawford, I.P. (1989) Evolution of a biosynthetic pathway: the tryptophan paradigm. *Annu. Rev. Microbiol.* **43**, 567–600.

Darimont, B., Stehlin, C., Szadkowski, H. & Kirschner, K. (1998) Mutational analysis of the active site of indoleglycerol phosphate synthase from *Escherichia coli*. *Protein Sci.*, **7**, 1221–1232.

Fani, R., Lio, P., Chiarelli, I. & Bazzicalupo, M. (1994) The evolution of the histidine biosynthetic genes in prokaryotes: a common ancestor for the *hisA* and *hisF* genes. *J. Mol. Evol.*, **38**, 489–495.

Gerlt, J.A. & Babbitt, P.C. (2001) Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annu. Rev. Biochem.*, **70**, 209–246.

Gust, B., Challis, G.L., Fowler, K., Kieser, T. & Chater, K.F. (2003) PCR-targeted *Streptomyces* gene replacement identifies a protein domain needed for biosynthesis of the sesquiterpene soil odour geosmin. *Proc. Natl Acad. Sci. USA*, 31 January 2003, 10.1073/pnas.0337542100.

Henn-Sax, M., Hocker, B., Wilmanns, M. & Sterner, R. (2001) Divergent evolution of $(\beta\alpha)_8$-barrel enzymes. *Biol. Chem.*, **382**, 1315–1320.

Henn-Sax, M., Thoma, R., Schmidt, S., Hennig, M., Kirschner, K. & Sterner, R. (2002) Two $(\beta\alpha)_8$-barrel enzymes of histidine and tryptophan biosynthesis have similar reaction mechanisms and common strategies for protecting their labile substrates. *Biochemistry*, **41**, 12032–12042.

Hodgson, D.A. (2000) Primary metabolism and its control in streptomycetes: a most unusual group of bacteria. *Adv. Microb. Physiol.*, **42**, 47–238.

Hu, D.S.-J., Hood, D.W., Heidstra, R. & Hodgson, D.A. (1999) The expression of the *trpD*, *trpC* and *trpBA* genes of *Streptomyces coelicolor* A3(2) is regulated by growth rate and growth phase but not by feedback repression. *Mol. Microbiol.*, **32**, 869–880.

Jensen, R.A. (1976) Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.*, **30**, 409–425.

Jensen, R.A. & Gu, W. (1996) Evolutionary recruitment of biochemically specialized subdivisions of family I within the protein superfamily of aminotransferases. *J. Bacteriol.*, **178**, 2161–2171.

Jürgens, C., Strom, A., Wegener, D., Hettwer, S. & Wilmanns, M. (2000) Directed evolution of a $(\beta\alpha)_8$-barrel enzyme to catalyze related reactions in two different metabolic pathways. *Proc. Natl Acad. Sci. USA*, **97**, 9925–9930.

Karplus, K., Barrett, C. & Hughey, R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.

Kieser, T., Bibb, M.J., Buttner, M.J., Chater, K.F. & Hopwood, D.A. (2000) *Practical* Streptomyces *Genetics*. The John Innes Foundation, Norwich, UK.

Lang, D., Thoma, R., Henn-Sax, M., Sterner, R. & Wilmanns, M. (2000) Structural evidence for evolution of the $\beta\alpha$-barrel scaffold by gene duplication and fusion. *Science*, **289**, 1546–1550.

Limauro, D., Avitabile, A., Capellano, C., Puglia, M.A. & Bruni, C.B. (1990) Cloning and characterization of the histidine biosynthetic gene cluster of *Streptomyces coelicolor* A3(2). *Gene*, **90**, 31–41.

Matney, T.S., Goldschmidt, E.P., Erwin, N.S. & Scroggs, R.A. (1964) A preliminary map of genomic sites for F-attachment in *Escherichia coli* K12. *Biochem. Biophys. Res. Commun.*, **17**, 278–281.

Matsui, K., Sano, K. & Ohtsubo, E. (1986) Complete nucleotide and deduced amino acid sequences of the *Brevibacterium lactofermentum* tryptophan operon. *Nucleic Acids Res.*, **14**, 10113–10114.

Miller, J.H. (1972) in *Experiments in Molecular Genetics*, 431–432, Cold Spring Harbor Laboratory Press, New York, USA.

Nagano, N., Orengo, C.A. & Thornton, J.M. (2002) One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J. Mol. Biol.*, **321**, 741–765.

Nester, E.W. & Montoya, A.L. (1976) An enzyme common to histidine and aromatic amino acid biosynthesis in *Bacillus subtilis*. *J. Bacteriol.*, **126**, 699–705.

O'Brien, P.J. & Herschlag, D. (1999) Catalytic promiscuity and the evolution of new enzymatic activities. *Chem. Biol.*, **6**, 91–105.

Parsot, C. (1987) A common origin for enzymes involved in the terminal step of the threonine and tryptophan biosynthetic pathways. *Proc. Natl Acad. Sci. USA*, **84**, 5207–5210.

Priestle, J.P., Grütter, M.G., White, J.L., Vincent, M.G., Kania, M., Wilson, E., Jardetzky, T.S., Kirschner, K. & Jansonius, J.N. (1987) Three-dimensional structure of the bifunctional enzyme *N*-(5′-phosphoribosyl)anthranilate isomerase-indole-3-glycerol-phosphate synthase from *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **84**, 5690–5694.

Redenbach, M., Kieser, H.M., Denapaite, D., Eichner, A., Cullum, J., Kinashi, H. & Hopwood, D.A. (1996) A set of ordered cosmids and a detailed genetic and physical map for the 8-Mb *Streptomyces coelicolor* A3(2) chromosome. *Mol. Microbiol.*, **21**, 77–96.

Sambrook, J., Fritsch, E. & Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, NY, USA.

Sano, K. & Matsui, K. (1987) Structure and function of the *trp* operon control regions of *Brevibacterium lactofermentum*, a glutamic-acid-producing bacterium. *Gene*, **53**, 191–200.

Teichmann, S.A., Rison, S.C.G., Thornton, J.M., Riley, M., Gough, J. & Chothia, C. (2001) The evolution and structural anatomy of the small molecule metabolic pathways in *Escherichia coli*. *J. Mol. Biol.*, **311**, 693–708.

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

Umbarger, H.E. (1996) In Escherichia coli *and* Salmonella: *Cellular and Molecular Biology* (ed. Neidhardt, F.C.) 442–457. ASM, Washington, DC, USA.

Wilmanns, M., Hyde, C.C., Davies, D.R., Kirschner, K. & Jansonius, J.N. (1991) Structural conservation in parallel β/α-barrel enzymes that catalyze three sequential reactions in the pathway of tryptophan biosynthesis. *Biochemistry*, **30**, 9161–9169.

Appendix 2

PCR targeting system in *Streptomyces coelicolor* A3(2)

Bertolt Gust, Tobias Kieser and Keith Chater, John Innes Centre, Norwich Research Park, Colney, Norwich NR47UH, UK, Tel: +44 (0)1603 452751 Fax: +44 (0)1603 456844

Introduction

Many bacteria are not readily transformable with linear DNA because of the presence of the intracellular *recBCD* exonuclease that degrades linear DNA. However, the λ RED (*gam, bet, exo*) functions promote a greatly enhanced rate of recombination when using linear DNA. By exploiting this, Datsenko and Wanner (2000) made 40 different disruptions on the *E. coli* chromosome by replacing the wild-type sequences with a selectable marker generated by PCR using primers with 36 nt homology extensions.

The strategy for PCR-targeting for mutagenesis of *Streptomyces coelicolor* is to replace a chromosomal sequence within a *S. coelicolor* cosmid (Redenbach *et al.*, 1996) by a selectable marker that has been generated by PCR using primers with 39 nt homology extensions. The inclusion of *oriT* (RK2) in the disruption cassette allows conjugation to be used to introduce the PCR targeted cosmid DNA into *S. coelicolor.* Conjugation is much more efficient than transformation of protoplasts and it is readily applicable to many actinomycetes (Matsushima *et al.*, 1994). The potent methyl-specific restriction system of *S. coelicolor* is circumvented by passaging DNA through a methylation-deficient *E. coli* host such as ET12567 (MacNeil *et al.*, 1992). Vectors containing *oriT* (RK2; Pansegrau *et al.*, 1994) are mobilisable *in trans* in *E. coli* by the self-transmissible pUB307 (Bennett *et al.*, 1977, Flett *et al.*, 1997) or the non-transmissible pUZ8002, which lacks a *cis*-acting function for its own transfer (Kieser *et al.*, 2000).

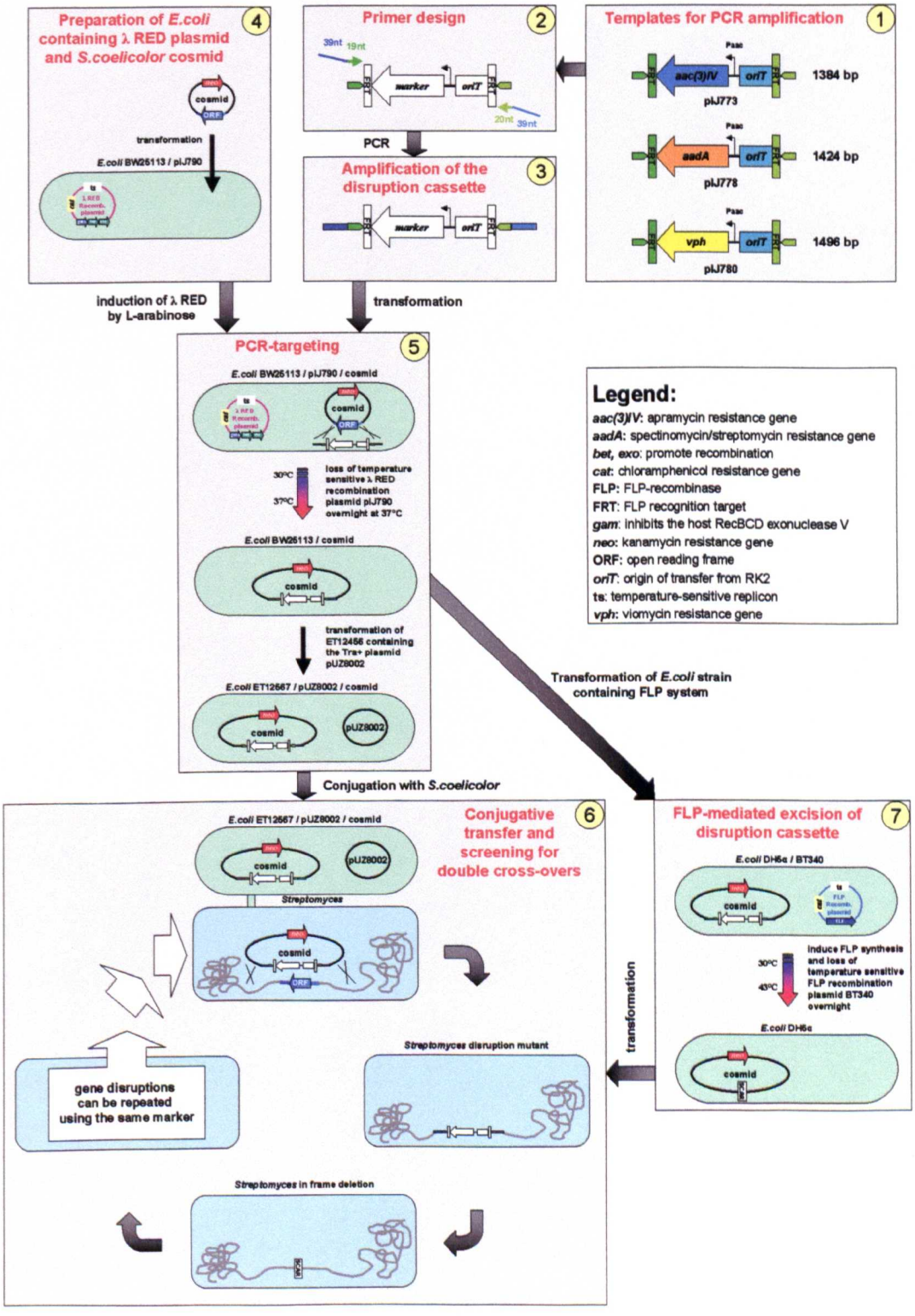
To adapt the procedure of λ RED mediated recombination for *Streptomyces*, cassettes for gene disruptions were constructed that can be selected both in *E. coli* and in *Streptomyces* (Table 1). After a single disruption with an *oriT*-containing cassette, further disruptions can be performed on the same cosmid using *oriT*-free cassettes containing alternative selective markers. The λ RED recombination plasmid pKD20

(*E. coli* Genetic Stock Center CGSC Strain # 7637) was modified by replacing the ampicillin resistance gene *bla* with the chloramphenicol resistance gene *cat*, generating pIJ790, to permit selection in the presence of Supercos1-derived cosmids (ampicillin and kanamycin resistance).

Name of plasmid	Resistance-marker	Resistance	Concentration for *E. coli*	*oriT*	Size of cassette
pIJ773 Fig. 5	*aac(3)IV*	apramycin	50 µg/ml LB	+	1384 bp
pIJ778 Fig. 6	*aadA*	spectinomycin streptomycin	50 µg/ml LB 50 µg/m LB	+	1424 bp
pIJ779.	*aadA*	spectinomycin- streptomycin	50 µg/ml LB 50 µg/ml LB	-	1118 bp
pIJ780 Fig.7	*vph*	viomycin	30 µg/ml DNA	+	1496 bp
pIJ781	*vph*	viomycin	30 µg/ml DNA	-	1540 bp

Table 1: Disruption cassettes containing different resistance markers with and without *oriT*: All disruption cassettes were cloned into the *Eco*RV site of pBluescript SK II (+) allowing the isolation of a *Eco*RI/*Hind*III fragment for use as template for the PCR reaction. The size of the cassettes includes the 19 bp and 20 bp primer site (see section 2: "primer design") which are identical in all disruption cassettes. The resistance genes with or without *oriT* are flanked by FRT sites (FLP recognition targets) which allows FLP-mediated excision of the cassette (see section 7: "FLP-mediated excision of the disruption cassette").

Fig. 1: Flowchart of gene disruption by PCR-targeting



Protocol (see Flowchart Fig. 1)

1 Purification of the PCR template (resistance (-*oriT*) cassette)

Using whole plasmids as templates for the PCR can result in a high proportion of antibiotic-resistant transformants without gene disruption. This is caused by traces of CCC DNA that compete with the linear PCR fragment and result in the occurrence of false positive transformants. Using gel-purified disruption cassettes as templates prevents the occurrence of false positives.

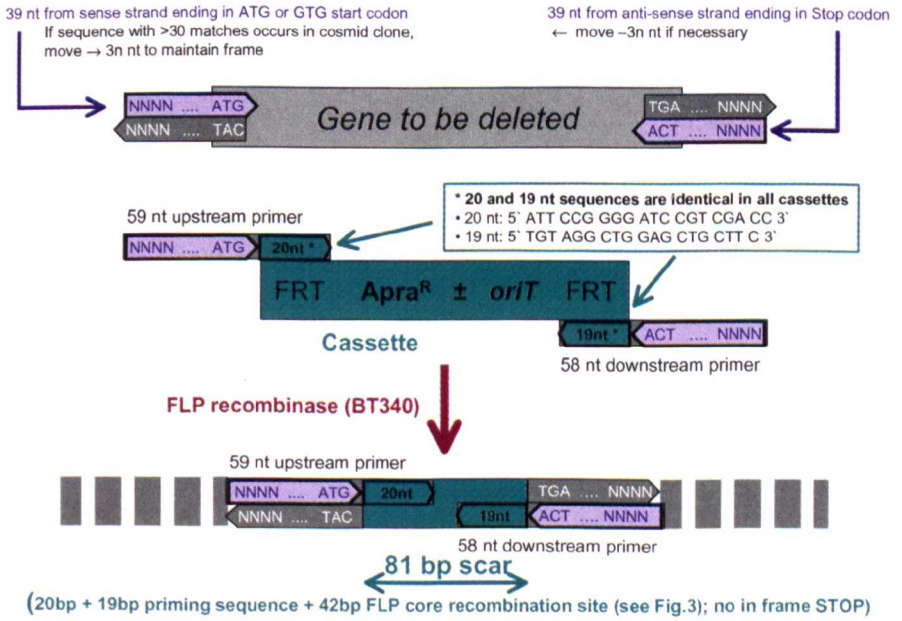
1. Digest ~ 10 µg plasmid DNA (see Table 1) with 50 U *Eco*RI (Roche) and 50 U *Hin*dIII (Roche) in 1 X buffer B (Roche) in a 100 µl reaction.
   - **A 2938 bp vector fragment and a fragment 14 bp larger than the size of the cassette given in Table 1 should be generated.**
2. Run the digest on a 20 x 20 x 0.25 cm (100 ml) 1% TAE (1x) agarose gel at 5V/cm for 2 - 3 h in 1x TAE buffer.
   - **Longer runs exhaust the buffer capacity and destroy the gel unless the buffer is recycled.**
3. Cut out the cassette band from the gel and purify using the Qiagen gel extraction kit. The purified fragment is stored in 10 mM Tris.HCl (pH 8) at a concentration of 100 ng / µl at –20°C.
4. Absence of plasmid DNA is tested by using 1µl (100 ng) of purified cassette DNA to transform highly competent *E. coli* DH5α cells (10^8/µg). Plate on LB agar containing 100 µg/ml carbenicillin. If any transformants appear, repeat steps 2-4.

2 Design of long PCR primers

For each gene disruption, two long PCR primers (58 nt and 59 nt) are required. Each has at the 5′end 39 nt matching the *S. coelicolor* sequence adjacent to the gene to be inactivated, and a 3′sequence (19 nt or 20 nt) matching the right or left end of the disruption cassette (all cassettes have the same "right" and "left" ends). The precise positioning of the 39 nt sequence as indicated in Fig. 2 is important for creating in-frame deletions by FLP recombinase-induced excision of the resistance marker (see section 7).

- **The 5′- 39 nt sequence of the forward primer (upstream primer; Fig. 2) must be from the coding strand of the gene of interest and its 3' end must be in the correct reading frame with respect to the replaced gene. The 5′- 39 nt sequence of the reverse primer (downstream primer; Fig. 2) must be from the *complementary* strand.**
- **To prevent unwanted recombination, a BlastN search is performed comparing each 39 nt sequence with the "real cosmid" (sequences at the Sanger Centre Homepage in the folder ftp.sanger.ac.uk/pub/S_coelicolor/cosmid_inserts and on the CD in the folder /S_coelicolor/cosmid inserts). The perfect match should be found but no other matches >30 bp. If necessary, the 39 nt sequence is shifted in 3 nt steps until the above criteria are met.**

Fig.2: Designing PCR primers for making an *in-frame* deletion
(the example illustrates a complete deletion)



3 PCR amplification of the extended resistance cassette

All PCR amplifications are performed using the Expand high fidelity PCR system according to the manufacturer's instructions (Roche). Reaction conditions:

• Primers (100 pmoles/µl)	0.5 µl each	50 pmoles each
• Template DNA (100 ng/µl)	0.5 µl	50 ng ≈ 0.06 pmoles
• Buffer (10x)	5 µl	1 x
• dNTPs (10 mM)	1 µl each	50 µM each
• DMSO (100 %)	2.5 µl	5%
• DNA polymerase (2.5 U/µl)	1 µl	2.5 Units
• Water	36 µl	
• Total volume	50 µl	

Cycle conditions:

1. Denaturation:	94°C, 2 min	
2. Denaturation:	94°C, 45 sec	10 cycles
3. Primer annealing:	**50°C**, 45 sec	
4. Extension:	72°C, 90 sec	
5. Denaturation:	94°C, 45 sec	15 cycles
6. Primer annealing:	**55°C**, 45 sec	
7. Extension:	72°C, 90 sec	
8. Final extension:	72°C, 5 min	

5 µl of the PCR product is used for analysis by gel electrophoresis. The expected sizes are 78 bp larger than the sizes of the disruption cassettes listed in Table **1** (because of the 2 x 39 bp 5′-primer extensions). The remaining 45 µl of the PCR product is purified using the Qiagen PCR purification kit according to the manufacturer's instructions. The PCR product is finally eluted from the columns with 12 µl of water (~200 ng/µl).

4 Introduction of *S. coelicolor* cosmid clone into *E. coli* BW25113/pIJ790 (λ RED recombination plasmid) by electroporation

pIJ790 contains the resistance marker *cat* (chloramphenicol resistance) and a temperature sensitive origin of replication (requires 30°C for replication).

1. Grow *E. coli* BW25113/pIJ790 overnight at 30°C in 10 ml LB (Luria-Bertani medium; Sambrook *et al.*, 1998) containing chloramphenicol (25 µg/ml).
2. Inoculate 100 µl *E. coli* BW25113/pIJ790 from overnight culture in 10 ml SOB (Hanahan, 1983) containing 20 mM $MgSO_4$ (add 200 µl of 1M stock to 10 ml SOB) and chloramphenicol (25 µg/ml).
3. Grow for 3-4 h at 30°C shaking at 200 rpm to an OD_{600} of ~ 0.6.
4. Recover the cells by centrifugation at 4000 rpm for 5 min at 4°C in a Sorvall GS3 rotor (or equivalent).
5. Decant medium and resuspend the pellet by gentle mixing in 10 ml ice-cold 10 % glycerol.
6. Centrifuge as above and resuspend pellet in 5 ml ice-cold 10 % glycerol, centrifuge and decant. Resuspend the cell pellet in the remaining ~ 100 µl 10 % glycerol.
7. Mix 50 µl cell suspension with ~ 100 ng (1-2 µl) of cosmid DNA. Carry out electroporation in a 0.2 cm ice-cold electroporation cuvette using a BioRad GenePulser II set to: 200 Ω, 25 µF and 2,5 kV. The expected time constant is 4.5 – 4.9 ms.
8. Immediately add 1 ml ice cold LB to shocked cells and incubate shaking for 1h at 30°C.
9. Spread onto LB agar containing carbenicillin (100 µg/ml), kanamycin (50 µg/ml) and chloramphenicol (25 µg/ml).
10. Incubate overnight at 30°C.
11. Transfer one isolated colony into 5 ml LB containing antibiotics as in (9) above.
12. Incubate overnight at 30°C. This culture will be used as a pre-culture for generating competent cells to be transformed with the extended resistance cassette.

5 PCR targeting of the *S. coelicolor* cosmid

E. coli BW25113/pIJ790 containing a *S. coelicolor* cosmid is electro-transformed with the extended resistance cassette. **The example described uses the apramycin – *oriT* disruption cassette from pIJ773.** Table **1** lists alternative cassettes and their resistance determinants.

1. Inoculate a 10 ml SOB – $MgSO_4$ culture containing carbenicillin (100 µg/ml), kanamycin (50 µg/ml) and chloramphenicol (25 µg/ml) with 1% of the overnight culture of *E. coli* BW25113/pIJ790 and the *S. coelicolor* cosmid. Add 100 µl 1M L-arabinose stock solution (final concentration is 10 mM, induces *red* genes).
2. Grow for 3-4 h at 30°C shaking at 200 rpm to an OD_{600} of ~ 0.6.
3. Recover the cells by centrifugation at 4000 rpm for 5 min at 4°C in a Sorvall GS3 rotor (or equivalent).
4. Decant medium and resuspend the pellet by gentle mixing in 10 ml ice-cold 10% glycerol.
5. Centrifuge as above and resuspend pellet in 5 ml ice-cold 10 % glycerol, centrifuge and decant. Resuspend the cell pellet in remaining ~ 100 µl 10 % glycerol.
6. Mix 50 µl cell suspension with ~ 100 ng (1-2 µl) of PCR product. Carry out electroporation in a 0.2 cm ice-cold electroporation cuvette using a BioRad GenePulser II set to: 200 Ω, 25 µF and 2,5 kV. The expected time constant is 4.5 – 4.9 ms.
7. Immediately add 1 ml ice cold LB to shocked cells and incubated shaking 1 h at 37°C (or 30°C if further gene disruptions will be made on the same cosmid; see below).
8. Spread onto LB agar containing carbenicillin (100 µg/ml), kanamycin (50 µg/ml) and apramycin (50 µg/ml). If no further gene disruptions will be made on this cosmid, incubate overnight at 37°C to promote the loss of pIJ790. (If further disruptions are planned propagate overnight at 30°C and include chloramphenicol (25 µg/ml) so that pIJ790 is retained).

- **If no colonies are obtained after 16 h growth at 37°C, repeat the experiment starting with a 50 ml SOB culture instead of 10 ml culture for generating electrocompetent cells. Try to concentrate the cells as much as possible by removing all of the remaining 10% glycerol. Resuspend the cell pellet in 50 µl 10% glycerol and use for electroporation.**
- **After 12 – 16 h growth at 37°C different colony-sizes are observed. Cultivating for longer time results in an increased background of small colonies, which are false positives. It is important to note that at this stage wild-type and mutant cosmids exist within one cell. The transformation with a PCR product and its integration in the cosmid DNA by homologous recombination will not occur in all copies of the cosmid molecules in one cell. One copy of a cosmid containing the incoming resistance marker is sufficient for resistance to this antibiotic. Normally, the larger the size of a colony, the more copies of mutagenised cosmids are present. Inoculating a large colony in 5 ml LB liquid cultures containing carbenicillin (100 µg/ml), kanamycin (50 µg/ml) and apramycin (50 µg/ml) result in a growth at 37°C to a cell density (OD_{600} ~ 0.1 – 0.3) within 3-4 h (*E. coli* BW25113 without pIJ790 grows very fast). After 6 h plasmid DNA can be isolated and tested by restriction analysis and/or PCR using the primers described below.**
- **PCR analysis with a primer pair (test primers) priming just ~ 100 bp outside the region affected by homologous recombination will generate the expected fragment after gene disruption, but will usually also generate the wild-type fragment, caused by remaining wild-type copies within the same transformant. These will be lost during the subsequent transformation step into the methylation-deficient *E. coli* host ET12567 containing the non-transmissible plasmid pUZ8002 (this is not a problem anyway because wild-type copies lack the *oriT*).**
- **Notes on viomycin selection: selecting for viomycin^R depends critically on the amount of salt in the medium; more viomycin is required at higher salt concentrations. For a clean selection of *E. coli* clones, use DNA agar or 2xYT agar containing 30 µg/ml viomycin (see Kieser *et al.*, 2000).**

For multiple gene replacements, choose an *oriT*-containing disruption cassette for the first knock-out, and a cassette without *oriT* and different resistance markers for further gene disruptions.

The gene disruption is confirmed by restriction analysis and/or PCR. Cosmid DNA of transformants is isolated from a 6 h, 37°C, 5 ml LB culture containing carbenicillin (100 µg/ml), kanamycin (50 µg/ml) and apramycin (50 µg/ml). Alkaline lysis followed by phenol/chloroform extraction produces cosmid DNA suitable for restriction analysis.

Cosmid CCC DNA isolation

1. Resuspend the cell pellet from 1 ml culture by vortexing in 100 µl solution I (50 mM Tris/HCl, pH 8; 10 mM EDTA).
2. Immediately add 200 µl solution II (200 mM NaOH; 1% SDS) and mix by inverting the tubes 10x.
3. Immediately add 150 µl solution III (3 M potassium acetate, pH 5.5) and mix by inverting the tubes 5x.
4. Spin at full speed in a microcentrifuge for 5 min at room temperature.
5. Immediately extract supernatant with 400 µl phenol/chloroform, vortex 2 min and spin at full speed in a micro centrifuge for 5 min.
6. Transfer the upper phase and add 600 µl 2-propanol. Leave the tubes on ice for 10 min.
7. Spin as above and wash the pellet with 200 µl 70% ethanol.
8. Spin as above and leave the tube open for 5 min at room temperature to dry the pellet. Resuspend the pellet in 50 µl 10mM Tris/HCl (pH 8) and use 10 µl for restriction digest.
    - **Omitting the phenol/chloroform extraction step results in degradation of the cosmid DNA. Use of miniprep-columns without including a phenol/chloroform extraction is not recommended.**

Verification of positive transformants by PCR requires an additional pair of 18 – 20 nt test primers which anneal 100 – 200 bp upstream and downstream of the 39 bp recombination region. (These primers can also be used later to verify the FLP-mediated excision of the resistance cassette.)

• Primers (100 pmoles/µl)	0.2 µl each	20 pmoles each
• Template DNA (~50 ng/µl)	1 µl	50 ng
• Buffer (10x)	5 µl	1 x
• dNTPs (10 mM)	1 µl each	50 µM each
• DMSO (100 %)	2.5 µl	5%
• DNA polymerase (2.5 U/µl)	1 µl	2.5 Units
• Water	36.1 µl	
• Total volume	50 µl	

Cycle conditions:

1. Denaturation:	94°C, 2 min	
2. Denaturation:	94°C, 45 sec	⎫
3. Primer annealing:	55°C, 45 sec	⎬ 30 cycles
4. Extension:	72°C, 90 sec	⎭
5. Final extension:	72°C, 5 min	

5 µl of the PCR product is used for gel electrophoresis.

6 Transfer of the mutant cosmids into *Streptomyces*

If the target *Streptomyces* for mutagenesis carries a methyl-sensing restriction system (as is the case for *S. coelicolor* and *S. avermitilis*), it is necessary to passage the cosmid containing an apramycin resistance-*oriT* cassette through a non-methylating *E. coli* host. To achieve this, it is introduced by transformation into the non-methylating *E. coli* ET12567 containing the RP4 derivative pUZ8002. The cosmid is then transferred to *Streptomyces* by intergeneric conjugation (see Table 2 for resistance markers). If the target *Streptomyces* for mutagenesis does **not** carry a methyl-sensing restriction system (as is the case for *S. lividans*), common *E. coli* strains such as DH5α containing pUZ8002 can be used instead.

Description	**Name**	**Replication**	**Carb^R**	**Cml^R**	**Kan^R**	**Tet^R**
S. coelicolor cosmid clones	Supercos 1		Carb^R		Kan^R	
λ Red plasmid	pIJ790	t^s		Cml^R		
FLP recombinase plasmid	BT340	t^s	Carb^R	Cml^R		
Tra$^-$ RP4 derivative	pUZ8002				Kan^R	
Tra$^+$ RP4 derivative	pUB307				Kan^R	
Non-methylating *E. coli*	ET12567			Cml^R		Tet^R

Table 2. Resistance markers of vectors, helper plasmids and strains (carbenicillin resistance (Carb^R), chloramphenicol resistance (Cml^R), kanamycin resistance (Kan^R), tetracycline resistance (Tet^R), temperature sensitive replicon (t^S)). See Table **1** for replacement cassettes.

1. Prepare competent cells of *E. coli* ET12567/pUZ8002 grown at 37°C in LB containing kanamycin (25 µg/ml) and chloramphenicol (25 µg/ml) to maintain selection for pUZ8002 and the *dam* mutation, respectively. (ET12567 has a doubling time > 30 min.)
   - **High competence is required when Dam-methylated plasmids are introduced into a dam$^-$ strain.**
2. Transform competent cells with the *oriT*-containing cosmid clone, and select for the incoming plasmid only using apramycin (50 µg/ml) and carbenicillin (100 µg/ml) .

3. Inoculate a colony into 10 ml LB containing apramycin (50 μg/ml), chloramphenicol (25 μg/ml) and kanamycin (50 μg/ml). Grow overnight at 37°C.
   - **Chloramphenicol^S or Kanamycin^S segregants arise frequently among transformants, so set up more than one culture. The kanamycin selection is probably ineffective because both the cosmid and pUZ8002 confer resistance (Table 2).**
4. Inoculate 100 μl overnight culture into 10 ml fresh LB plus antibiotics as above and grow for ~ 4 h at 37°C to an OD_{600} of 0.4 - 0.6.
5. Wash the cells twice with 10 ml of LB to remove antibiotics that might inhibit *Streptomyces*, and resuspend in 1 ml of LB.
6. While washing the *E. coli* cells, for each conjugation add 10 μl (10^8) *Streptomyces* spores to 500 μl 2 × YT broth. Heat shock at 50°C for 10 min, then allow to cool.
7. Mix 0.5 ml *E. coli* cell suspension and 0.5 ml heat-shocked spores and spin briefly. Pour off most of the supernatant, then resuspend the pellet in the c. 50 μl residual liquid.
8. Make a dilution series from 10^{-1} to 10^{-4} each step in a total of 100 μl of water.
9. Plate out 100 μl of each dilution on MS agar + 10mM $MgCl_2$ (without antibiotics) and incubate at 30°C for 16-20 h.
10. Overlay the plate with 1 ml water containing 0.5 mg nalidixic acid (20 μl of 25 mg/ml stock; selectively kills *E. coli*) and 1.25 mg apramycin (25 μl of 50 mg/ml stock). Use a spreader to lightly distribute the antibiotic solution evenly. Continue incubation at 30°C.
11. Replica-plate each MS agar plate with single colonies onto DNA plates containing nalidixic acid (25 μg/ml) and apramycin (50 μg/ml) with and without kanamycin (50 μg/ml). Double cross-over exconjugants are kanamycin^S and apramycin^R. (DNA gives fast, non-sporulating growth.)
12. Kanamycin^S clones are picked from the DNA plates and streaked for single colonies on MS agar (promotes sporulation) containing nalidixic acid (25 μg/ml) and apramycin (50 μg/ml).
13. Confirm kanamycin sensitivity by replica-plating onto DNA plates containing nalidixic acid (25 μg/ml) with and without kanamycin (50 μg/ml).
14. Purified kanamycin sensitive strains are then verified by PCR and Southern blot analysis.

- **Typically, ~ 10 % of the exconjugants are double cross-over recombinants. The frequency of double cross-overs depends on the length of the flanking regions of homologous DNA on the cosmid. If < 1 kb is left on one side of the disrupted gene, obtaining kanamycin^S double cross-over types directly on the conjugation plates may be difficult. It may be necessary to streak out several exconjugants for single colonies on MS agar without antibiotics. After 3-5 days growth replica-plate onto DNA with and without kanamycin.**

Antibiotic	Stock mg/ml	µl for 1 ml overlay	Final conc. after flooding µg/ml	Concentration in	
				MS, DNA µg/ml	R2YE µg/ml
Apramycin	50	25	50	50	50
Kanamycin	50	100	200	50	200
Spectinomycin	200	25	200	400	400
Streptomcyin	10	25	10	10	10
Viomycin	30	25	30	30	NA
Nalidixic acid	25 in 0.3 M NaOH	20	20	25	25

Table 3: Antibiotic concentrations for selection on *S. coelicolor* MS conjugation plates, DNA replica plates or R2YE protoplast regeneration plates (Note some small differences from Kieser *et al.*, 2000).

7 FLP-mediated excision of the disruption cassette

The disruption cassettes are flanked by FRT sites (FLP recognition targets). Expression of the FLP-recombinase in *E. coli* removes the central part of the disruption cassette, leaving behind a 81 bp "scar" sequence which, in the preferred reading frame (bold in Fig. 3), lacks stop codons.

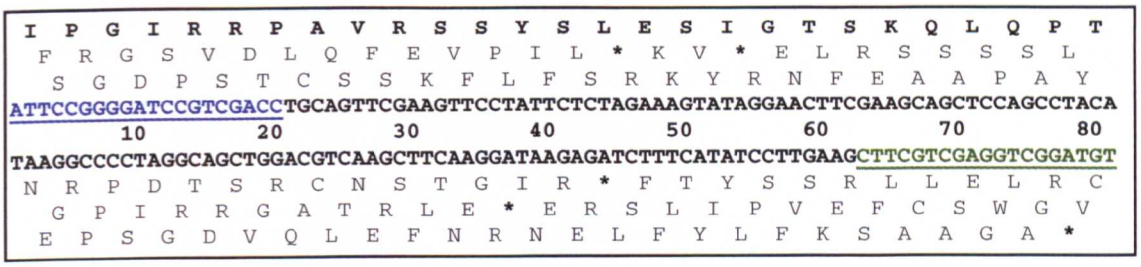


Fig. 3: Sequence of the 81 bp "scar" sequence remaining after FLP-mediated excision of the disruption cassette. The translation of the preferred reading frame is printed bold. The 20 and 19 nt priming sites are underlined and printed in colour. (Fig. **2** explains the determination of the reading frame.)

★ indicate stop codons,

priming site (20 nt) priming site (19 nt)

This allows the generation of (hopefully) non-polar, unmarked in-frame deletions and repeated use of the same resistance marker for making multiple knock-outs in the same cosmid or in the same strain. *E. coli* DH5α cells containing the temperature sensitive FLP recombination plasmid BT340 (Datsenko and Wanner, 2000; can be obtained from the *E. coli* Genetic Stock Center: CGSC Strain# 7629) are transformed with the mutagenised cosmid DNA (obtained in section 5). BT340 contains ampicillin and chloramphenicol resistance determinants and is temperature sensitive for replication (replicates at 30°C). FLP synthesis and loss of the plasmid are induced at 42°C (Cherepanov and Wackernagel, 1995).

1. Grow *E. coli* DH5α/BT340 overnight at 30°C in 10 ml LB containing chloramphenicol (25 µg/ml).
   - **Transforming *E. coli* BW25113/cosmid::apramycin (mutagenised cosmid) with the plasmid BT340 is not recommended because the isolates after PCR targeting may still contain copies of undisrupted cosmid DNA (see page 10, second paragraph).**
2. Inoculate 100 µl *E. coli* DH5α/BT340 from overnight culture into 10 ml LB containing chloramphenicol (25 µg/ml).
3. Grow for 3-4 h at 30°C shaking at 200 rpm to an OD_{600} of ~ 0.6.
4. Recover the cells by centrifugation at 4000 rpm for 5 min at 4°C in a Sorvall GS3 rotor (or equivalent).
5. Decant medium and resuspend the pellet by gentle mixing in 10 ml ice-cold 10 % glycerol.
6. Centrifuge as above and resuspend pellet in 5 ml ice-cold 10 % glycerol, centrifuge and decant. Resuspend the cell pellet in remaining ~ 100 µl 10% glycerol.
7. Mix 50 µl cell suspension with ~ 100 ng (1-2 µl) of mutagenised cosmid DNA. Carry out electroporation in a 0.2 cm ice-cold electroporation cuvette using a BioRad GenePulser II set to: 200 Ω, 25 µF and 2,5 kV. The expected time constant is 4.5 – 4.9 ms.
8. Immediately add 1 ml ice cold LB to shocked cells and incubate shaking for 1 h at 30°C.
9. Spread onto LB agar containing apramycin (50 µg/ml) and chloramphenicol (25 µg/ml).
10. Incubate for 2 d at 30°C (*E. coli* DH5α/BT340 grows slowly at 30°C).
11. A single colony is streaked on an LB agar plate without antibiotics for single colonies and grown overnight at 42°C to induce expression of the FLP recombinase followed by the loss of plasmid BT340.
12. Make two masterplates by streaking 20 – 30 single colonies with a toothpick first on a LB agar plate containing apramycin (50 µg/ml) and then on a LB agar plate containing kanamycin (50 µg/ml).

13. Grow the masterplates overnight at 37°C. Apramycin^S kanamycin^R clones indicate the successful loss of the resistance cassette and are further verified by restriction and PCR analysis.
    - **Typically, ~ 10 % of the single colonies after non-selective growth lose the incoming resistance marker and the plasmid BT340 simultaneously.**
    - **Using the same test primers as in section 5 (annealing ~ 100 bp upstream and downstream of the 39 nt primer sequence) should produce a PCR product of ~ 300 bp (200 bp + 81 bp "scar"). PCR fragments can be sequenced using the amplification primers for verification.**

Replacing resistance cassette inserts in *S. coelicolor* with the unmarked "scar" sequence

The chromosomal apramycin resistance cassette insert in *S. coelicolor* is replaced by the "scar" sequence. This is achieved by homologous recombination between the chromosome and the corresponding "scar cosmid" prepared in 7. The procedure differs from section 6 because the cosmid lacks *oriT*, and the desired product is antibiotic sensitive. Therefore, it is necessary to introduce the scar cosmid into *Streptomyces* by protoplast transformation, and then select for kanamycin resistant *Streptomyces* containing the entire scar cosmid integrated by a single crossover. Restreaking to kanamycin-free medium, followed by screening for concomitant loss of kanamycin resistance and apramycin resistance, then identifies the desired *Streptomyces* clones.

Preparation of *Streptomyces coelicolor* protoplasts

1. Add 25 ml YEME medium to a baffled flask. Add ~ 0.1 ml spore suspension and required growth factors. Incubate 36-40 h at 30°C in an orbital incubator shaker.
    - **Cultures of *S. lividans* and *S. coelicolor* are ready for harvesting when they start to produce red pigment**
2. Pour culture broth into a 20 ml screw cap bottle and spin in the bench centrifuge (~ 1000 x g, 10 min).
    - **Before centrifugation, examine the culture for contamination by unicellular bacteria, usually indicated by turbidity: the *Streptomyces* mycelium sediments**

quickly while unicellular contaminants remain suspended. In case of doubt, use the microscope.

3. Discard the supernatant carefully; the pellet is easily disturbed.
   - **If the mycelium does not pellet add 5 ml sterile water to reduce the density of the medium and centrifuge again.**
4. Resuspend pellet in 15 ml 10.3% sucrose and spin in bench centrifuge as above. Discard supernatant.
5. Repeat step 4.
   - **The mycelial pellet, without added liquid, can be stored frozen at –20°C**
6. Resuspend mycelium in 4 ml lysozyme solution (1 mg/ml P buffer, filter sterilised); incubate at 30°C, 15-60 min.
7. Draw in and out of a 5 ml pipette three times and incubate for a further 15 min.
   - **This helps to free protoplasts from the mycelium so that they will pass through the cotton wool filter used in step 9. At least with *S. lividans*, it is possible to obtain transformants with unfiltered material, but the washing (steps 9-10) is still needed to remove lysozyme.**
8. Add 5 ml P buffer. Repeat step 7.
9. Filter protoplasts through cotton wool (using a filter tube) and transfer to a plastic tube.
10. Sediment protoplasts gently by spinning in a bench centrifuge (~ 1000 x g, 7 min).
11. Discard supernatant and suspend protoplasts in 1 ml P buffer.
    - **At this and any other steps when pelleted protoplasts are to be resuspended, resuspend in the remaining drop of liquid by tapping the side of the tube repeatedly with a finger until the protoplasts are dispersed to form a creamy suspension, then add the suspending P buffer (otherwise the protoplast pellet is difficult to disperse). Avoid vortexing, which induces foaming and consequent lysis. To freeze the protoplasts for storage, place samples of the protoplast suspension in small plastic tubes, close them and place them in ice in a plastic beaker. Place the beaker at –70°C overnight. Free the frozen protoplasts in their tubes from the ice and store at –70°C. To thaw, shake the frozen tube under running warm water (i.e. freeze slowly, thaw quickly). To assess the proportion of non-protoplasted units in the suspension, samples can be diluted in parallel in P buffer and in dilute detergent (~ 0.01% SDS) and plated on regeneration plates. Any colonies arising after dilution in detergent are likely to have arisen from non-protoplasted units.**

Rapid small-scale transformation of *Streptomyces coelicolor*

1. Dispense 50 µl samples of protoplasts (~ 10^{10}/ml) into as many tubes as there are transformations.
   - **We usually spin the protoplasts down immediately before the transformation experiment. This eliminates substances that may have leaked out of the protoplasts during storage and the contents of protoplasts which have lysed spontaneously (which may include nucleases).**
2. Complete steps 2a-c for each transformation individually.
   a. Add up to 5 µl DNA solution to protoplasts and mix immediately by tapping tube.
   b. Add 200 µl 25% PEG 1000 in P buffer and mix by pipetting up and down four times (be careful not to contaminate the barrel of the pipette).
   c. Spread protoplast suspension (100-200 µl) on two dried R2YE plates. Use P buffer to make dilutions if required.
      - **1 ml glass pipettes can be used instead of spreaders. The solution will spread to some extent by itself if the plates are left on a horizontal surface.**
3. Incubate plates at 30°C. After 14-20 h, flood for kanamycin selection. Score for resistant colonies after 3 d.
4. Select single colony and streak non-selectively for single colonies on MS agar plates and grow 3-4 d at 30°C.
5. Replica-plate to DNA agar plates with apramycin (50 µg/ml) or kanamycin (50 µg/ml) to screen for apramycin^S and kanamycin^S transformants.

The λ RED recombination plasmid pIJ790

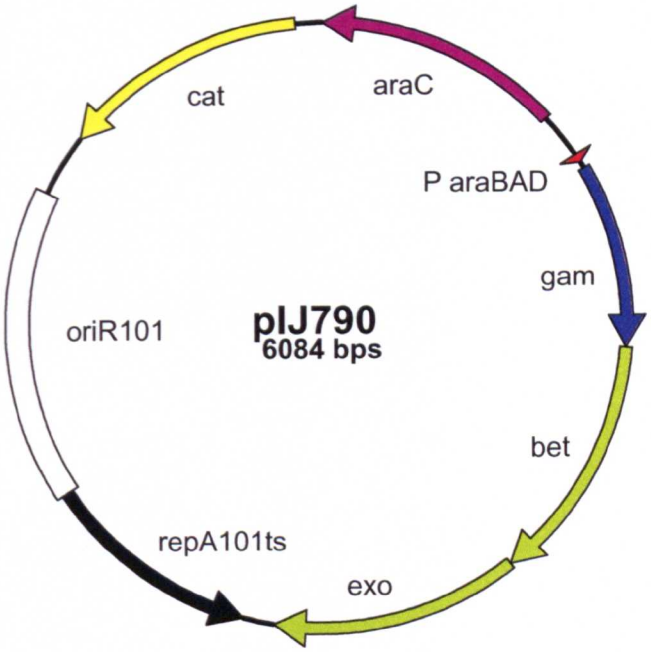


Fig. 4: The λ RED recombination plasmid pKD20 (*E. coli* Genetic Sock Center CGSC Strain# 7637; Datsenko and Wanner, 2000) was modified by replacing the ampicillin resistance gene *bla* by the chloramphenicol resistance gene *cat*, generating pIJ790. Unlike pKD20, pIJ790 can be selected in presence of Supercos1.

Legend:

araC: encodes arabinose activator

bet: encodes single strand DNA binding protein

cat: chloramphenicol resistance gene

exo: exonuclease gene, promotes recombination together with *bet*

gam: gene product inhibits the host exonuclease V

oriR101: origin of replication

P araBAD: L-arabinose inducible promotor

repA101ts: temperature-sensitive replication

The following digestions can be used to verify this plasmid.

*Bam*HI: 6084 bp; *Eco*RI: 2872 bp, 1703bp, 1509 bp;

*Nco*I: 3927 bp, 2157 bp; *Pst*I: 5873 bp, 247 bp

Template plasmids

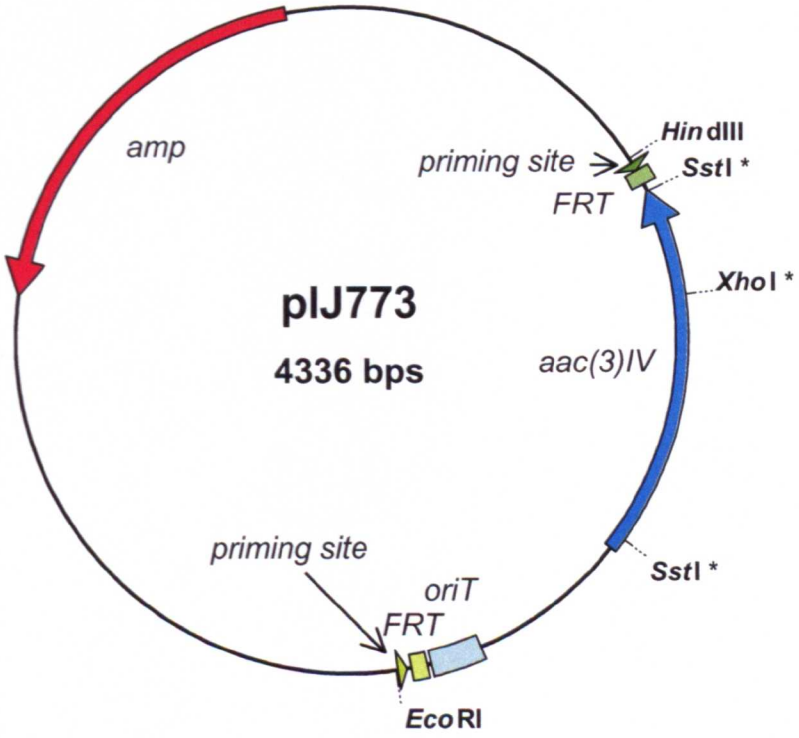


Fig. 5: Template plasmid pIJ773 containing the apramycin resistance gene *aac(3)IV* (AC=X99313) and the *oriT* of plasmid RP4 (=RK2) (AC=L27758), flanked by FRT sites (FLP recognition targets, see Datsenko and Wanner, 2000). The disruption cassette was cloned into the *Eco*RV site of pBluescript KS (+) allowing its isolation as a 1398 bp *Eco*RI/*Hin*dIII fragment.

* indicates suitable restriction sites for verification of mutagenised cosmid DNA by restriction analysis (for example: *Sst*I generates a 751 bp internal fragment within the disruption cassette).

```
LOCUS       pIJ773       4336 bp    DNA    CIRCULAR SYN        17-OCT-2001
DEFINITION  Ligation of Apra-oriT disruption cassette into the EcoRV site of
pBluescript SK(+)
ACCESSION   pIJ773
REFERENCE   1  (bases 1 to 4336)
  AUTHORS   Self
  JOURNAL   Unpublished.
FEATURES             Location/Qualifiers
     CDS             699..717
                     /marker="Primer site reverse"
                     /product="TGT AGG CTG GAG CTG CTT C"
     CDS             complement (718..751)
                     /region="FRT"
```

```
                     /product="Natural FRT site"
     CDS             complement (764..1549)
                     /gene="aac(3)IV"
                     /product="aminoglycoside acetyltransferase inactivating
apramycin"
     CDS             1890..1999
                     /region="oriT"
                     /product="origin of transfer from plasmid RP4 (also designated
RP1/RK2)"
     CDS             2008..2041
                     /region="FRT"
                     /product="Natural FRT site"
     CDS             complement (2050..2069)
                     /marker="Primer site forward"
                     /product="ATT CCG GGG ATC CGT CGA CC"
     CDS             complement (3345..4205)
                     /gene="amp"
                     /product="b-lactamase"
BASE COUNT     1048 a   1121 c   1143 g   1024 t
ORIGIN
        1 GGAAATTGTA AACGTTAATA TTTTGTTAAA ATTCGCGTTA AATTTTTGTT AAATCAGCTC
       61 ATTTTTTAAC CAATAGGCCG AAATCGGCAA AATCCCTTAT AAATCAAAAG AATAGACCGA
      121 GATAGGGTTG AGTGTTGTTC CAGTTTGGAA CAAGAGTCCA CTATTAAAGA ACGTGGACTC
      181 CAACGTCAAA GGGCGAAAAA CCGTCTATCA GGGCGATGGC CCACTACGTG AACCATCACC
      241 CTAATCAAGT TTTTTGGGGT CGAGGTGCCG TAAAGCACTA AATCGGAACC CTAAAGGGAG
      301 CCCCCGATTT AGAGCTTGAC GGGGAAAGCC GGCGAACGTG GCGAGAAAGG AAGGGAAGAA
      361 AGCGAAAGGA GCGGGCGCTA GGGCGCTGGC AAGTGTAGCG GTCACGCTGC GCGTAACCAC
      421 CACACCCGCC GCGCTTAATG CGCCGCTACA GGGCGCGTCG CGCCATTCGC CATTCAGGCT
      481 GCGCAACTGT TGGGAAGGGC GATCGGTGCG GGCCTCTTCG CTATTACGCC AGCTGGCGAA
      541 AGGGGGATGT GCTGCAAGGC GATTAAGTTG GGTAACGCCA GGGTTTTCCC AGTCACGACG
      601 TTGTAAAACG ACGGCCAGTG AATTGTAATA CGACTCACTA TAGGGCGAAT TGGGTACCGG
      661 GCCCCCCCTC GAGGTCGACG GTATCGATAA GCTTGATGTG TAGGCTGGAG CTGCTTCGAA
      721 GTTCCTATAC TTTCTAGAGA ATAGGAACTT CGGAATAGGA ACTTCATGAG CTCAGCCAAT
      781 CGACTGGCGA GCGGCATCGC ATTCTTCGCA TCCCGCCCTC TGGCGGATGC AGGAAGATCA
      841 ACGGATCTCG GCCCAGTTGA CCCAGGGCTG TCGCCACAAT GTCGCGGGAG CGGATCAACC
      901 GAGCAAAGGC ATGACCGACT GGACCTTCCT TCTGAAGGCT CTTCTCCTTG AGCCACCTGT
      961 CCGCCAAGGC AAAGCGCTCA CAGCAGTGGT CATTCTCGAG ATAATCGACG CGTACCAACT
     1021 TGCCATCCTG AAGAATGGTG CAGTGTCTCG GCACCCCATA GGGAACCTTT GCCATCAACT
     1081 CGGCAAGATG CAGCGTCGTG TTGGCATCGT GTCCCACGCC GAGGAGAAGT ACCTGCCCAT
     1141 CGAGTTCATG GACACGGGCG ACCGGGCTTG CAGGCGAGTG AGGTGGCAGG GGCAATGGAT
     1201 CAGAGATGAT CTGCTCTGCC TGTGGCCCCG CTGCCGCAAA GGCAAATGGA TGGGCGCTGC
     1261 GCTTTACATT TGGCAGGCGC CAGAATGTGT CAGAGACAAC TCCAAGGTCC GGTGTAACGG
     1321 GCGACGTGGC AGGATCGAAC GGCTCGTCGT CCAGACCTGA CCACGAGGGC ATGACGAGCG
     1381 TCCCTCCCGG ACCCAGCGCA GCACGCAGGG CCTCGATCAG TCCAAGTGGC CCATCTTCGA
     1441 GGGGCCGGAC GCTACGGAAG GAGCTGTGGA CCAGCAGCAC ACCGCCGGGG GTAACCCCAA
     1501 GGTTGAGAAG CTGACCGATG AGCTCGGCTT TTCGCCATTC GTATTGCACG ACATTGCACT
     1561 CCACCGCTGA TGACATCAGT CGATCATAGC ACGATCAACG GCACTGTTGC AAATAGTCGG
     1621 TGGTGATAAA CTTATCATCC CCTTTTGCTG ATGGAGCTGC ACATGAACCC ATTCAAAGGC
     1681 CGGCATTTTC AGCGTGACAT CATTCTGTGG GCCGTACGCT GGTACTGCAA ATACGGCATC
     1741 AGTTACCGTG AGCTGCATTT TCCGCTGCAT AACCCTGCTT CGGGGTCATT ATAGCGATTT
     1801 TTTCGGTATA TCCATCCTTT TTCGCACGAT ATACAGGATT TTGCCAAAGG GTTCGTGTAG
     1861 ACTTTCCTTG GTGTATCCAA CGGCGTCAGC CGGGCAGGAT AGGTGAAGTA GGCCCACCCG
     1921 CGAGCGGGTG TTCCTTCTTC ACTGTCCCTT ATTCGCACCT GGCGGTGCTC AACGGGAATC
     1981 CTGCTCTGCG AGGCTGGCGG GAACTTCGAA GTTCCTATAC TTTCTAGAGA ATAGGAACTT
     2041 CGAACTGCAG GTCGACGGAT CCCCGGAATA TCGAATTCCT GCAGCCCGGG GGATCCACTA
     2101 GTTCTAGAGC GGCCGCCACC GCGGTGGAGC TCCAGCTTTT GTTCCCTTTA GTGAGGGTTA
     2161 ATTCCGAGCT TGGCGTAATC ATGGTCATAG CTGTTTCCTG TGTGAAATTG TTATCCGCTC
     2221 ACAATTCCAC ACAACATACG AGCCGGAAGC ATAAAGTGTA AAGCCTGGGG TGCCTAATGA
     2281 GTGAGCTAAC TCACATTAAT TGCGTTGCGC TCACTGCCCG CTTTCCAGTC GGGAAACCTG
     2341 TCGTGCCAGC TGCATTAATG AATCGGCCAA CGCGCGGGGA GAGGCGGTTT GCGTATTGGG
     2401 CGCTCTTCCG CTTCCTCGCT CACTGACTCG CTGCGCTCGG TCGTTCGGCT GCGGCGAGCG
     2461 GTATCAGCTC ACTCAAAGGC GGTAATACGG TTATCCACAG AATCAGGGGA TAACGCAGGA
     2521 AAGAACATGT GAGCAAAAGG CCAGCAAAAG GCCAGGAACC GTAAAAAGGC CGCGTTGCTG
     2581 GCGTTTTTCC ATAGGCTCCG CCCCCCTGAC GAGCATCACA AAAATCGACG CTCAAGTCAG
     2641 AGGTGGCGAA ACCCGACAGG ACTATAAAGA TACCAGGCGT TTCCCCCTGG AAGCTCCCTC
     2701 GTGCGCTCTC CTGTTCCGAC CCTGCCGCTT ACCGGATACC TGTCCGCCTT TCTCCCTTCG
     2761 GGAAGCGTGG CGCTTTCTCA TAGCTCACGC TGTAGGTATC TCAGTTCGGT GTAGGTCGTT
     2821 CGCTCCAAGC TGGGCTGTGT GCACGAACCC CCCGTTCAGC CCGACCGCTG CGCCTTATCC
     2881 GGTAACTATC GTCTTGAGTC CAACCCGGTA AGACACGACT TATCGCCACT GGCAGCAGCC
     2941 ACTGGTAACA GGATTAGCAG AGCGAGGTAT GTAGGCGGTG CTACAGAGTT CTTGAAGTGG
     3001 TGGCCTAACT ACGGCTACAC TAGAAGGACA GTATTTGGTA TCTGCGCTCT GCTGAAGCCA
     3061 GTTACCTTCG GAAAAAGAGT TGGTAGCTCT TGATCCGGCA AACAAACCAC CGCTGGTAGC
     3121 GGTGGTTTTT TTGTTTGCAA GCAGCAGATT ACGCGCAGAA AAAAAGGATC TCAAGAAGAT
     3181 CCTTTGATCT TTTCTACGGG GTCTGACGCT CAGTGGAACG AAAACTCACG TTAAGGGATT
     3241 TTGGTCATGA GATTATCAAA AAGGATCTTC ACCTAGATCC TTTTAAATTA AAAATGAAGT
     3301 TTTAAATCAA TCTAAAGTAT ATATGAGTAA ACTTGGTCTG ACAGTTACCA ATGCTTAATC
     3361 AGTGAGGCAC CTATCTCAGC GATCTGTCTA TTTCGTTCAT CCATAGTTGC CTGACTCCCC
```

```
3421 GTCGTGTAGA TAACTACGAT ACGGGAGGGC TTACCATCTG GCCCCAGTGC TGCAATGATA
3481 CCGCGAGACC CACGCTCACC GGCTCCAGAT TTATCAGCAA TAAACCAGCC AGCCGGAAGG
3541 GCCGAGCGCA GAAGTGGTCC TGCAACTTTA TCCGCCTCCA TCCAGTCTAT TAATTGTTGC
3601 CGGGAAGCTA GAGTAAGTAG TTCGCCAGTT AATAGTTTGC GCAACGTTGT TGCCATTGCT
3661 ACAGGCATCG TGGTGTCACG CTCGTCGTTT GGTATGGCTT CATTCAGCTC CGGTTCCCAA
3721 CGATCAAGGC GAGTTACATG ATCCCCCATG TTGTGCAAAA AAGCGGTTAG CTCCTTCGGT
3781 CCTCCGATCG TTGTCAGAAG TAAGTTGGCC GCAGTGTTAT CACTCATGGT TATGGCAGCA
3841 CTGCATAATT CTCTTACTGT CATGCCATCC GTAAGATGCT TTTCTGTGAC TGGTGAGTAC
3901 TCAACCAAGT CATTCTGAGA ATAGTGTATG CGGCGACCGA GTTGCTCTTG CCCGGCGTCA
3961 ATACGGGATA ATACCGCGCC ACATAGCAGA ACTTTAAAAG TGCTCATCAT TGGAAAACGT
4021 TCTTCGGGGC GAAAACTCTC AAGGATCTTA CCGCTGTTGA GATCCAGTTC GATGTAACCC
4081 ACTCGTGCAC CCAACTGATC TTCAGCATCT TTTACTTTCA CCAGCGTTTC TGGGTGAGCA
4141 AAAACAGGAA GGCAAAATGC CGCAAAAAAG GGAATAAGGG CGACACGGAA ATGTTGAATA
4201 CTCATACTCT TCCTTTTTCA ATATTATTGA AGCATTTATC AGGGTTATTG TCTCATGAGC
4261 GGATACATAT TTGAATGTAT TTAGAAAAAT AAACAAATAG GGGTTCCGCG CACATTTCCC
4321 CGAAAAGTGC CACCTG
//
```

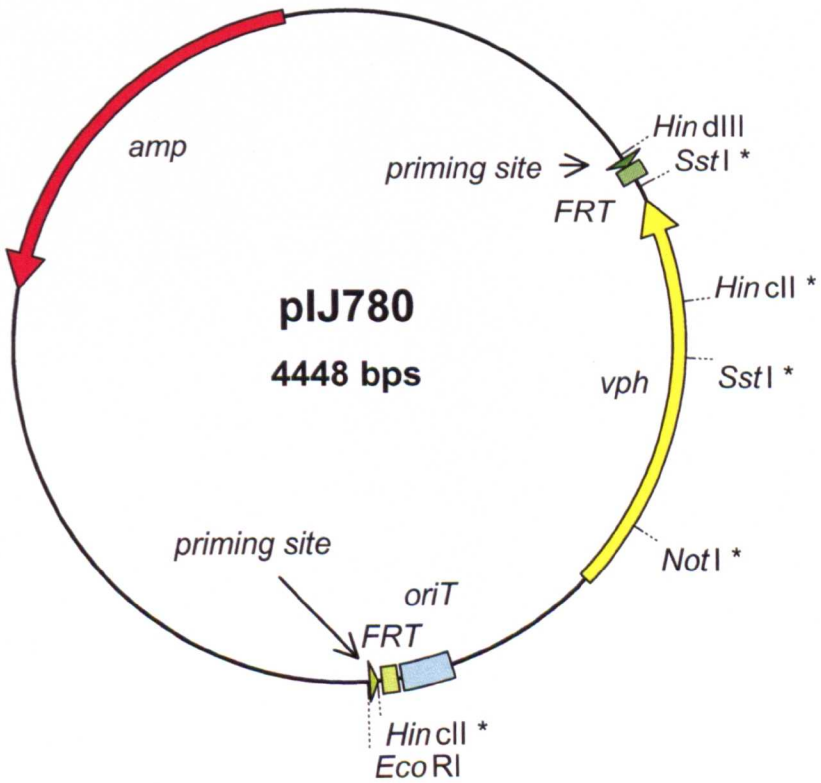


Fig. 7: Template plasmid pIJ780 containing the viomycin resistance gene *vph* (AC=X99314) and the *oriT* of plasmid RP4 (=RK2) (AC=L27758), flanked by FRT sites (FLP recognition targets, see Datsenko and Wanner, 2000). The disruption cassette was cloned into the *Eco*RV site of pBluescript KS (+) allowing its isolation as a 1510 bp *Eco*RI/*Hin*dIII fragment.

* indicates suitable restriction sites for verification of mutagenised cosmid DNA by restriction analysis (for example: *Hinc*II generates a 1137 bp internal fragment within the disruption cassette).

```
LOCUS       pIJ780       4448 bp    DNA   CIRCULAR SYN        17-OCT-2001
DEFINITION  Ligation of Vio-oriT disruption cassette into the EcoRV site of
pBluescript SK(+)
ACCESSION   pIJ780
KEYWORDS    .
REFERENCE   1  (bases 1 to 4448)
  AUTHORS   Self
  JOURNAL   Unpublished.
FEATURES             Location/Qualifiers
     CDS             699..717
                     /marker="Primer site reverse"
                     /product="TGT AGG CTG GAG CTG CTT C"
     CDS             complement (718..751)
                     /region="FRT"
                     /product="Natural FRT site"
     CDS             complement (798..1661)
                     /gene="vph"
                     /product="viomycin phosphotransferase gene from S.vinaceus"
     CDS             2002..2111
                     /region="oriT"
                     /product="origin of transfer from plasmid RP4 (also designated
RP1/RK2)"
     CDS             2120..2153
                     /region="FRT"
                     /product="Natural FRT site"
     CDS             complement (2162..2181)
                     /marker="Primer site forward"
                     /product="ATT CCG GGG ATC CGT CGA CC"
     CDS             complement (3457..4317)
                     /gene="amp"
                     /product="b-lactamase"
BASE COUNT     1002 a   1223 c   1211 g   1012 t
ORIGIN
        1 GGAAATTGTA AACGTTAATA TTTTGTTAAA ATTCGCGTTA AATTTTTGTT AAATCAGCTC
       61 ATTTTTTAAC CAATAGGCCG AAATCGGCAA AATCCCTTAT AAATCAAAAG AATAGACCGA
      121 GATAGGGTTG AGTGTTGTTC CAGTTTGGAA CAAGAGTCCA CTATTAAAGA ACGTGGACTC
      181 CAACGTCAAA GGGCGAAAAA CCGTCTATCA GGGCGATGGC CCACTACGTG AACCATCACC
      241 CTAATCAAGT TTTTTGGGGT CGAGGTGCCG TAAAGCACTA ATCGGAACCC CTAAAGGGAG
      301 CCCCCGATTT AGAGCTTGAC GGGGAAAGCC GGCGAACGTG GCGAGAAAGG AAGGGAAGAA
      361 AGCGAAAGGA GCGGGCGCTA GGGCGCTGGC AAGTGTAGCG GTCACGCTGC GCGTAACCAC
      421 CACACCCGCC GCGCTTAATG CGCCGCTACA GGGCGCGTCG CGCCATTCGC CATTCAGGCT
      481 GCGCAACTGT TGGGAAGGGC GATCGGTGCG GGCCTCTTCG CTATTACGCC AGCTGGCGAA
      541 AGGGGGATGT GCTGCAAGGC GATTAAGTTG GGTAACGCCA GGGTTTTCCC AGTCACGACG
      601 TTGTAAAACG ACGGCCAGTG AATTGTAATA CGACTCACTA TAGGGCGAAT TGGGTACCGG
      661 GCCCCCCCTC GAGGTCGACG GTATCGATAA GCTTGATGTG TAGGCTGGAG CTGCTTCGAA
      721 GTTCCTATAC TTTCTAGAGA ATAGGAACTT CGGAATAGGA ACTTCATGAG CTCAGCCAAT
      781 CGACTGGCGA GCGGCATCTA CCGGTAGCCG CTGAGGCCGT CGGCGAGTTC CTCCTCGTCG
      841 CCGTCGCGCT GCGCGTAGAG GGCCTGCTGG AGTGCGAAGG TGCCTCGGAT CGCCGAGATC
      901 CGCTCGGCCG TTCCGTTGTC GGCCCAGCCG CCGAGCGCGA GCACTCGGCC CAGCAGTTCC
      961 TCGCCGTAGC TCGCCCCGAT GGCGGCCAGG TCCTCAGCCG GGTCGCCGAT GCCGACCTCG
     1021 TCCCAGTCGA CGACGCCGCT CATGCGCGGC ACTCCGTCCA CCGTCTCCCA CAGGACGTTC
     1081 TCGCCGCCGA GGTCACCGTG GACCACCGCG GAGGTGAGAT GGGGCAGGGC GTCGAGCGCG
     1141 GCGAGCTCGC GCTCGGCACG CTCCCGGCCG CCGTCGGACA TCAGCGGGAA CAGTTCGGTA
     1201 CGCACCCCCG TGGCGAACTC CTGCCACTCG TTCGCGGGAG CCTCCGGCAG CGCGGCGCGC
     1261 ACCTTCTCCT CGTCGCCCGC CGCCGCGAGC CCGGACAGCA GGGTCGCGTA CTGTCGGGCG
     1321 ACGGCCTCCG CCACCTCCGG GCTGGTGAGC ACATCGTCCT CCAACGGTGC TCCGGGAATG
     1381 CGGCTCAGCA CCAGGTACGG CGGCTCGTCC GTGCCCTGGG CGCCGCCCTC GGACAGCGGC
     1441 TGCGGCGTGC GAAACCCGAG GTCGATCCCG GCAAGAGCGC GCAGGACGTC CGCCCTGCCG
     1501 GGCAGACGGT CGGCGGCCGC CCGGGTGCGG GCGAAGCAGA CCACCCGGTG CGATCCGATC
     1561 ACCACATGGT GGAACTGCCC CTCGTGGACG GCGAGTCCGC CCACGGTGTC CCCGGGCAGG
     1621 AGCCGGCTCA GCAGATCGCG GTGCGTCTCA ATGATTCTCA TGACATTGCA CTCCACCGCT
     1681 GATGACATCA GTCGATCATA GCACGATCAA CGGCACTGTT GCAAATAGTC GGTGGTGATA
     1741 AACTTATCAT CCCCTTTTGC TGATGGAGCT GCACATGAAC CCATTCAAAG GCCGGCATTT
     1801 TCAGCGTGAC ATCATTCTGT GGGCCGTACG CTGGTACTGC AAATACGGCA TCAGTTACCG
     1861 TGAGCTGCAT TTTCCGCTGC ATAACCCTGC TTCGGGGTCA TTATAGCGAT TTTTTCGGTA
     1921 TATCCATCCT TTTTCGCACG ATATACAGGA TTTTGCCAAA GGGTTCGTGT AGACTTTCCT
     1981 TGGTGTATCC AACGGCGTCA GCCGGGCAGG ATAGGTGAAG TAGGCCCACC CGCGAGCGGG
     2041 TGTTCCTTCT TCACTGTCCC TTATTCGCAC CTGGCGGTGC TCAACGGGAA TCCTGCTCTG
     2101 CGAGGCTGGC GGGAACTTCG AAGTTCCTAT ACTTTCTAGA GAATAGGAAC TTCGAACTGC
     2161 AGGTCGACGG ATCCCCGGAA TATCGAATTC CTGCAGCCCG GGGGATCCAC TAGTTCTAGA
     2221 GCGGCCGCCA CCGCGGTGGA GCTCCAGCTT TTGTTCCCTT TAGTGAGGGT TAATTCCGAG
     2281 CTTGGCGTAA TCATGGTCAT AGCTGTTTCC TGTGTGAAAT TGTTATCCGC TCACAATTCC
     2341 ACACAACATA CGAGCCGGAA GCATAAAGTG TAAAGCCTGG GGTGCCTAAT GAGTGAGCTA
     2401 ACTCACATTA ATTGCGTTGC GCTCACTGCC CGCTTTCCAG TCGGGAAACC TGTCGTGCCA
     2461 GCTGCATTAA TGAATCGGCC AACGCGCGGG GAGAGGCGGT TTGCGTATTG GGCGCTCTTC
```

```
2521 CGCTTCCTCG CTCACTGACT CGCTGCGCTC GGTCGTTCGG CTGCGGCGAG CGGTATCAGC
2581 TCACTCAAAG GCGGTAATAC GGTTATCCAC AGAATCAGGG GATAACGCAG GAAAGAACAT
2641 GTGAGCAAAA GGCCAGCAAA AGGCCAGGAA CCGTAAAAAG GCCGCGTTGC TGGCGTTTTT
2701 CCATAGGCTC CGCCCCCCTG ACGAGCATCA CAAAAATCGA CGCTCAAGTC AGAGGTGGCG
2761 AAACCCGACA GGACTATAAA GATACCAGGC GTTTCCCCCT GGAAGCTCCC TCGTGCGCTC
2821 TCCTGTTCCG ACCCTGCCGC TTACCGGATA CCTGTCCGCC TTTCTCCCTT CGGGAAGCGT
2881 GGCGCTTTCT CATAGCTCAC GCTGTAGGTA TCTCAGTTCG GTGTAGGTCG TTCGCTCCAA
2941 GCTGGGCTGT GTGCACGAAC CCCCCGTTCA GCCCGACCGC TGCGCCTTAT CCGGTAACTA
3001 TCGTCTTGAG TCCAACCCGG TAAGACACGA CTTATCGCCA CTGGCAGCAG CCACTGGTAA
3061 CAGGATTAGC AGAGCGAGGT ATGTAGGCGG TGCTACAGAG TTCTTGAAGT GGTGGCCTAA
3121 CTACGGCTAC ACTAGAAGGA CAGTATTTGG TATCTGCGCT CTGCTGAAGC CAGTTACCTT
3181 CGGAAAAAGA GTTGGTAGCT CTTGATCCGG CAAACAAACC ACCGCTGGTA GCGGTGGTTT
3241 TTTTGTTTGC AAGCAGCAGA TTACGCGCAG AAAAAAAGGA TCTCAAGAAG ATCCTTTGAT
3301 CTTTTCTACG GGGTCTGACG CTCAGTGGAA CGAAAACTCA CGTTAAGGGA TTTTGGTCAT
3361 GAGATTATCA AAAAGGATCT TCACCTAGAT CCTTTTAAAT TAAAAATGAA GTTTTAAATC
3421 AATCTAAAGT ATATATGAGT AAACTTGGTC TGACAGTTAC CAATGCTTAA TCAGTGAGGC
3481 ACCTATCTCA GCGATCTGTC TATTTCGTTC ATCCATAGTT GCCTGACTCC CCGTCGTGTA
3541 GATAACTACG ATACGGGAGG GCTTACCATC TGGCCCCAGT GCTGCAATGA TACCGCGAGA
3601 CCCACGCTCA CCGGCTCCAG ATTTATCAGC AATAAACCAG CCAGCCGGAA GGGCCGAGCG
3661 CAGAAGTGGT CCTGCAACTT TATCCGCCTC CATCCAGTCT ATTAATTGTT GCCGGGAAGC
3721 TAGAGTAAGT AGTTCGCCAG TTAATAGTTT GCGCAACGTT GTTGCCATTG CTACAGGCAT
3781 CGTGGTGTCA CGCTCGTCGT TTGGTATGGC TTCATTCAGC TCCGGTTCCC AACGATCAAG
3841 GCGAGTTACA TGATCCCCCA TGTTGTGCAA AAAAGCGGTT AGCTCCTTCG GTCCTCCGAT
3901 CGTTGTCAGA AGTAAGTTGG CCGCAGTGTT ATCACTCATG GTTATGGCAG CACTGCATAA
3961 TTCTCTTACT GTCATGCCAT CCGTAAGATG CTTTTCTGTG ACTGGTGAGT ACTCAACCAA
4021 GTCATTCTGA GAATAGTGTA TGCGGCGACC GAGTTGCTCT TGCCCGGCGT CAATACGGGA
4081 TAATACCGCG CCACATAGCA GAACTTTAAA AGTGCTCATC ATTGGAAAAC GTTCTTCGGG
4141 GCGAAAACTC TCAAGGATCT TACCGCTGTT GAGATCCAGT TCGATGTAAC CCACTCGTGC
4201 ACCCAACTGA TCTTCAGCAT CTTTTACTTT CACCAGCGTT TCTGGGTGAG CAAAAACAGG
4261 AAGGCAAAAT GCCGCAAAAA AGGGAATAAG GGCGACACGG AAATGTTGAA TACTCATACT
4321 CTTCCTTTTT CAATATTATT GAAGCATTTA TCAGGGTTAT TGTCTCATGA GCGGATACAT
4381 ATTTGAATGT ATTTAGAAAA ATAAACAAAT AGGGGTTCCG CGCACATTTC CCCGAAAAGT
4441 GCCACCTG
//
```

Strains:

Strain designation: BW25113/pIJ790

Plasmid: pIJ790 *[oriR101], [repA101*(ts)*], araBp-gam-be-exo*

Chromosome: (Δ(*araD-araB*)*567*, Δ*lacZ4787*(::*rrnB-4*), *lacIp-4000*($lacI^Q$), λ^-, *rpoS369*(Am), *rph-1*, Δ(*rhaD-rhaB*)568, *hsdR514;* Datsenko and Wanner, 2000);

this strain should be grown on rich medium containing chloramphenicol (25 µg/ml) at 30°C.

Strain designation: DH5α/pIJ773

Plasmid: pBluescript KS (+), *aac(3)IV*, *oriT* (RK2), FRT sites

chromosomal marker : see Stratagene,

this strain should be grown on rich medium containing carbenicillin (100 µg/ml) and apramycin (50 µg/ml) at 37°C.

Strain designation: DH5α/pIJ778

Plasmid: pBluescript KS (+), *aadA*, *oriT* (RK2), FRT sites

chromosomal marker : see Stratagene,

this strain should be grown on rich medium containing carbenicillin (100 µg/ml), streptomycin (50 µg/ml) and spectinomycin (50 µg/ml) at 37°C.

Strain designation: DH5α/pIJ780

Plasmid: pBluescript KS (+), *vph*, *oriT* (RK2), FRT sites

chromosomal marker : see Stratagene,

this strain should be grown on DNA (Difco Nutrient Agar), 2XYT broth or DNB (Difco Nutrient broth) containing carbenicillin (100 µg/ml) and viomycin (30 µg/ml) at 37°C.

Strain: DH5α/BT340 (Datsenko and Wanner, 2000) **can be ordered from:**

E. coli Genetic Stock Center

830 Kline Biology Tower

MCD Biology Department

266 Whitney Ave. Box 208103

Yale University

New Haven, CT 06520-81033

Tel. (0044) (203) 432-9997 Fax –6161

email: berlyn@cgsc.biology.yale.edu

webpage: http://cgsc.biology.yale.edu

Strains DH5α/pIJ779 and DH5α/pIJ781 will be send separately on request, please contact:

Plant Bioscience Limited, Dr Karin Schofield

Technology Acquisition & Technology ManagerNorwich Research Park, Colney, Norwich, NR4 7UH, UKEmail: karin@plantbioscience.com

Tel: +44 (0) 1603 456500; Fax: +44 (0) 1603 456552

www.plantbioscience.com

The CD-ROM includes:

1. Manual as Microsoft Word document and Adobe Acrobat document (X:/Protocol/)

2. Sequences of template plasmids as plain sequence files and as EMBL files (X:/Template plasmids/)
3. Sequences of the *S. coelicolor* real cosmids (X:/S_coelicolor/cosmid inserts)

Literature

Bennett, P.M., Grinsted, J. and Richmond, M.H. (1977) *Mol.Gen.Genet.*, **154**, 205-211

Cherepanov , P.P and Wackernagel, W. (1995) *Gene*, **158**, 9-14

Datsenko, K.A. and Wanner, B.L. (2000) *Proc. Natl. Acad. Sci. USA* **97**, 6640-6645

Flett, F., Mersinias, V. and Smith, C.P. (1997) *FEMS Microbiol.Lett.*, **155**, 223-229

Hobbs G, Frazer CM, Gardner DCJ, Cullum JA, Oliver SG (1989) *Appl.Microbiol.Biotechnol.*, **31**, 272-277.

Hodgson DA (1982) *J.Gen.Microbiol.*,**128**, 2417-2430.

Hopwood DA (1967) *Bacteriol.Rev.*, **31**, 373-403.

Hopwood DA, Bibb MJ, Chater KF, Kieser T, Bruton CJ, Kieser HM, Lydiate DJ, Smith CJ, Ward JM, and Schrempf H. (1985) Norwich: The John Innes Foundation

Kieser, T., Bibb, M.J., Buttner, M.J., Chater, K.F. and Hopwood, D.A. (2000) *Practical Streptomyces Genetics*: John Innes Foundation, Norwich Research Park, Colney, Norwich NR4 7UH, UK

MacNeil, D.J., Gewain, K.M., Ruby, C.L., Dezeny, G., Gibbons, P.H. and MacNeil, T. (1992) *Gene*, **111**, 61-68

Matsushima ,P., Broughton, M.C., Turner, J.R. and Baltz, R.H. (1994) *Gene*, **146**, 39-45

Redenbach, M., Kieser, H.M., Denapaite, D., Eichner, A., Cullum, J., Kinashi, H. and Hopwood, D.A. (1996) *Mol. Microbiol.*, **21**, 77-96

Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) Cold Spring Harbor Laboratory Press, Cold Spring Habor, NY.

Strauch E, Takano E, Baylis H, Bibb MJ (1991) *Mol.Microbiol.*, **5**, 289-298.