INVESTIGATION OF *STREPTOMYCES* PROMOTERS

BY: WILLIAM RICHARD BOURN

Thesis Presented for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Microbiology

UNIVERSITY OF CAPE TOWN

DECEMBER 1995

CERTIFICATION OF SUPERVISOR

In terms of paragraph GP 8 of "General Rules for the degree of Doctor of Philosophy (PhD)" I, supervisor of the candidate William Richard Bourn, certify that I approve of the incorporation in this thesis of material that has already been published or submitted for publication.

Signed by candidate

Signature Removed

Thomson
Head of Department
Department of Microbiology
University of Cape Town

AUTHOR'S FOREWORD

It is difficult to produce any document without first determining the audience for which it is intended. The work here is intended to be understandable (without frequent recourse to a library) to readers familiar with molecular biological techniques and bacterial genetics in general, but not necessarily the details of *Streptomyces* biology and bacterial development. This has resulted in a broad literature review being presented, for which I make no apology. I believe a broad information base is essential for critical analysis. I have also included speculation to stimulate thought and discussion.

With regard to the research reported here, I have included, as briefly as possible, a number of lines of work that have yielded little of value. I do this simply so others do not follow the same fruitless routes. Finally, some of the work reported here has duplicated work by other researchers. This is because their work was unpublished when my work was undertaken. I have noted this when it occurs, and report my own results as corroboration and because it assists in the understanding of why certain paths were taken.

ACKNOWLEDGEMENTS

This work would have been impossible but for the help given by many members of the Microbiology Department of UCT. I wish to thank them for their friendship and all they have taught me. Particular thanks are due to Prof. Ed Rybicki, Dr Vernon Coyne, Ant Smith, Joe Santangelo, Jono Lane, Renate Scholle and Gordon Brown. Special mention and thanks must be accorded to Brendan Babb, who wrote all the novel computer programs used here, Di James who performed some of the sequencing and also Shelly Deane for proofreading.

I would also like to express my gratitude to both my supervisors, Prof. Ralph Kirby and Prof. Jennifer Thomson. Both have shown much kindness and deserve individual mention, Ralph for his role as unfailing friend and mentor and Jennifer for her support so freely given when it was needed most. Thanks are also due to scientists from other institutions, among them Prof. Sir David Hopwood, Dr Keith Chater and Dr Mervyn Bibb, who have sent me many useful strains and plasmids.

No tribute can be complete unless it includes those who have aided me in more personal ways. Thanks are due to my parents for their help and faith, as always. Finally I would like to express my appreciation to my wife Anne, who ensured that I finished this work (instead of vice versa), and to whom I dedicate this work.

Hindsight is an exact science.

(anon.)

SUMMARY

The work described here had multiple aims: to create a promoter probe that was suitable for the isolation of developmentally regulated *Streptomyces* promoters, to isolate such promoters, to develop a computer assisted analysis system whereby potential promoter sequences could be determined and to use this in the analysis of the cloned promoters. Initially the suitability of the *Streptomyces antibioticus melC* operon for use as a reporter system in *Streptomyces* was investigated. It was established that late-expressed promoters could be identified and that it was possible to use the *melC2* gene alone for this purpose. However, it was shown that the use of both *melC1* and *melC2* resulted in a more sensitive reporter system. High copy number promoter probe vectors were constructed and tested. A low copy number promoter probe (which used the *Streptomyces penemefaciens* pSPN1 origin of replication) was also constructed. The characteristics (copy number, stability and mobility) of the probe were established. The conditions in which sporulation was induced by phosphate limitation were identified. Under such conditions late expressing, phosphate dependent promoters were isolated, using the promoter probes previously developed. The expression of these promoters was tested in *Streptomyces coelicolor bldA* mutants, and the *bldA* dependent promoters identified. These were sequenced. Computer assisted analysis of DNA sequence bias was conducted, with the intention of using bias patterns to identify potential regulatory regions. The initial approach of using the sequence bias of protein coding regions (based on the premise that regulatory sites are likely to be under represented in these regions) was unsuccessful. Further analysis in which the positional preference of sequences that were over represented in regulatory regions was conducted. Based on this the known promoters of *Streptomyces* were partially classified. The sequence bias of protein coding DNA regions was used to develop a novel method to identify the protein coding regions of *Streptomyces* DNA. The computer programs were then used to identify protein coding and potential regulatory regions.

CONTENTS

CHAPTER ONE

GENERAL INTRODUCTION AND LITERATURE REVIEW

1.1 General introduction

One of the most taxing riddles for biologists of our time concerns the question of cellular differentiation. Scientists are beginning to investigate what governs the fate of genetically identical cells, such that they appear in morphologically different forms. Questions concern how this is regulated on a genetic level and how individual cells communicate, so that multicellular organisms are formed. Considerable progress has been made in this area for both prokaryotic and eukaryotic organisms.

Among the prokaryotes, one genus that has been selected for intense study is *Streptomyces*. These bacteria display an unusual morphology and undergo a complex life cycle that involves both cellular differentiation and multicellular development.

Commercial interest plays a large part in *Streptomyces* research, as this genus produces the majority of the world's known antibiotics (Hopwood and Merrick, 1977). The production of these antibiotics is linked, in ways that are unclear, to the process of sporulation (Chater, 1984).

The advent of recombinant DNA technology has opened opportunities for the manipulation of the streptomycetes and their ability to produce antibiotics. Chater (1990) has reasoned that the positive regulatory features of antibiotic production by streptomycetes are suited for alteration to promote yield increases. This approach has proven valid, and complements the classic methods of mutation and selection which is suited to negatively regulated systems.

Furthermore, projects to create novel antibiotics have been undertaken. Initially a "shotgun" approach of combining similar *Streptomyces* antibiotic production pathways yielded unique products (Hopwood *et al*, 1985b). More recently, individual components of different *Streptomyces* polyketide antibiotic synthesis pathways have been amalgamated and shown to be functional (Sherman *et al*, 1992; Khosla *et al*, 1993). In *Saccharopolyspora erythraea*, a member of the same order as *Streptomyces*, directed mutation has been used to inactivate an enzyme activity involved at a specific stage

during the synthesis of erythromycin, thus generating a predicted novel compound (Donadio *et al*, 1991).

Research is now directed towards a more complete understanding of the regulation of *Streptomyces* antibiotic production and the coupled process of differentiation. Such knowledge will be both academically and commercially valuable.

1.2 General characteristics of the streptomycetes

1.2.1 *Streptomyces*, an introduction

The Gram-positive bacteria can be loosely grouped into those with either a low or high proportion of G+C in their genome. The latter group, to which the streptomycetes belong, are termed the actinomycetes (Chater and Hopwood, 1993).

Streptomycetes live in soil or plant litter, existing there predominantly as spores, which germinate in the intermittent presence of nutrients. Following germination, they penetrate and degrade complex organic compounds, utilising extracellular enzymes in the process (McCarthy and Williams, 1992). Streptomycetes face challenges typical of such a niche, such as changes in type and availability of their nutrient source. Other factors include variations in pH, salinity, oxygen level, water tension and temperature. Different species are known to show different tolerances to each of these (Goodfellow and Williams, 1983).

Due to the focus of research upon *Streptomyces* secondary metabolism and cellular differentiation (section 1.3) little is known about the vegetative growth stage. Research is further hampered by the commercially motivated tendency to focus upon a large variety of species and the reluctance of commercial laboratories to reveal their results. Most information available concerns *Streptomyces coelicolor*, *Streptomyces lividans* and *Streptomyces griseus*.

1.2.2 Carbon metabolism and catabolite repression

This subject has been reviewed by Chater and Hopwood (1993). The streptomycetes are able to utilise simple sugars, some of which (glucose being the best studied) repress transcription of the extracellular enzymes agarase, α-amylase and chitinase (Bibb *et al.*, 1987; Virolle and Bibb, 1988; Delic *et al.*, 1992). Glucose in the growth medium also represses the expression of the genes for galactose and glycerol utilisation (Fornwald *et al.*, 1987; Seno and Chater, 1983; Smith and Chater, 1988) and can affect both sporulation and antibiotic production (see section 1.3).

Carbon catabolite repression is a general feature of microbial metabolism and usually occurs at the level of transcription. Several different mechanisms of achieving this are known or suspected to exist in different microorganisms (reviewed by Saier, 1991).

It seems that the enzymes involved in the phosphorylation of glucose play a central role in most organisms. In *Escherichia coli* protein IIA^{glc} regulates the level of cyclic AMP, and thus ultimately the activity of the CRP transcription regulator, which controls expression of some catabolite repressed genes. Protein IIA^{glc} also regulates the levels of specific inducers of a number of genes by allosteric inhibition of permeases and catabolic enzymes. In this way, metabolic genes are repressed in the presence of glucose by the exclusion of their inducers (Saier, 1991). In *Bacillus subtilis* it appears that catabolite repression is mediated, at least in part, by a repressor protein that recognises binding sites at a number of genes but it is not known how this is regulated by glucose (reviewed by Chambliss, 1993).

The mechanism by which *Streptomyces* glucose repression is achieved is poorly understood, however the ATP dependent enzyme glucose kinase, responsible for phosphorylating glucose, appears to play a central role. Mutants of *S.coelicolor* that are able to grow in the presence of 2-deoxyglucose have been isolated and shown to be deficient

in glucose repression (Hodgson, 1982). This defect could be corrected by complementation with a cloned DNA fragment shown to carry the *glk*A gene that encodes glucose kinase (Angell *et al.*, 1992). The catabolite repression effect is not, however, dependent solely on the glucose kinase catalytic ability (Angell *et al.*, 1994). The simplest model is that the system is dependent on measuring the glucose flux through the kinase. This is, however, untenable because *glk*A dependent catabolite repression is also caused by other carbon sources, some of which are not catabolised via glucose. This observation has lead to the proposal that the presence of an inducing sugar may generate a signal that would cause some form of conformational change in the glucose kinase. The glucose kinase in this activated form might then directly control catabolite repression sensitive genes, or do so in conjunction with another protein (Kwakman and Postma, 1994). In favour of the latter model is the fact that although GlkA is homologous to a family of repressor proteins, this homology probably represents a sugar binding domain and there appear to be no DNA binding domains (Angell *et al.* 1992).

1.2.3 The general genetics of *Streptomyces*

The approximately 8-Mb *Streptomyces* chromosome is large compared to other bacteria and is known to be linear in the case of *S.lividans* and six other species, including *S.coelicolor*. In *S.lividans* telomeric structures at the chromosomal ends have been characterised although the cells remained viable when the chromosome was artificially circularised (Lin *et al.*, 1993). Over 100 loci have been mapped on the chromosome of *S.coelicolor*, showing that there is a large (roughly 2 Mb) "silent region" where few of the loci that have been mapped occur (Hopwood *et al.*, 1993). In many species the chromosome contains long regions, up to 2 Mb in *Streptomyces ambiofaciens* (Leblond *et al.*, 1991), that are not essential for growth under laboratory conditions. These can undergo spontaneous or DNA damage induced deletions causing pleiotropic phenotypic changes (Simonet *et al.*, 1992; Lin *et al.*, 1993).

In protein coding regions of the *Streptomyces* genome, the distribution of the G and C nucleotides has a bias with a periodicity of three. The percentage of occurrence of these nucleotides was shown to be approximately 70%, 50% and 90% in positions 1, 2 and 3 of the codons respectively. This has been used to determine the position of protein coding sequences (Bibb *et al.*, 1984).

Many *Streptomyces* plasmids, insertion elements and phages have been isolated and studied (reviewed by Chater and Hopwood, 1993). In general, plasmids have been found to vary in character from small circular high copy number plasmids, to extremely large low copy number linear plasmids. Most cloning vectors are based on the high copy number pIJ101 (Kendall and Cohen, 1988), the low copy number SCP2 plasmid or phage ϕC31 (Chater and Hopwood, 1993).

The *Streptomyces* are amenable to various methods of genetic manipulation. These include strain crosses, conjugation, transfection and transformation (reviewed by Chater and Hopwood, 1993). No natural competence has been demonstrated for the *Streptomyces*, and transformation is based on the ability of polyethyleneglycol to cause the uptake of DNA by *Streptomyces* protoplasts (Hopwood *et al.*, 1985a).

1.2.3.1 Transcription and promoters

The *Streptomyces* RNA polymerase appears to be typical of prokaryotic RNA polymerases, with a core (termed E) consisting of the β, β' and two α subunits, which form a complex with the σ factor (Eσ). The ability of the RNA polymerase to recognise a specific sequence of DNA as a promoter resides in the σ factor. By producing different σ factors the cell is able to direct expression from a series of different promoters. Control of σ factor expression allows for tight regulation of large numbers of genes of a particular promoter class. Typically, bacteria have a single major σ factor that directs recognition of the majority of promoters, while other minor σ factors have limited and specialised functions. This is widely used in bacteria to regulate groups of genes as diverse as those involved in

nitrogen fixation (Hirschman *et al.*, 1985), flagella production (Helmann, 1991), heat shock response (Grossman *et al.*, 1984) and differentiation (see sections 1.3.1 and 1.3.4.6).

Strohl (1992) has compiled a list of 139 *Streptomyces* promoters, for 87 genes, that have been identified on the basis of transcript mapping. It was found that approximately 31% of the genes were transcribed from multiple promoters and roughly 16% (probably an underestimation) had overlapping divergent promoters. This arrangement has been postulated to govern interdependent expression of the genes involved and is not unusual in bacteria (Collado-Vides *et al.*, 1991). At least 6 of the 139 promoters studied were found within protein coding sequences. The distance between the transcription initiation site and the start codon of the first gene was found to vary from 9 to 345 nucleotides (excluding the genes that have leaderless mRNA, see section 1.2.3.2), with the longer untranslated regions often showing secondary structure. This is unlike the short leader sequences displayed by *E.coli*.

An attempt was made to classify the *Streptomyces* promoters on the basis of sequence similarity (Strohl, 1992). Approximately 21% of the promoters were those described as *E.coli*-like *Streptomyces* promoters, on the basis of their similarity to the promoters recognised by the *E.coli* major σ factor, σ^A. Some of these have been shown to be functional in *E.coli*. This type of promoter was primarily found in conjunction with "housekeeping" genes but was also found upstream of a number of genes involved in secondary metabolism. No other separate classes were obvious among the remaining promoters. These might only be classified correctly once the *Streptomyces* σ factors, and the sequences they recognise, have been defined.

Different experimental approaches have implied the existence of a number of σ factors in streptomycetes, and combination of these data has led to the conclusion that there are at least seven different σ factors in *S.coelicolor* (Buttner, 1989).

Westpheling *et al.*, (1985) were historically the first to show that *S.coelicolor* produced σ factors that recognised different promoter classes. Two σ factors were isolated, each of which directed transcription from a different *B.subtilis* promoter. The two σ factors, termed σ^{35} and σ^{49} (molecular masses of 35 kDa and 49 kDa), allowed Eσ recognition of the *veg* and *ctc* gene promoters respectively. In *B.subtilis* the *veg* gene promoter is of the σ^{A} (principle σ factor) class, while the ctc-p promoter is of the σ^{B} class (involved in stationary phase phenomena; Boylan *et al.*, 1993).

The existence of another *S.coelicolor* σ factor was shown in work on the agarase producing *dag*A gene, which has four heterogeneous promoters, dagA-p1-4 (Buttner *et al.*, 1987). A separate transcribing activities for each of dagA-p2 dagA-p3 and dagA-p4 were distinguished. Purification of the σ factors, and *in vitro* transcription using reconstituted Eσ showed that dagA-p2 was recognised by the novel σ^{28} while dagA-p3 was probably recognised by σ^{49} (the estimated molecular mass in this work was 52 kDa). The latter conclusion was based on the copurification of dagA-p3 and ctc-p transcribing activities. Although *in vitro* transcription from dagA-p4 was not achieved, copurification of the veg-p and dagA-p4 transcribing activities indicated that the dagA-p4 promoter was transcribed using σ^{35} (Buttner *et al.*, 1988). Later work (Brown *et al.*, 1992; see below) showed that the dagA-p4 promoter was transcribed by Eσ containing a 66 kDa σ factor.

Studies on the *gal* operon have implied the existence of a further σ factor (Westpheling and Brawner, 1989). The *gal* operon is transcribed from two promoters, the upstream gal-p1 and the intercistronic gal-p2 (Fornwald *et al.*, 1987). It was shown that in *S.coelicolor* neither of these promoters is recognised by $E\sigma^{35}$ or $E\sigma^{49}$, and the transcribing activities of the two *gal* gene promoters could be separated by chromatography. This indicates that *S.coelicolor* has at least two other σ factors in addition to σ^{35} and σ^{49}. It was argued that on the basis of sequence similarity between the gal-p2 and dagA-p2, that gal-p2 is

probably recognised by $E\sigma^{28}$ (Westpheling and Brawner, 1989). The gene encoding σ^{28} (*sig*E) has been sequenced and appears to encode a σ factor of a class thought to regulate extracytoplasmic functions (Lonetto *et al.*, 1994).

In order to identify the principle σ factor genes of a variety of bacteria, Tanaka *et al.* (1988) used Southern blotting and an oligonucleotide probe consisting of a DNA sequence conserved only in the major σ factor genes. It was found that *Micrococcus*, *Pseudomonas* and *Streptomyces* strains had multiple *rpo*D homologous loci. This was also true of *Anabaena* and *Mycobacterium smegmatis* (Brahamsha and Haselkorn, 1992; Predich *et al.*, 1995). In the case of *S.coelicolor* the four homologous regions were subsequently cloned and sequenced (Shiina *et al.*, 1991; Tanka *et al.*, 1991), revealing that each of the four loci could encode a homologue of the major σ factor class. The sequence similarity was not of the type that would occur between σ factors of different classes. The loci were termed *hrd*A-D. A number of important residues are conserved in the 2.4 and 4.2 regions, (which are implicated in the specific recognition of the -35 and -10 regions of promoters) of the predicted σ^{Hrd} factors. This indicates that the σ^{Hrd} proteins recognise similar, though possibly subtly different sequences (Buttner, 1989).

Gene disruption experiments have shown that mutations in the *hrd*A, *hrd*C and *hrd*D genes, and all the double and the triple combinations thereof, give rise to viable cells. None of the mutants showed any obvious phenotypic changes in growth, antibiotic production or sporulation, however the published reports of this do not make clear the growth conditions under which this was tested. In contrast *hrd*B mutants could not be isolated, presumably because the lesion is lethal (Buttner *et al.*, 1990; Buttner and Lewis, 1992). The *hrd*B locus therefore appears to encode an essential σ factor, probably the equivalent of *E.coli* σ^{A}. Other streptomycetes also have multiple *hrd* genes, however only *hrd*B has been found in all species tested so far (Buttner, 1989).

Mapping of the *hrd* gene transcriptional start points proved that *hrd*B and *hrd*D mRNA is produced when cultures of

S.coelicolor are grown in rich liquid medium (Buttner *et al.*, 1990). Small amounts of *hrd*A RNA were also detected when the cultures were grown in liquid minimal medium (Buttner, 1989), but no transcription of *hrd*C has yet been reported. Only *hrd*B is known to produce a viable σ factor. Brown *et al.* (1992) have isolated a 66 kDa peptide which, in association with the RNA polymerase core, can direct transcription from the veg-p and dagA-p4 promoters. N-terminal amino acid sequencing indicated that this protein was the *hrd*B gene product. It is unclear whether σ^{35} (which also directs *veg* transcription) described by Westpheling *et al.*, (1985), is a different σ factor or a breakdown product of σ^{hrdB}.

Westpheling and Brawner (quoted by Buttner, 1989) have noted that the transcribing activities for the *veg* promoter and the XP55-p promoter from *S.lividans* are independent entities. This is remarkable because both of these promoters are highly similar in sequence to the *E.coli* σ^{70} type promoter consensus sequence, and differ from each other only in a single base pair in each of the -10 and -35 regions. It was argued that the two promoters are probably recognised by different homologues of the Hrd proteins.

The multiple *hrd* genes of *Streptomyces aureofaciens* have also been extensively investigated. As in *S.coelicolor*, four *hrd* genes have been found, however a *hrd*C-like homologue is absent, while the novel *hrd*E gene is present (Kormanec *et al.*, 1992). Later work proved that *hrd*A, B and D were each transcribed from tandem promoters. In surface culture the transcription was dependent upon the developmental stage. The *hrd*B gene was transcribed throughout growth, however the *hrd*D gene was active only during vegetative growth while transcription of *hrd*A was found only at the onset of sporulation. In each case both promoters of the tandem set were regulated in this manner. Surprisingly, none of the promoters of the *hrd* genes themselves appeared to be of the *E.coli* σ^{70} like class that the σ^{hrd} factors are thought to recognise (Kormanec and Fakasovsky, 1993).

Finally, it is known that a further σ factor (termed σ^{whiG}) is involved in *Streptomyces* sporulation. This is described

elsewhere (section 1.3.4.5). It has also been reported that in *S.aureofaciens* a different σ factor, encoded by *sigF*, is essential for spore maturation (Potuckova *et al*, 1995).

There have been several studies involving mutation of *Streptomyces* promoters followed by expression level tests (Forsman and Granstrom, 1992; Strohl, 1992, and references therein). In general, the mutation of *E.coli*-like promoters resulted in predictable changes in expression, given the similarity to the *E.coli* promoter consensus. Notable exceptions include gal-p1 and blaF-p. In the case of gal-p1 the -35 region appears to be different to that of *E.coli*-like promoters although the -10 region is similar, while blaF-p has no similarity to *E.coli*-like promoters.

The fact that a number of genes are transcribed from multiple promoters recognised by different σ factors indicates that σ factors may play a much larger and more central role in gene regulation in *Streptomyces* than in organisms such as *E.coli*.

1.2.3.2 Translation

The ribosome binding sites of a number of *Streptomyces* genes have been analysed by Strohl (1992). From this work it appears that the consensus sequence for *Streptomyces* ribosome binding is (a/g)GGAGG, however a wide range of sites are functional. This is in conflict with the observation of McLauglin *et al.* (1981) that Gram-positive bacteria, unlike Gram-negative, require extended and conserved sequences for ribosome binding.

The *Streptomyces*, and other actinomycetes, are also unusual in that there are reports of a number of genes in which translation and transcription are initiated at the same or adjacent nucleotides. This means there is no leader sequence or ribosome binding site, and searches for binding sites both upstream and downstream of the start codon have revealed no obvious candidates in most cases (reviewed by Strohl, 1992). The majority of the genes displaying this feature are involved in secondary metabolism.

1.3 Differentiation of the streptomycetes

1.3.1 Differentiation of the prokaryotes

Research on development has focused on several eubacteria, besides the streptomycetes and the subject has been extensively reviewed. The organisms include *Bacillus subtilis* (Errington, 1993), *Caulobacter crescentus* (Ely and Shapiro, 1984; Newton and Ohta, 1990), the myxobacteria, primarily *Myxococcus xanthus* (Shimkets, 1990; Kim and Kaiser, 1992) and the cyanobacterium *Anabaena* (Dworkin, 1985, Haselkorn *et al.*, 1986). The details of their development are unlikely to have much in common as each has evolved separately to exist in a different niche. However, bacteria have common aspects of physiology and face common dilemmas, so solutions similar in form may have evolved and indeed, are observed. *C.crecentus* and *Anabaena* are very different to the streptomycetes as they are both aquatic organisms with specific needs. *B.subtilis*, however, is a Gram-positive soil dwelling bacterium. Like the streptomycetes, it sporulates in response to starvation although the structure and development of the *Bacillus* endospore is very different to the *Streptomyces* arthrospore. The myxobacteria also have features in common with the streptomycetes as they too live in organic debris and differentiate in response to starvation. In this case multicellular structures develop and diffusible factors are crucial to the process. The fact that *E.coli* displays a highly organised response to starvation has only recently been recognised (reviewed by Hengge-Aronis, 1993) and *Streptomyces* development may show similar features to *E.coli* stationary phase survival.

With development in response to environmental conditions, the process can be considered in two parts. The first stage is the initiation of the process and the second is the differentiation itself. Starvation is a common trigger for differentiation. Little is known about how the bacterium determines that starvation is imminent. Probably the overall state of cell metabolism is determined via global metabolic

regulatory systems, such as the stringent response (see section 1.3.3). In *B.subtilis* it has been shown that sporulation can be induced by the stringent response, however, differentiation is not dependent on this mechanism as relaxed (stringent response negative) mutants are still able to sporulate (Lopez *et al.*, 1981). A decrease in the internal level of GTP appears to be crucial for sporulation induction (Lopez *et al.*, 1979).

The use of diffusible factors is a common theme in developmental regulation. For the myxobacteria this may fulfil the requirements for orchestrated cell movement, and at least four different intercellular signaling systems operate in *M.xanthus* (Kim and Kaiser, 1992). *Anabaena* spp may also use diffusible factors to regulate heterocyst numbers and spacing (Black and Wolk, 1994). *B.subtilis* however, undergoes no multicellular development, so the diffusible factor EDF-1 (Grossman and Losick, 1988) must be used to measure cell density.

It is a fundamental concept that sporulation occurs in response to a number of different stimuli. Thus starvation implies the limitation of nitrogen, carbon, phosphate and energy sources. Furthermore, diffusible factor concentration, pH, temperature, toxin presence, desiccation etc. may play a role.

It is a logical assumption that the many factors involved in induction of development are measured and balanced in a process of information integration (Errington, 1993). Thereafter the point of commitment is reached and a resource expensive course of action taken (Dworkin, 1985). It must be rigidly controlled, probably via a small number of proteins present in the cells prior to starvation. These are likely to be acted upon by sensor proteins, which sense the parameters that trigger differentiation. An example of this type of regulation is found in *B.subtilis* sporulation, where the crucial regulator, Spo0A is activated by a phosphorylation cascade initiated by sensor kinases (reviewed by Errington, 1993). Signal transduction systems of this nature are widespread throughout bacterial species

(reviewed by Parkinson, 1993; Bourret *et al.*, 1991) and are expected to play a major role in bacterial differentiation.

Once the bacteria is committed to development, a complex cascade of gene expression must ensue, such that each step in the process is complete before the next is initiated. There are examples of expression regulation at every genetic level.

The replication of the chromosome is known to be essential for the development of both *C.crescentus* and *Anabaena* (Newton and Ohta, 1990; Adams and Carr 1989). Regulation has also been shown to exist at the level of the DNA itself. DNA rearrangements enable the translation of the *B.subtilis* mothercell-specific σ^K and the *Anabaena* heterocyst-specific *nif* gene products (reviewed by Haselkorn, 1992).

The onset of transcription is the last stage at which control can be exerted without the commitment of metabolic resources. Positive regulation has some advantages over negative regulation, as regulatory proteins need be present only when the gene they control is active. This is one of the advantages incurred by bacteria when the controlled production of RNA polymerases that recognise different classes of promoters is used as a regulatory device. In *B.subtilis* there are five sporulation specific σ factors (Errington, 1993), and a developmentally regulated σ factor is also involved in *M.xanthus* differentiation (Apelian and Inouye, 1990).

Regulation of development also occurs at the level of protein activation, either by post translational processing or by blocking mechanisms (for example the *B.subtilis* σ^F, σ^E, σ^G and σ^K; Losick and Stragier, 1992).

Developmental regulation at the level of translation may occur in the case of the streptomycetes (see section 1.3.4.2).

1.3.2 Morphological development of the streptomycetes

The morphological changes of streptomycetes growing on solid media have been well documented (reviewed by Hodgson, 1992). For most species sporulation has only been observed on solid media, however there are several reports of its occurrence in liquid media in the case of specific species and specific nutrient conditions (Kendrick and Ensign, 1983; Koepsel and Ensign, 1984; Daza *et al.*, 1989; Glazebrook *et al.*, 1990).

Spore germination in *Streptomyces antibioticus* has been described as having three stages. The first is dependent on divalent cations and the formation of ATP using internal energy reserves, and results in spores becoming phase dark. Following this, the spore swells, the spore wall becomes stratified, and protein and RNA synthesis begin. This stage requires an external carbon source. The final stage is nitrogen source dependent and involves the production of DNA, followed by the germ tube pushing through the broken cell wall (Hardisson *et al.*, 1978).

The hyphae grow only at the tip (Brana *et al.*, 1982; Gray *et al.*, 1990) but branching occurs, by which mechanism semi-exponential growth is achieved. The hyphae are multinucleate with few septa. Microscopic and autoradiograpic studies on liquid cultures of *Streptomyces granaticolor* have shown that chromosome replication occurs in all parts of the mycelium, but in older subapical parts the chromosomes do not segregate, but form multi-chromosome nucleiods. Individual chromosomes are found only in the apical compartment and in new branches. During branch formation the nucleoid segregates and supplies two individual chromosomes to the new branch (Kretschmer and Kummer, 1987; Kretschmer, 1987). Extracellular enzymes break down the growth substrate which is quickly colonised by the mycelial mat, or vegetative mycelium (Chater, 1984). The multicellular nature of the colony probably allows for the more efficient action of pooled enzymes as described by Dworkin (1985) for the myxobacteria. In the older parts of the colony storage

compounds such as glycogen, polyphosphates and lipids collect (Brana *et al.*, 1986; Chater, 1993).

The first steps of differentiation occur in response to starvation and, probably, other factors (see section 1.3.3). The uppermost part of the vegetative mycelium then produces a number of vertical aerial hyphae. Chater (1989b) has proposed that the osmotic pressure needed to drive growth at this stage is derived from the breakdown of macromolecules, such as glycogen, which are stored by the vegetative mycelium and degraded during sporulation. The aerial hyphae develop at the expense of the substrate mycelium, which degrades and is cannibalised (Wildermuth, 1970; Mendez *et al.*, 1985).

Granozzi *et al.*, (1990) have used wet weight measurements to demonstrate that there is a pause in growth prior to the development of aerial hyphae. This was accompanied by a transient decrease in the incorporation of radiolabeled precursors, supplied in the medium, into DNA, RNA and protein. These authors have argued that this reflects a period of growth cessation. However, it is possible that the apparent decrease of growth and macromolecular synthesis could merely reflect that the aerial mycelium is growing at the expense of the vegetative mycelium. This assumes that the breakdown products are channeled to the aerial mycelium, and the vegetative mycelium has ceased to take up nutrients.

At this point secondary metabolites, including antibiotics, are produced. Chater (1984) has proposed that these antibiotics prevent other microorganisms from colonising the degrading substrate mycelium.

The densely packed aerial hyphae are sheathed in a fibrous layer that is absent from the vegetative hyphae (Williams *et al.*, 1972). As the colony ages, the aerial hyphae are divided into long cells which then become coiled. These are then further divided into smaller compartments by specialised sporulation septae (Wildermuth and Hopwood, 1970) of which there are at least three species dependent types (Hardisson and Manzanal, 1976; Manzanal and Hardisson, 1978). Each compartment develops into a mature spore by

rounding, followed by thickening of the wall from inside the old hyphal wall (Wildermuth and Hopwood, 1970; Manzanal and Hardisson, 1978). Glycogen, which accumulates in the sporulating hyphae is absent from the spores, but other storage compounds such as trehalose and polyphosphate are present (Brana *et al.*, 1986). McBride and Ensign (1987) have shown that the trehalose reserves are rapidly depleted upon *S.griseus* spore germination.

The spore, which was white in colour, finally becomes a dark grey. The newly formed spores, some of them already germinating, are soon overgrown by more aerial hyphae, which sporulate in turn. Several rounds of this occur, resulting in a mass of chains of spores, each chain being loosely held within a fibrous sheath (Wildermuth, 1970). The sheath is very hydrophobic and thought to enable dispersal of the spores on the surface of water droplets (Williams *et al.*, 1972).

1.3.3 Initiation of development

The question of what signals trigger the onset of sporulation and antibiotic production is unresolved. It has been suggested that the response is mediated by a starvation induced increase of certain internal nucleotides, as for *B.subtilis* (section 1.3.1). In *E.coli*, the stringent response describes a general regulatory mechanism which is induced by amino acid starvation (reviewed by Cashel and Rudd, 1987). It is controlled by the ribosome associated *rel*A gene product, which generates ppGpp when an uncharged tRNA binds to the A site of the ribosome. The presence of ppGpp causes complex changes in the overall gene expression, typically causing decreased ribosomal gene expression and increased amino acid synthesis. The stringent response is also thought to play a role in growth rate control. These characteristics ideally suit the stringent response for a role in the regulation of differentiation.

Streptomycetes also display the stringent response (Bascaran *et al.*, 1991; Ochi, 1987b; Reisenberg *et al.*, 1984; Strauch *et al.*, 1991). Ochi (1986b) has demonstrated that formycin

synthesis in a formycin producing *Streptomyces* strain was induced by the stringent response. Stringent response negative mutants (tentatively identified as *relC*) did not produce formycin under those amino acid starvation conditions used. Similar results were obtained with a streptomycin producing strain of *S.griseus* (Ochi, 1987c). In contrast, Strauch *et al.* (1991) showed that in *S.coelicolor*, while transcription of the *act*III gene (which is involved in production of the antibiotic actinorhodin) was preceded by an increase of intracellular ppGpp under certain growth regimes, in other cases it was not. Similarly, no obligatory relationship between ppGpp level and cephalosporin biosynthesis was found in *Streptomyces clavuligerus* (Bascaran *et al.*, 1991). Thus while it is possible that ppGpp levels play a role in antibiotic synthesis induction, other factors may be involved and differences between species could occur.

Ochi (1986a, 1987a,b) has shown that in five strains of *Streptomyces*, a decrease in the internal level of GTP is sufficient to initiate sporulation. Decoyinine, an inhibitor of GTP synthesis, was added to the growth media and shown to cause sporulation under conditions in which it was normally repressed. However, this treatment did not result in the production of antibiotics by the strains tested. Furthermore, although relaxed mutants sporulated weakly, this defect could be overcome by the addition of decoyinine. This implies that the signal that initiates sporulation is not an increase in ppGpp but a drop in GTP level. A decrease of GTP always accompanies the onset of the stringent response, but might also occur independently of it (Ochi, 1987a,b; Ochi 1986a; Riesenberg *et al.* 1984; Strauch *et al.* 1991).

1.3.4 The genetics of *Streptomyces* development

For detailed reviews of the genetics of *Streptomyces* development the reader is referred to Chater (1989a), Chater (1989b), Chater (1993) and Hodgson (1992).

Apart from those concerned solely with antibiotic production, over 20 loci that are involved in the development of *S.coelicolor* have been identified (Chater, 1993). Mutations in several of these loci give rise to two distinct phenotypes, so the mutants are termed either bald (*bld* genes) or white (*whi* genes).

1.3.4.1 The *bld* developmental mutants

The *bld* mutants derive their name from the appearance of the mutant colonies, which produce either no aerial hyphae or hyphae that are ill formed and not vertical to the colony. Consequently the colonies are shiny, and produce no spores. The phenotype of these mutants is usually conditional, as sporulation can be induced by changes in the growth medium, primarily the carbon source, or by growth near another strain. The latter effect is presumably due to diffusible factors (see section 1.3.5). A summary of the characteristics of the *bld* mutants is given in Table 1.1.

Little is known about the functions of the *bld* genes so their role can only be speculated upon. It is possible that they act prior to starvation in a sensory capacity analogous to the *spo*0 genes of *B.subtilis*. The fact that late sporulation genes are expressed in *bld* mutants (section 1.3.4.6) argues against this. Alternatively they may play a part in the orchestration of multicellular development. Whatever their function, it is puzzling that while *bld* mutants that are deficient in antibiotic production are common (Table 1.1), *bld* mutants that display normal antibiotic production are rare. Chater (1993) has suggested a number of possible reasons for this: such genes may be functionally redundant, or required during vegetative growth so that mutations are lethal, or antibiotic production may require a positive feedback signal from aerial mycelium formation. To this can be added the possibility that such

Table 1.1 The characteristics of *S.coelicolor bld* mutants

bld gene[a]	Colony morphology on glucose MM[b]	Antibiotic production on glucose MM[b]	Spore formation on mannitol	Other comments	Reference[c]
A	Wrinkled soft fragmenting Prostrate aerial hyphae	Only undecylprodigeosin on low phosphate	Yes	A tRNA gene (section 1.3.4.2)	4
B	Smooth hard non-fragmenting No aerial hyphae	Only on complex media	No	In *S.coelicolor* only. Two loci	2, 4
C	Smooth non-fragmenting No aerial hyphae	No CDA production, weak pigment production	Yes	Single mutant isolated restored by diffusible factor	4
D	Soft wrinkled fragmenting Prostrate aerial hyphae	Impaired	Yes	Single mutant isolated	4
E	Smooth sculpted	Undecylprodigeosin production	Yes	Agarase overproduction possibly same as *bldF*	3
F	Smooth sculpted	Undecylprodigeosin production	Yes	Agarase overproduction possibly same as *bldE*	5
G	Smooth soft fragmenting Sparse aerial hyphae	Impaired	Yes		1
H	Smooth hard fragmenting	Impaired	Yes	Antibiotic production restored on mannitol	1
I	Smooth hard non-fragmenting	Impaired	No	Maps near *bldB* but fails to complement	2
221	?[d]	Impaired	Yes		6
17	?[d]	Impaired	?[d]		6

[a]Other less well characterised *bld* loci are described by Chater (1993); Schauer *et al.* (1991) and Sohaskey *et al.* (1992).

[b]See section 1.3.6.

[c]References, 1, Champness *et al.* (1988); 2, Harasym *et al.* (1990); 3, Hodgson (1992); 4, Merrick (1976); 5, Puglia and Cappelletti (1984); 6, Willey *et al.* (1991). See also Chater (1984); Chater (1989); Chater (1993); Hodgson, (1992).

[d]Unclear from published reports.

genes are involved in negative regulatory steps so that mutants would sporulate constitutively.

1.3.4.2 The role and regulation of *bldA*

The *bldA* mutants of *S.coelicolor* produce aberrant aerial hyphae which undergo frequent septation and fragmentation. The cell wall of these hyphae appears to be deficient in galactose (Chater, 1984). The aerial hyphae fail to develop spores, however this asporogenous phenotype is only displayed on complex media or on minimal media containing glucose, mannose or cellobiose. *BldA* mutants will sporulate normally on minimal medium when either galactose, arabinose, mannitol, glycerol or maltose is supplied as the carbon source. They will not sporulate if a combination of glucose and mannitol is used. These mutants also lack the ability to synthesise antibiotics, and this defect is not corrected by changes in the carbon source (Merrick, 1976). Undecylprodigiosin synthesis is, however, restored on low phosphate medium (Guthrie and Chater, 1990).

The *bldA* locus was originally identified and cloned on the basis of its ability to complement mutants (Piret and Chater, 1985). Subsequently, sequencing of wild type and mutant *bldA* loci revealed that the lesions causing the bald phenotype were located in a tRNA gene with a UAA anticodon. The tRNA thus recognises the leucine UUA codon, which occurs extremely rarely in the protein coding sequences of streptomycetes (Lawlor *et al.* 1987).

These findings prompted questioning of whether the *bldA* gene product was the only tRNA that could recognise the UUA codon. Leskiw *et al.* (1991b) have monitored the expression levels of five genes that contain TTA codons. The results were essentially the same on complex medium and minimal medium with mannitol as the carbon source. Three of the genes tested were *ampC*, *lacZ* and *carB* (encoding β-lactamase from *E.coli*, β-galactosidase from *E.coli* and an antibiotic resistance, ribosome methylating enzyme from *Streptomyces thermotolerans*, respectively). In each of these cases, expression in a *bldA* mutant was abolished. Expression was

restored when the TTA codons in *carB* were converted to CTC codons. For two other genes, *aad* and *hyg* (encoding the antibiotic resistance enzymes spectinomycin adenyltransferase from an R-plasmid and hygromycin phosphotransferase from *Streptomyces hygroscopicus*, respectively) resistance to the test antibiotic was recorded in *bldA* mutants. The authors postulate that this was due to translation of the UUA codon by a noncognate tRNA. They further argue that the low level of expression recorded for *hyg* would be expected, and sufficient to confer antibiotic resistance, due to the efficiency of the resistance mechanisms.

It is interesting to note that Leskiw *et al.* (1991b), using an indicator reaction, could only detect expression of the *ampC* gene in *bldA*$^+$ strains after two days. However, when the actual antibiotic resistance due to cloned, TTA codon carrying resistance genes was tested, it was found that *bldA*$^+$ strains grew in the presence of the antibiotic, which indicated that the necessary tRNA was present even at the earliest stages of growth. This can be explained if the antibiotic itself was inducing *bldA* expression, possibly indirectly due to the stress caused by its action.

In direct tests of *bldA* gene expression, RNA from a *bldA*$^+$ strain of *S.coelicolor*, grown on the surface of the complex medium R2YE, was isolated. Subsequent dot blotting and strand specific probing showed that the tRNA was expressed at a high level only after 36 hours of growth, the time at which aerial hyphae first appear (Lawlor *et al.*, 1987).

Promoter probing and S1 mapping experiments on *S.coelicolor*, using cultures grown in either the rich YEME liquid medium or on the rich solid R2YE medium were later performed (Leskiw *et al.*, 1993). This showed the existence of a promoter capable of directing the transcription of RNA that is antisense to the 5' end of the tRNA. The tRNA itself was shown to be expressed from a single promoter and could be detected early during growth. However, at this stage, these transcripts appeared to be in the form of immature tRNA, while the mature, processed form of the tRNA was only abundant after 24 hours. In contrast, Gramajo *et al.* (1993)

have shown that the mature tRNA was abundant in young cultures grown in liquid minimal medium supplemented with casamino acids, with glucose as the carbon source.

The theory that the tRNA plays a regulatory role in development has yet to be proved. The alternative view is that the TTA codon is absent in crucial vegetatively expressed genes and present in developmental genes, where it has no regulatory function. However, the TTA codon appears to be found mainly in those genes involved early in the production of secondary metabolites and antibiotics. These are summarized in Table 1.2. The positioning of the TTA codons within these genes is also suggestive. For the most part they are found within the first codons of the coding sequence. Experimental work has so far failed to resolve the question.

The *act*II-ORF4 gene product, which has a TTA encoded leucine as its fifth amino acid, is a transcriptional regulator for the actinorhodin biosynthetic genes (see section 1.3.6, Table 1.2). Thus it is possible that the failure of *bldA* mutants to synthesise this actinorhodin is solely due to their inability to produce the activator and not because other regulatory steps have been impaired. Fernandez-Moreno *et al.* (1991) have investigated whether this is the case. The TTA codon of *act*II-ORF4 was replaced with another leucine codon, TTG, and the production of actinorhodin in a *bldA* mutant was monitored. It was found that the replacement of the TTA codon caused actinorhodin production to be fully restored. This strongly implies that the TTA codon in *act*II-ORF4 is the only target of the *bldA* gene product during the induction stages of actinorhodin biosynthesis. It does not, however, prove that the regulation depends on this feature.

The work of Gramajo *et al.* (1993) has indicated that actinorhodin biosynthesis is regulated at the level of transcription of the *act*II-ORF4 gene. In this study, the transcription of the *act*II-ORF4 gene was monitored during growth in glucose minimal medium supplemented with amino acids. Actinorhodin biosynthesis and transcription of actinorhodin biosynthetic genes closely followed the pattern of *act*II-ORF4 transcription. It did not correlate with the

Table 1.2 *Streptomyces* genes that contain TTA codons[a]

Species	Gene	Gene function	TTA codons Position
S.coelicolor	*actII-ORF4*	Actinorhodin biosynthesis activator	5/256
S.coelicolor	*actII-ORF2*	Actinorhodin export	19/579
S.coelicolor	*orf1590*	Unknown, homologue of *S.griseus* gene	206/473
S.coelicolor	unnamed	Transposase for IS*466*	122/514
S.griseus	*aphE*	Streptomycin resistance	47/298
S.griseus	*strR*	Streptomycin biosynthesis regulation	30/351
S.griseus	*orf1590*	Unknown, mutants have *bldA* phenotype	263/530
S.griseus	*suaC*	Cytochrome P450	26/383
S.glaucescens	*sph*	Streptomycin resistance	39/308
S.hygroscopicus	*hyg*	hygromycin resistance	7/333
S.hygroscopicus	*brpA*	Bialaphos biosynthesis regulation	250/257
S.thermotolerans	*carB*	Carbomycin resistance	95 & 112/300
S.nigrescens	SMPI	Metalloproteinase inhibitor	7/131

[a]Information from: Leskiw *et al*, 1991a (and references therein).

appearance of the mature *bldA* product, although it was shown that the *bldA* tRNA was essential for expression.

Clearly, the question of the role of the *bldA* gene product is more complex than at first supposed. The *bldA* gene must be expressed when the genes that contain a TTA codon are expressed, and so its expression may be regulated by the same mechanisms that regulate developmental genes. However, it is possible that *bldA* expression is regulated in a different manner, and as such the tRNA it encodes would then act as a regulator which would only allow expression of certain pathways under conditions which allow *bldA* expression. This would give the bacteria a greater degree of flexibility in its potential responses to environmental conditions. The discussion is clouded by several factors. Firstly, as development is dependent upon environmental conditions, the confusing difference in results obtained by different laboratories using different growth media is not unexpected. Furthermore, although development is thought to be a response to starvation, little is known about the growth limiting factors in any of these media. The situation is further confused by the consideration that TTA codons may act as regulatory sites in some cases and not others. Any gene which is only expressed at a time when *bldA* is expressed would experience no selective pressure for the elimination of TTA codons. Finally, in sporulating colonies, different parts of the colony may have different regulatory features which are difficult to distinguish.

Whatever the case, mutations in *bldA* are highly pleiotropic, and a potential target for the action of the *bldA* gene on morphological development has been identified in *S.griseus*. A class of mutants that display the characteristics of *bldA* mutants, but which are not restored to a wild type phenotype by complementation with the cloned *bldA* locus, have been isolated. However another gene, termed *orf*1590, which is able to complement this class of mutants, has been identified and sequenced. It was found that *orf*1590 contains a TTA codon and a potential DNA-binding domain. The locus has unusual features, in that the gene is transcribed from two promoters with the downstream transcription start site

falling within the open reading frame. Translation of the shorter transcript would generate a shorter polypeptide in the same reading frame. This polypeptide would lack the potential DNA binding domain. The relative amounts of each transcript have been shown to change as the organism enters the first stages of sporulation. A homologous gene has been isolated from *S.coelicolor*, and this too carries a TTA codon (Babcock and Kendrick, 1990; McCue *et al.*, 1992).

There has recently been evidence that the streptomycetes are not the only bacteria in which a tRNA might regulate a major aspect of the life cycle. Sauer and Durre (1992) have shown that in the anaerobic, Gram-positive, endospore forming *Clostridium acetobutylicum*, a mutation in a gene for a threonine tRNA prevents solvent production. To date, in *C.acetobutylicum*, codons of this type have only been found in genes expressed at the onset of stationary phase, when solvent production is initiated.

For further information on the *bld*A locus the reader is referred to Leskiw *et al.*, 1991a.

1.3.4.3 The *whi* developmental mutants

Whi mutants appear to be blocked at a later stage of development than *bld* mutants, and they form white fluffy colonies, unlike the matte grey wild type colonies. In this case, sporulation is initiated but not completed and the spores never fully mature. The phenotypes of these mutants are described in Table 1.3, which summarises the observations of Hopwood *et al.*(1970), McVittie (1974), Chater (1972) and Chater (1989). Eight of the *whi* loci have been mapped on the chromosome (Chater, 1993).

On the basis of its ability to complement mutants, the *whi*G gene of *S.coelicolor* has been cloned. It was shown that the presence of this gene at a high copy number caused vigorous and early sporulation and the appearance of unusual branched spore chains. Sporulation in liquid culture and below the agar surface was also noted, as was a considerable decrease in pigment production (Mendez and Chater, 1987; Chater *et al.*, 1989).

Table 1.3 The *whi* genes and their mutant phenotypes

Gene	Phenotype of mutants (references, section 1.3.4.3)
whiA	Coiled aerial hyphae, sporulation septa absent
whiB	Coiled aerial hyphae, sporulation septa absent
whiC	Straight aerial hyphae, weak sporulation
whiD	Thin walled spores formed
whiE	Spores normal but unpigmented
whiF	Rod shaped spores
whiG	Straight aerial hyphae, sporulation septa absent
whiH	Loosely coiled, fragmenting aerial hyphae
whiI	Tightly coiled, fragmenting aerial hyphae

Table 1.4 *Whi* gene promoter activity in developmental mutants[a]

	Mutant background					
Promoters	*bldA*[b]	*whiA*	*whiB*	*whiG*	*whiH*	*whiI*
whiB P1	+	+	+	+	+	
whiB P2	+	+	+	+	+	
whiE P1	+	-	-	+	-	+
whiE P2	+	-	-	+	-	+
whiG	+	-	-	+	-	+
P_{TH4}	+	+	+	-	-	+
P_{TH270}	+	+	+	-	-	+

+, Normal transcription; -, Weak or no transcription.

[a]Summarised from: Chater, 1993; Soliveri *et al.*, 1992; Tan and Chater, 1993.

[b]Under aerial mycelium inhibiting conditions.

Subsequent sequence analysis revealed that the *whiG* gene encodes a σ factor, termed σ^{whiG} (Chater *et al.*, 1989). This σ factor shows strong sequence similarity with the σ factors from *Bacillus subtilis* (σ^{D}) and *Salmonella typhimurium* (σ^{F}) that direct transcription of genes involved in chemotaxis and flagellar synthesis (reviewed by Helmann, 1991). The similarity is readily apparent in the two regions that determine recognition of the promoter -10 and -35 sequences (Chater, 1989; Chater *et al.*, 1989). The promoters recognised by each of these sigma factors, including those from *Streptomyces coelicolor* are to some degree similar in sequence (Chater 1993; see section 1.3.4.5).

Homologues of the *whiG* gene have been detected in a number of streptomycetes but appear to be absent in actinomycetes that are not closely related to that genus, and which do not form a sporulating aerial mycelium (Chater *et al.*, 1989; Blanco *et al.*, 1993; Soliveri *et al.*, 1993).

In contrast, hybridisation studies have shown that *whiB* homologues occur in a number of diverse actinomycetes and all streptomycetes tested (Blanco *et al.*, 1993; Soliveri *et al.*, 1993). The *whiB* gene has been isolated, sequenced and shown to encode a small highly charged protein (Davis and Chater, 1992; Kormanec and Homerova, 1993). The function of the gene product is unknown, however, it has features common to various transcription factors.

The *whiE* locus has also been sequenced and consists of a group of seven open reading frames. On the basis of sequence similarity six of these have been identified as encoding homologues of the enzymatic components of the polyketide antibiotic biosynthesis pathway (Davis and Chater, 1990).

Recently Potuckova *et al.* (1995) have isolated a σ factor encoding gene (*sigF*) from *S.coelicolor*. Disruption of this gene resulted in a *whi* mutant phenotype and small spores that were deformed and sensitive to detergents.

1.3.4.4 The spore associated proteins

To identify genetic loci involved in *S.coelicolor* spore formation, spore associated proteins have been isolated, and partially sequenced. The structural genes that encode most of them have been cloned. The proteins were termed SapA,B,C,D and E (spore associated proteins). Of these, the SapC,D and E encoding loci were found on the SCP1 plasmid, the loss of which has no effect on development (Guijarro *et al.*, 1988; Willey *et al.*, 1991).

The gene encoding SapA is chromosomally located, and its function is unknown. Transcript mapping indicated that the *sap*A gene was transcribed in older cultures grown on the surface of rich medium, coincident with the first appearance of aerial hyphae. However, transcription was not limited to the aerial hyphae and was recorded in liquid medium. Furthermore, *sap*A transcription was not abolished in several developmental mutants, but was severely reduced in *bld*C, *bld*H and *whi*H mutant strains. Gene fusion with the *lux*AB reporter system indicated that expression was both temporally and spatially controlled, as light emission from the reporter gene was greatest at the edge of the colonies (Guijarro *et al.*, 1988). It should be noted that the results obtained using the *lux*AB fusion system are questionable, as the indicator operon contains several TTA codons (see section 1.4). Unpublished work by Im and Schauer (quoted by Chater, 1993) indicates that transcription and regulation of the *sap*A gene does not require DNA sequence upstream of the -8 position.

One of the most fascinating questions concerning bacterial differentiation concerns the molecular basis of spatial organisation. In *S.coelicolor* this may be mediated, at least in part, by SapB. SapB is a small 18 residue peptide, which is modified by a moiety bearing a vicinal hydroxyl group. Immunoassays have shown that on rich solid medium SapB is produced in large amounts at the time aerial hyphae first appear, and it is found both in the medium surrounding the colony and on the surfaces of the aerial hyphae and spores. On mannitol minimal medium however, SapB is undetectable in sporulating colonies (Willey *et al.*, 1991). Synthesis of

SapB is not abolished in the presence of chloramphenicol, indicating that it may be formed by a nonribosomal mechanism, as is the case with certain peptide antibiotics (Willey *et al.*, 1993). An alternative view is that SapB could appear in the form of inactive precursors which are later processed.

All the *bld* mutants tested were deficient in their ability to produce this peptide, however *whi* mutants were not. Application of exogenous SapB directly onto *bld* mutants resulted in a transient restoration of morphological development. These observations have led to the proposal that SapB is a morphogenic peptide which coats the hyphae that are destined to form the aerial mycelium. There it may act as a scaffolding agent or allow hyphae to break the surface tension of the air-water interface (Willey *et al.*, 1991, Willey *et al.*, 1993).

1.3.4.5 The *whiG* dependent promoters

It has been noted that introduction of a σ^D type *B.subtilis* promoter at a high copy number into *S.coelicolor* causes pale colonies and weak sporulation (Chater *et al.*, 1989). It was concluded that this was due to the titration of the sigma factor encoded by *whiG*, and subsequently two small DNA fragments from *S.coelicolor* that caused a similar effect were isolated. These proved to each carry a promoter capable of directing transcription of a *xylE* reporter gene. Transcript mapping demonstrated that transcription was abolished in *whiG* mutants, and was temporally regulated such that expression occurred when aerial mycelium first appeared. The promoters were termed P_{TH4} and P_{TH270}. They both appear to fall within protein coding regions, the stop codon of which overlaps with a potential start codon of the gene under the control of the σ^{WhiG} promoter. In the case of P_{TH270} translation using mRNA expressed from this promoter would require that another, downstream, start codon be utilised. The overlap of reading frames suggests that some kind of translational coupling exists and that genes under the control of P_{TH4} and P_{TH270} are also expressed as part of

a polycistronic operon under the control of another promoter (Tan and Chater, 1993).

The promoters themselves are similar to the consensus sequence of promoters utilised by σ^D of *B.subtilis* and σ^F of *S.typhimurium*. They both show a similarity of 2/4 in the -35 region and 6/8 in the -10 region. The genes that these promoters control are chromosomally located and their functions are unknown (Tan and Chater, 1993).

1.3.4.6 The *whi* network

Chater (1975) has examined a series of double *whi* mutants and originally concluded that there was a simple epistatic cascade of the form *whiG-whiH-whiA/whiB-whiI* reflected in the phenotype of the mutants. However, microscopy of wild type aerial hyphae has shown that the mutant phenotypes are unlike those that would be achieved by a cessation of development at any particular point (Chater, 1993).

The timing of expression, control and interaction of the *whi* genes is only now being unraveled. Promoter-probes and transcript mapping have been used to investigate the promoter positions and induction time. The *whiB* gene was found to be expressed from two promoters, with the upstream, P1, promoter being weakly and constitutively expressed, and the downstream, P2, promoter being induced upon the appearance of aerial mycelium (Soliveri, *et al.*, 1992). Both the *whiG* gene and the *whiE* gene cluster were also shown to be induced at the first appearance of aerial mycelium; in both cases only one transcriptional start point was detected. There is evidence of a second promoter within the *whiE* cluster and this has not been analysed (Chater, 1993).

Because sporulation is not synchronous it is impossible to determine the relative timing of the expression of the *whi* genes, however, expression interdependence can be determined by monitoring transcription in mutant strains. The results of these types of experiment are shown in Table 1.4. Several observations are pertinent. Although the *whiG* phenotype was epistatic to all the other *whi* mutant phenotypes, transcription of *whiG* is abolished in *whiA,B,* and H mutants.

However, the *whiG* dependent promoters P_{TH4} and P_{TH270} do not depend on the *whiA,B,* and H gene products for their expression. These two observations are incompatible with a simple cascade of gene expression. The wide range of *whi* genes that are not transcribed in *whiA,B,* and I mutants suggests that the products of these genes may function as transcriptional activators. This has some support from the observation that WhiB has features similar to transcriptional activators (Davis and Chater, 1992). Furthermore, the potential regulatory regions upstream of the regulated *whiE* cluster and the *whiG* gene are similar, and conserved in homologues from other species (Chater 1993).

Although Chater (1993) has proposed a tentative model to explain the interdependence pattern, it is probable that too little is known about the system for accurate interpretation of the results to be made. Furthermore, the author himself has questioned the validity of this model, and has argued that the use of high copy number vectors in the transcription studies has led to spurious results being obtained (K. Chater, personal communication).

1.3.5 Diffusible factors and *Streptomyces* development

The streptomycetes produce a wide range of secondary metabolites, including antibiotics (section 1.3.6) and others of unknown function. Certain of these secondary metabolites function as diffusible autoregulators that are essential for the production of antibiotics and sporulation. These include the butyrolactone type regulators like A-factor (see below) and its homologues which are found in numerous species. Other simple autoregulators are pamamycin, a macrolide type antimicrobial agent from *Streptomyces alboniger* and a pigment from *Streptomyces venezuelae*. A diffusible autoregulatory protein, factor C, from *Streptomyces griseus* has also been reported (reviewed by Beppu, 1992; Horinouchi and Beppu, 1992a, Horinouchi and Beppu, 1992b and Horinouchi and Beppu, 1994).

Of the autoregulatory diffusible factors known, A-factor is a typical example, and has been the most thoroughly investigated. Bioassays have been used to show that it is found widely among the streptomycetes, albeit in the form of different analogues (reviewed by Beppu, 1992; Horinouchi and Beppu, 1992a). However, its autoregulatory role appears to be limited to *S.griseus*, where it is crucial for development. A-factor non-producing mutants of *S.griseus* are unable to sporulate or produce streptomycin, although the mutants will revert to the wild type phenotype in the presence of exogenous factor. Szabo and Vitalis (1992) showed that sporulation occurred in the absence of A-factor but the formation of the aerial mycelium was blocked. Other species, such as *S.coelicolor* and *S.lividans*, produce A-factor, although it appears to have no crucial function in development or antibiotic synthesis (Horinouchi *et al.*, 1983). A-factor production is probably dependent upon the biosynthetic action of the *afsA* gene product. In *S.coelicolor* the *afsA* gene is found on the chromosome while in *S.griseus* it is almost certainly extrachromosomal (Hara *et al.*, 1983). Transcript mapping has shown that the *afsA* transcription and translation start points coincide (Horinouchi *et al.*, 1989b; section 1.2.3.2).

From *S.griseus* protein binding studies with tritium labelled A-factor, it was shown that A-factor associates in a molar ratio of 1:1 with a cytoplasmic protein that is present at a copy number of 30 to 40 per genome. No binding with membrane associated proteins was detected, implying that no surface receptor is required (Miyake *et al.*, 1989). Mutants deficient in this binding protein have been isolated, and for these A-factor is no longer essential for sporulation or streptomycin production. This indicates that the binding protein inhibits development, and this inhibition is relieved by A-factor (Miyake *et al.*, 1990). The structural gene encoding a similar butyrolactone type autoregulator binding protein from *Streptomyces verginiae* has been isolated and sequenced. On the basis of amino acid similarity to the *E.coli* NusG protein it was proposed that

the binding protein may have transcription antiterminator properties (Okamoto *et al.*, 1992).

Transcription studies on the *S.griseus strR* gene (which encodes a general regulator of streptomycin biosynthesis) have led to a model for A-factor action. It is thought that the unknown gene(s) that are regulated by A-factor binding protein in turn separately control antibiotic production and sporulation (Vujaklija *et al.*, 1991).

Willey *et al.* (1993) have determined that there are at least four diffusible factors at work in *S.coelicolor* development. Complementation tests utilising *bld* mutants which are unable to sporulate or produce SapB (section 1.3.4.1 and 1.3.4.4) were conducted. The ability of each class of *bld* mutant to restore SapB synthesis and sporulation to the other classes, when the two strains are grown in close proximity was tested. The results indicate that the four factors (termed signal A-D) act as a cascade, whereby A induces B, which in turn induces C, which finally induces D.

Discussion on the biological role of diffusible factors in *Streptomyces* is conspicuous by its paucity. While they cause almost simultaneous sporulation by the whole colony under laboratory culture conditions, the biological advantage of this (or if it even occurs in natural environments) is not clear. Perhaps diffusible factors allow measurement of the microenvironment volume and thus allow imminent starvation to be predicted. While certain factors are limited in their target species, others affect a wide range of streptomycetes. Certain *Streptomyces* products even induce fruiting body formation in fungi. These observations have lead Beppu (1992) to argue that intraspecies communication and symbiosis may occur. However it would seem more likely that diffusible factors are used to interfere with competing microorganisms, by disrupting their normal life cycle. Such behavior may explain why some streptomycetes produce diffusible factors that affect other species, yet have no known effect upon themselves.

1.3.6 Antibiotic production in streptomycetes

Many loci that are involved solely in the production of antibiotics, but not sporulation have been identified in *S.coelicolor*. These can be divided into two broad classes, those that encode structural and regulatory genes for specific antibiotic production pathways, and those that regulate the overall production of different antibiotics.

Of the general antibiotic production regulatory loci identified in *S.coelicolor*, *afs*B was (historically) the first. Mutants are unable to synthesise either A-factor or antibiotics, nor is this corrected by the addition of exogenous A-factor. It is thought that the *afs*B gene product is a cytoplasmic regulatory protein which controls the synthesis of both A-factor and antibiotics. Early cloning attempts resulted in the isolation of a gene that was incorrectly identified as *afs*B on the basis of its ability to complement *afs*B mutants (Horinouchi *et al.*, 1983; Hara *et al.*, 1983; Horinouchi *et al.*, 1984; Horinouchi and Beppu, 1984). Later work showed that the cloned gene was not allelic to *afs*B and probably complemented mutants by a bypass mechanism. The cloned gene was renamed *afs*R (Stein and Cohen, 1989).

The *afs*R gene has ATP and DNA binding domains and exhibits complex transcriptional features (Horinouchi *et al.*, 1990). The *afs*R gene product is known to be subject to phosphorylation by a membrane bound kinase, AfsK, in a manner reminiscent of two component regulatory systems (Hong *et al.*, 1991). Other loci that are involved in the global regulation of antibiotic production include the *afs*Q1,2 genes that also appear to code for a two component regulatory system (Ishizuka *et al.*, 1992) and the *abs*A and *abs*B genes (Champness *et al.*, 1992). The role of all these regulators is at present unclear.

The details of the biochemistry and regulation of the different *Streptomyces* antibiotics is too vast a subject to address here. For further information the reader is referred to Chater (1993) and Martin and Demain (1989). It is necessary however to briefly describe those antibiotics

produced by *S.coelicolor*, the organism used in the work reported here.

S.coelicolor is known to produce four antibiotics. The most well studied is the polyketide actinorhodin. This is a purple pigment which is produced by the enzymes of the *act* gene cluster (Rudd and Hopwood, 1979; Rudd and Hopwood, 1980; Malpartida and Hopwood, 1984; Malpartida and Hopwood, 1986). The other antibiotics include the red pigment undecylprodigiosin, produced by the products of the *red* gene cluster (Fietelson *et al.*, 1985; Tsao *et al.*, 1985; Fietelson *et al.*, 1986), CDA or calcium dependent antibiotic (Hopwood and Wright, 1983; Lakey *et al.*, 1983) and methylenomycin. The latter is encoded by plasmid borne genes (Kirby *et al.*, 1975; Chater and Bruton, 1985).

1.4 Promoter probes and their use in streptomycetes

The principle of fusing an indicator gene (of which the expression is easily monitored) to the promoter and controlling region of a gene under investigation has obvious merit. Thus, reporter gene fusions are a widespread and useful tool of molecular biologists. Fundamental to this type of work is the necessity for the recorded expression to accurately reflect the expression of the system being tested, and this is sometimes not the case. A common problem involves the effect of copy number of the gene fusion. Plasmid based promoter-probes with high copy numbers can conceivably titrate the regulatory proteins. In the case of positively acting regulators this will have little effect upon the timing (but not the level) of expression. However the titration of negatively acting control proteins can result in unregulated expression. Titration of regulatory proteins may also disturb the normal cell metabolism. Low copy number vectors overcome this to a degree, but systems that rely on single copy integration into the chromosome are preferable, although these may suffer from reduced sensitivity.

It is also essential that the indicator does not interfere with the normal host metabolism, either through the protein

itself, or the substrate it depends on. Consideration must also be taken of the effects of DNA supercoiling and the position of regulatory sites.

Many different indicator genes have been used in streptomycetes. Early promoter-probes relied on antibiotic resistance genes as indicators including the *amp*, *cat*, *vph* and *neo* genes (Forsman and Jaurin, 1987; Bibb and Cohen, 1982; Rodicio *et al.*, 1985; Ward *et al.*, 1986). These have proved useful but suffer a fundamental weakness in that antibiotic selection limits their use to promoters that are expressed in young colonies. Any promoter that is not activated at the time of plating and selection, for example those involved in sporulation, will not be detected, as the selection is lethal. To overcome this it is necessary to utilise indicator systems that produce visibly detectable signals. The ubiquitous *lacZ* gene of E.coli has proved to be poorly expressed in *Streptomyces*, (King and Chater, 1986), in addition to which most *Streptomyces* have endogenous β-galactocidase activity (Ingram *et al.*, 1989). Other visible indicators used in streptomycetes include a *Streptomyces* DNA fragment carrying pigment producing genes (Hourinouchi and Beppu, 1985; Feitelson, 1988), the *E.coli galK* gene (Brawner *et al.*, 1985), and the *luxAB* operon from *Vibrio harveyi* (Schauer *et al.*, 1988; Sohasky *et al.*, 1992). The pigment producing genes have since been shown to be part of the developmental *whiE* locus (section 1.3.4.3). Both the *galK* and *luxAB* dependent systems suffer the disadvantage of containing TTA codons (Cohn *et al.*, 1985; Johnston *et al.*, 1986), so the expression level of these indicators must depend upon the developmentally regulated *bldA* locus (section 1.3.4.2). They can not, therefore, be used in *bldA* mutants. The *luxAB* dependent system is also unsuitable for screening large numbers of colonies and requires exposure of the cells to X-ray film or the use of expensive equipment.

Recently Ingram *et al.* (1989) have tested the *xylE* gene from the *Pseudomonas putida* TOL plasmid as a reporter gene. It was found to be both sensitive and accurate. It contains no TTA codons. Phage vectors that utilise this system and which

can integrate into the chromosome in single copy have been developed (Bruton *et al.*, 1991).

It is obvious that an absence of TTA codons in the indicator gene of *Streptomyces* promoter probes is crucial. A suitable operon for use in this role is the *mel* operon (section 1.5), which has previously been used for gene fusion experiments in Gram negative bacteria (Altenbucher, 1988; Sugiyama *et al.*, 1990) Recently Paget *et al.* (1994) have developed a *mel* operon based promoter probe. This is discussed elsewhere (Chapter 2).

1.5 Tyrosinase and the *Streptomyces melC* operon

The enzyme tyrosinase (*o*-diphenol:oxygen oxidoreductase, EC 1.14.18.1) is found in a wide range of organisms, including plants, animals, fungi and bacteria. It catalyses the orthohydroxylation of monophenols and aromatic amines to *o*-diphenols, followed by oxidation of the *o*-diphenols to *o*-quinones. Melanin is formed by the oxidation of tyrosine to form DOPA and dopaquinone which condense to form an insoluble black pigment.

The ability to produce melanin is found in about one third of streptomycetes (Williams *et al.*, 1983), and research on the subject has focused on the tyrosinase of two species, *S.antibioticus* and *S.glaucescens*. On the basis of mutational studies there appear to be multiple loci that are involved in the regulation and production of tyrosinase, including *melA*, *melB* and *melC* (Hintermann *et al.*, 1985). For both *S.antibioticus* and *S.glaucescens* the operon responsible for the production of an active tyrosinase, termed the *melC* operon, has been cloned and sequenced (Katz *et al.*, 1983; Bernan *et al.*, 1985, Huber *et al.*, 1985; Huber *et al.*, 1987). The two loci show similarity in the organisation of the operon and the sequence of the genes therein. In *S.glaucesens* the *melC* phenotype is highly unstable, and loss of the ability to produce melanin has been shown to be accompanied by large deletions (Hintermann *et al.*, 1985; Yu and Chen, 1993). There are also reports that *melC* is

unstable on recombinant vectors (Chen *et al.*, 1992; Yu and Chen, 1993).

The induction and regulation of tyrosinase has been studied in several laboratories. It has been established that in *S.antibioticus* and *S.glaucescens* expression is induced by the presence of a small range of amino acids, with methionine exerting strong induction at low concentrations. Induction by phenylalanine and leucine was less marked. This regulation is at the level of transcription (Hintermann *et al.*, 1985; Katz and Betancourt, 1988; Betancourt *et al.*, 1992). It has been noted that expression of *S.glaucescens* tyrosinase was constitutive when the operon was cloned in *S.lividans* (Hintermann *et al.*, 1985). Expression of the natural tyrosinase was shown not to be induced by amino acids in *Streptomyces michiganensis*, but by copper ions, and this induction was repressed by ammonium (Held and Kutzner, 1990). Ochi (1987c) has noted that in *S.antibioticus* melanin production is halted in *rel* (stringent response negative) mutants (section 1.3.3). For both *S.antibioticus* and *S.glaucescens* transcript mapping has identified the transcription start point and promoter regions. The promoters bear no resemblance to the *E.coli* σ^{70} promoter class. Both genes within the operon (see below) are transcribed from the same single promoter. Deletion analyses have indicated that in both cases at least some of the regulatory regions lie upstream of the promoter (Geistlich *et al.*, 1989; Leu *et al.*, 1989).

The *melC* operon of both *S.antibioticus* and *S.glaucescens* consists of two genes, *melC1* and *melC2*. The downstream *melC2* codes for the tyrosinase enzyme. The tyrosinase activity is largely secreted upon expression (Crameri *et al.*, 1982), however when the *mel* operon of *S.antibioticus* was cloned in *S.lividans* (which does not produce melanin) it was reported that the majority of activity remained cell bound (Katz *et al.*, 1983). Later work, however, contradicted this, indicating that tyrosinase activity, from a cloned *mel* operon, was found extracellularly in *S.lividans*. (Betancourt *et al.*, 1992) The enzyme carries no signal-peptide sequence

at its N-terminal end (Crameri *et al.*, 1982; Bernan *et al.*, 1985).

The *S.antibioticus melC1* product is also secreted, but in this case the protein has a 30 amino acid signal-peptide that is removed after secretion (Chen *et al.*, 1992). Insertion inactivation studies upon this gene indicated that it might be essential for the expression of melanin (Katz *et al.*, 1983) although possible polar effects were not accounted for.

Bernan *et al.* (1985) have suggested that there are three possible ways that the MelC1 protein may exert its effect. Firstly, MelC1 might regulate the induction of tyrosinase, although there is no evidence yet that this is the case. Secondly, it is possible that MelC1 plays a role in the secretion of the tyrosinase. Finally, MelC1 may facilitate copper transfer and activation of the tyrosinase. There is experimental support for both of these last two proposals.

It has been shown, using combinations of cloned mutant and wild type *S.antibioticus melC1* and *melC2* genes, that inactivation of *melC1* did not inhibit *melC2* expression. It did, however, result in the production of tyrosinase that was enzymatically inactive. However, crude extracts of the inactive apotyrosinase could be converted *in vitro* to the active form by incubation with copper ions or, more efficiently, with copper ions and mycelial extract containing the wild type *melC1* product (Lee *et al.*, 1988). Later work using purified *S.antibioticus mel* operon products proved that MelC1 and MelC2 could form a copper free complex with a stoichiometric ratio of 1:1 between the two components. Addition of copper ions to the complex resulted in the incorporation of two copper molecules into, and the release of, active tyrosinase. These observations, in addition to site directed mutagenesis studies, lead to the proposal that MelC1 is a molecular chaperone that directs the correct folding of, and incorporation of copper ions into tyrosinase (Chen *et al.*, 1992; Chen *et al.*, 1993).

The *melC1* product also appears to play a role in the secretion of tyrosinase. Leu *et al.*, (1992) created

S.antibioticus mutants with lesions in the MelC1 signal peptide sequence. This, in some cases, resulted in the accumulation of intracellular tyrosinase due to a failure in the secretory mechanism. It was also noted that other *melC1*$^-$ mutants were deficient in tyrosinase export.

Considerable information concerning the *mel* operon has been gathered, however, the most important question has yet to be answered. The biological function of the operon has never been explained.

1.6 Motivation and aims

It is plain that the streptomycetes are complex and unusual organisms and the scope of potential research is diverse. Many fascinating questions are yet to be resolved. The control of basic metabolism, hyphal growth and branching patterns, sporulation, multicellular development and antibiotic production are just a few areas where some progress has been made.

Central to the understanding of these processes is their genetic regulation. It is in this area that streptomycetes display many unusual features, for example the temporally controlled expression of a tRNA (section 1.3.4.2) and the apparent complexity of promoters (section 1.2.3.1).

The *Streptomyces* promoters are so diverse that only one group has been classified, based on the similarity of its members to *E.coli* promoters. Subsequent derivation of a consensus sequence has led to the finding that it is the same as that of *E.coli*, which is not unexpected as that was the criterion by which the promoters were originally grouped (Strohl, 1992). This level of understanding about a subject as fundamental as transcription is unacceptable for a genus in which so much work is conducted and which is of such commercial value. The lack of knowledge in this area is surely inhibiting progress in other areas. A clear understanding of the roles of the different promoters would provide some guidance as to the function of various genes and thus give hints as to how research should further

proceed. The work described here was mainly aimed at addressing this question.

The work had the following aims:

1). To develop a *melC* operon based gene fusion system. At the time the work commenced, no promoter probes utilising a chromogenic indicator gene free of TTA codons existed. This meant that there was no reliable system for testing late promoters. Such a molecular tool is crucial for research on developmental genes.

2). To isolate late expressed promoters (among others) and determine which of these were developmentally regulated, by testing them in developmental mutants. Such promoters would then be sequenced. This approach was to serve two purposes. It is possible that in developmental mutants the promoters of a particular class may be inactive due to the lack of a particular σ factor (for example *whiG*, section 1.3.4.3 and 1.3.4.5). If large numbers of these promoters were isolated and sequenced then it might be possible that a consensus sequence could be derived. In addition, this approach would allow other workers to isolate developmentally regulated genes. It is remarkable that the isolation of developmental genes by classic mutagenesis has yielded so few developmental mutants, despite widespread efforts. The reasons for this are unknown. By isolating genes on the basis of their expression regulation, this problem may be overcome.

3). To attempt by computer assisted DNA sequence analysis to classify known *Streptomyces* promoters into groups based on their sequence similarity and other factors. The approach described above (aim number 2) relies on large numbers of promoters being isolated. If this was not achieved then further work (such as transcript mapping) would be unlikely to result in the classification of promoters. Such studies would simply supply costly additions to the growing list of streptomycete promoters, and be largely futile. The development of a classification system, and a less subjective method of determining promoter sequences would have many uses. Firstly it would shed some light on those

promoters already isolated by other workers. It would enable those researchers currently dissecting the role of σ factors in streptomycetes to choose which promoters to use in *in vitro* assays of purified σ factors. Furthermore it may give those researchers currently attempting site directed mutagenesis some clue as to which nucleotides to target. Finally, it may enable prediction of the position of promoters from sequence data alone.

CHAPTER TWO

CONSTRUCTION OF PROMOTER PROBES

Abstract

The feasibility of using the *S.antibioticus melC* operon as a reporter gene in *Streptomyces* promoter probes that were to be used for the isolation of developmentally regulated promoters was examined. Using *Streptomyces lividans* TK24 as a test organism growth conditions under which both sporulation and melanin production were possible were established. The effect of pH of the growth medium on melanin production was determined. A frame shift mutation was created in the *melC1* gene and it was shown that melanin production still occurred. Preliminary plasmid constructs designed for use as promoter probes were therefore made such that only the *melC2* gene would be expressed. Initial tests in which *Streptomyces coelicolor* M130 chromosomal DNA fragments were inserted upstream of the *mel* reporter showed that melanin production could be used as an expression indicator for genes that function late in the development of the colony. Several *Escherichia coli/Streptomyces* shuttle vectors with extended multiple cloning sites were created. These were ineffective as promoter probes, as it was shown that the multiple cloning site was acting as a promoter. A useful high copy number promoter probe with the multiple cloning site eliminated was constructed. This promoter probe utilised both genes *of the melC* operon. Low copy number promoter probes, which carried the *Streptomyces pememafaciens* plasmid pSPN1 origin of replication, were also constructed. These promoter probes also utilised two genes of the *melC* operon. The copy number, stability and mobility of the plasmids that utilised the pSPN1 origin of replication were established.

2.1 Introduction

The requirement that the activity of a reporter gene has to accurately reflect the activity of the controlling regions to which it is fused is obvious. This may not be the case for streptomycetes where that indicator gene contains TTA codons (section 1.4) due to the controlled expression of the cognate tRNA (sections 1.3.4.1 and 1.3.4.2). The following work was aimed at developing a *S.antibioticus melC* (section 1.5) operon based promoter probe.

2.2 Methods

General methods are described in Appendix C. Bacterial strains and plasmids are listed in Appendix B. Materials used are described in Appendix D.

2.2.1 DNA cloning strategies

DNA cloning procedures are detailed in Fig. 2.1

2.2.2 Testing melanin production on solid media

In each case the *S.lividans* TK24 was transformed with the plasmid to be tested (C.2.4). Thereafter spores were harvested and suspensions were diluted to between 1 x 10^5 and 1 x 10^6 colony forming units (cfu) ml^{-1} (C.1.2). These were either streaked to single colonies or spotted on the growth medium. In cases where the sample was spotted, 10 µl samples were placed on the medium without spreading. The *Streptomyces* strains, media and other procedures used are detailed in section 2.3.

2.2.3 Sequencing

Plasmid pBP223 was isolated from *E.coli* LKIII (C.3.2). The *S.coelicolor* M130 chromosomal DNA insert in plasmid pBP223 was sequenced (C.4.10) using the 1212 and MEL primers (Appendix D).

Plasmid pWB146 was isolated from *E.coli* LKIII (C.3.2). One strand of each of the outer limits of the pSPN1 origin of replication fragment was sequenced (C.4.10) using the 1212 and 1233 primers (Appendix D).

2.2.4 Plasmid stability tests

Tests were conducted using both *S.lividans* TK24 (pWB150) and *S.coelicolor* J1501 (pWB150). Spores of *Streptomyces* plasmid bearing strains were harvested from M3/7.0 (thiostrepton). After preliminary tests to determine the dilution required to generate countable colonies (between 50 and 1000 colonies/plate) the spore samples were each plated on 6 plates of M3/7.0 medium both with and without 50 $\mu g ml^{-1}$ thiostrepton. Colonies were counted after 1 week.

2.2.5 Plasmid mobility test

The following strains were cultured from spores on M3/7.0 medium with no antibiotic selection: *S.lividans* TK21, *S.lividans* TK24, *S.coelicolor* J1501, *S.lividans* TK24 (pWB150), *S.coelicolor* J1501 (pWB150), *S.lividans* TK21 (pWB150), mixed *S.lividans* TK21 (pWB150) and *S.coelicolor* J1501, mixed *S.lividans* TK21 (pWB150) and *S.lividans* TK24. Each inoculum was such that confluent growth was achieved. Mixed cultures were from approximately equal numbers of spores from each strain, as judged by cfu counts. Spores from each culture were harvested (C.1.3) and plated on M3/7.0 plates containing concentrations of 0 to 128 $\mu g ml^{-1}$ streptomycin, with and without 50 $\mu g ml^{-1}$ thiostrepton. Growth was recorded.

2.2.6 Plasmid copy number test

Chromosomal DNA from *S.coelicolor* J1501 was isolated and quantified (C.3.6 and C.3.13). Total DNA from *S.coelicolor* J1501 (pWB150) was isolated and quantified (C.3.7 and C.3.13). Plasmid pWB150 DNA was isolated from *E.coli* GM41 (pWB150) and quantified (C.3.2 and C.3.13). Aliquots of different quantities of DNA from the different sources were

subject to restriction digestion with *Cla*I and *Apa*I (C.4.1). The digested DNA samples consisted of plasmid pWB150 (50.0 ng to 0.75 ng), total DNA from *S.coelicolor* J1501 (pWB150) (2.50 µg to 0.31 µg) and chromosomal DNA from *S.coelicolor* J1501 (5.00 µg). The samples were subjected to agarose gel electrophoresis (C.4.2), and Southern blotting (C.4.9). Probing was with plasmid pUC19 labelled with ^{32}P by nick translation (C.4.8). Blots were subjected to autoradiography without the use of enhancer screens, to ensure a linear relationship between signal and band intensity.

2.3 Results and Discussion

2.3.1 Construction of plasmids

For purposes of clarity the entire cloning strategy used is summarised in Fig. 2.1. Detailed maps of plasmid starting materials that are not in common use (pIJ702, pIJ4642, pFu11 and pBlue) and important constructs (pWB150, pWB151 and pWB152) are given in Appendix F.

2.3.2 Initial questions

Before *mel* based promoter probes could be developed several questions had to be answered. Firstly, the growth conditions that would allow both sporulation and detection of tyrosinase activity had to be established. The growth conditions in work on starvation induced developmental phenomena are clearly of prime importance, particularly in the streptomycetes where developmental mutants display different phenotypes on complex and minimal medium (section 1.3.4.1). For this reason both defined and complex media had to be tested.

Secondly, it was necessary to determine how the cloned I *S.antibioticus mel* gene was regulated in *S.lividans*. The answer to this question is important because it affects the interpretation of experiments directed at answering the previous question.

Figure 2.1 Construction of plasmids Section A

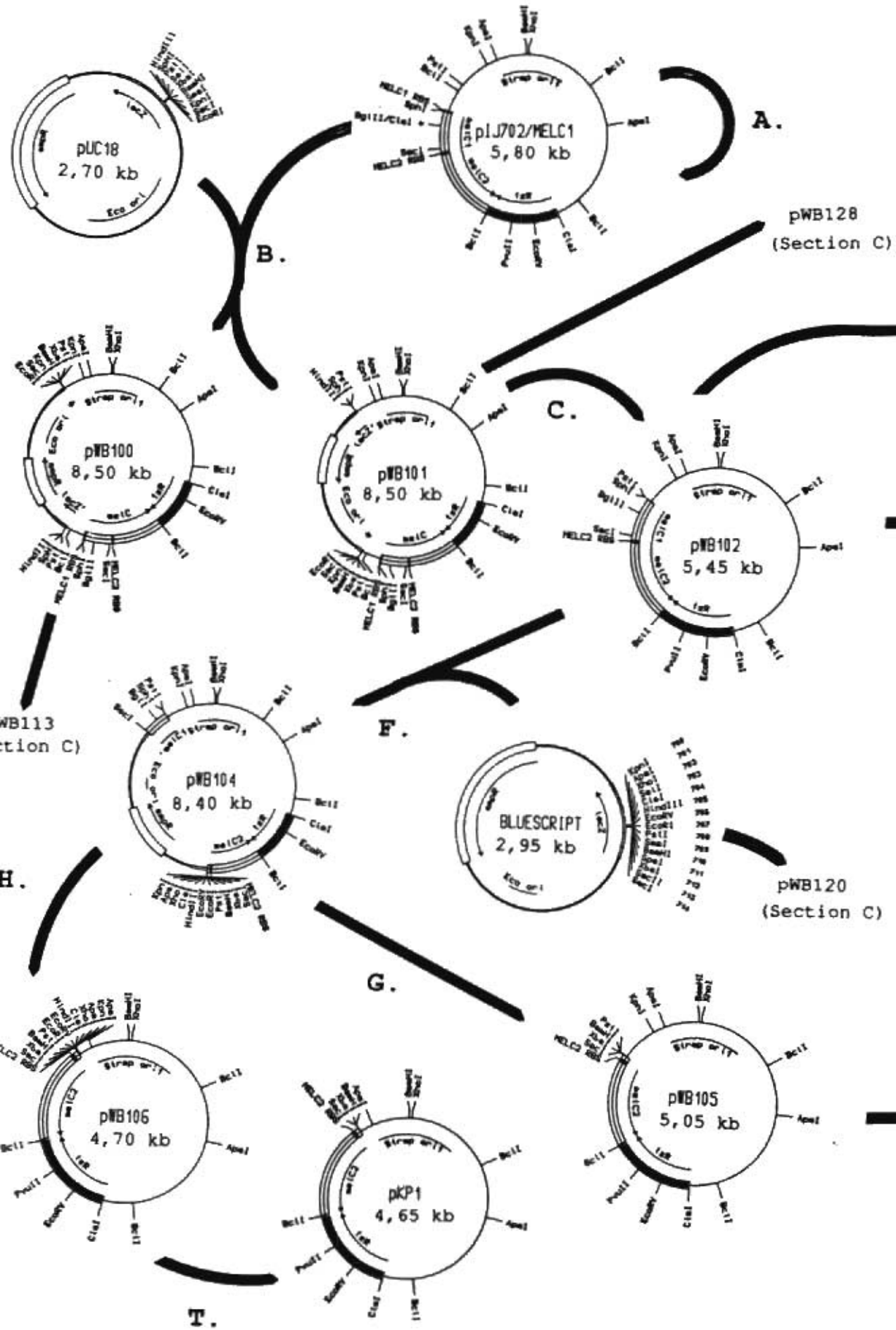


Figure 2.1 (continued) Section B

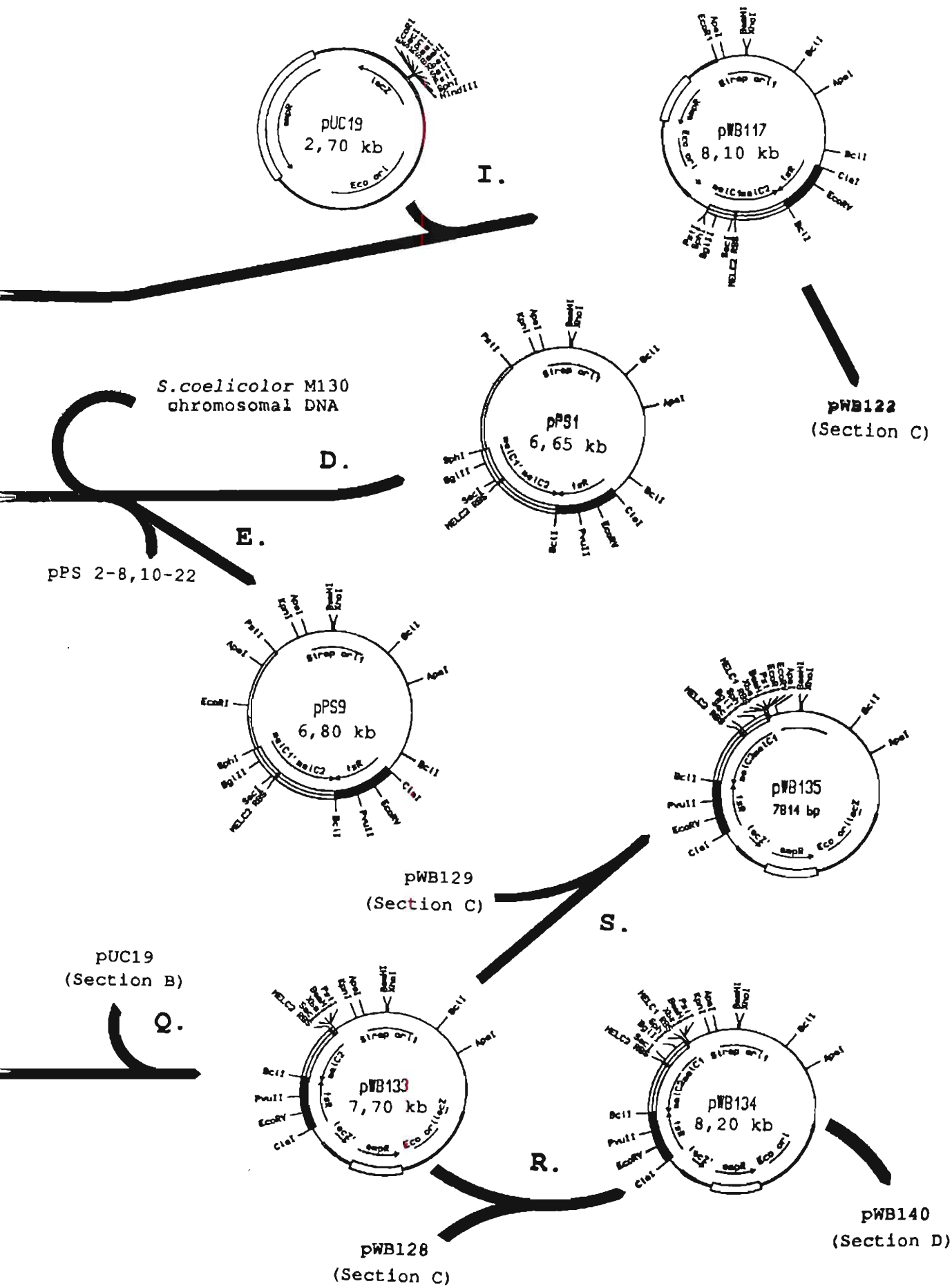


Figure 2.1 (continued) Section C

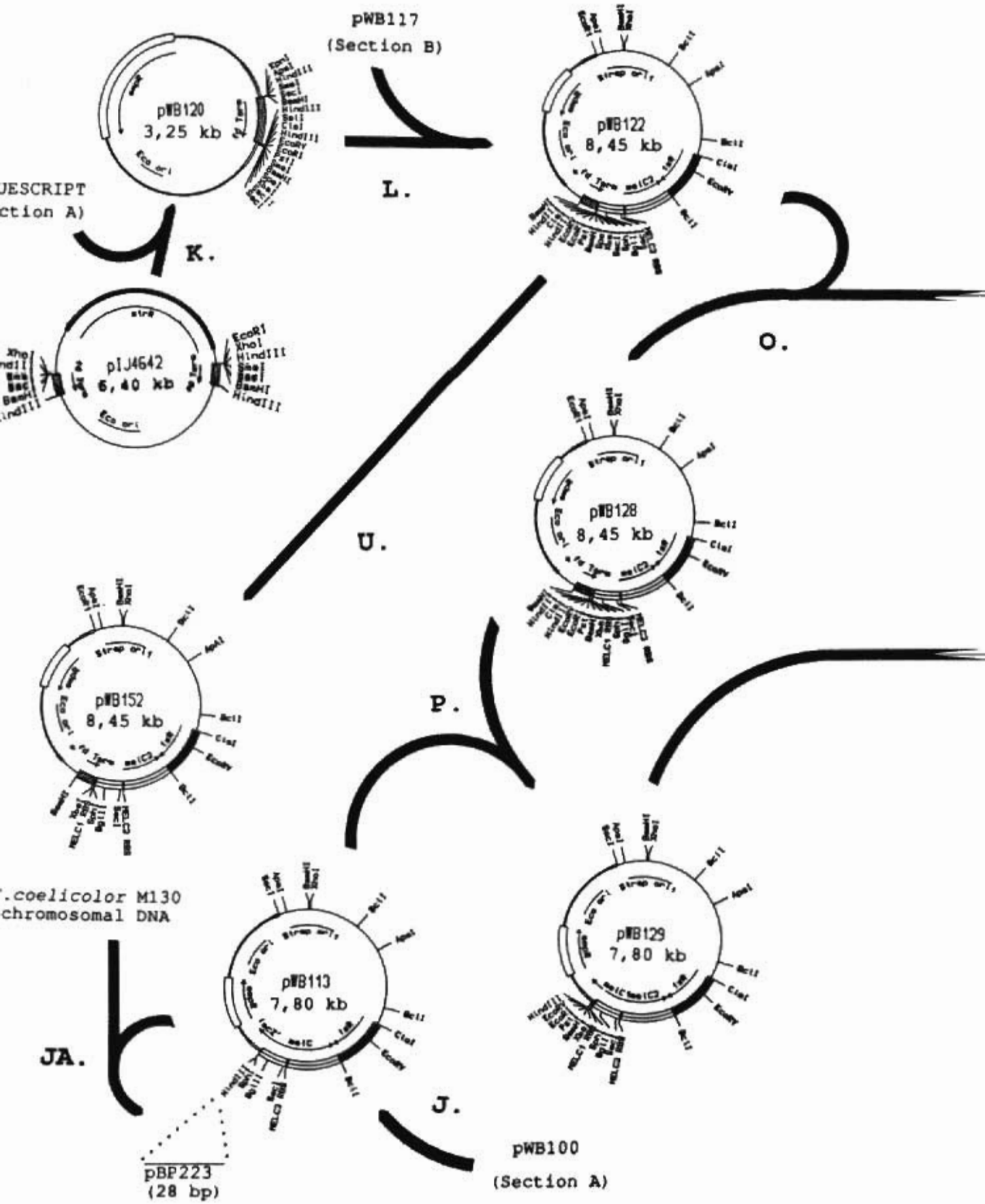
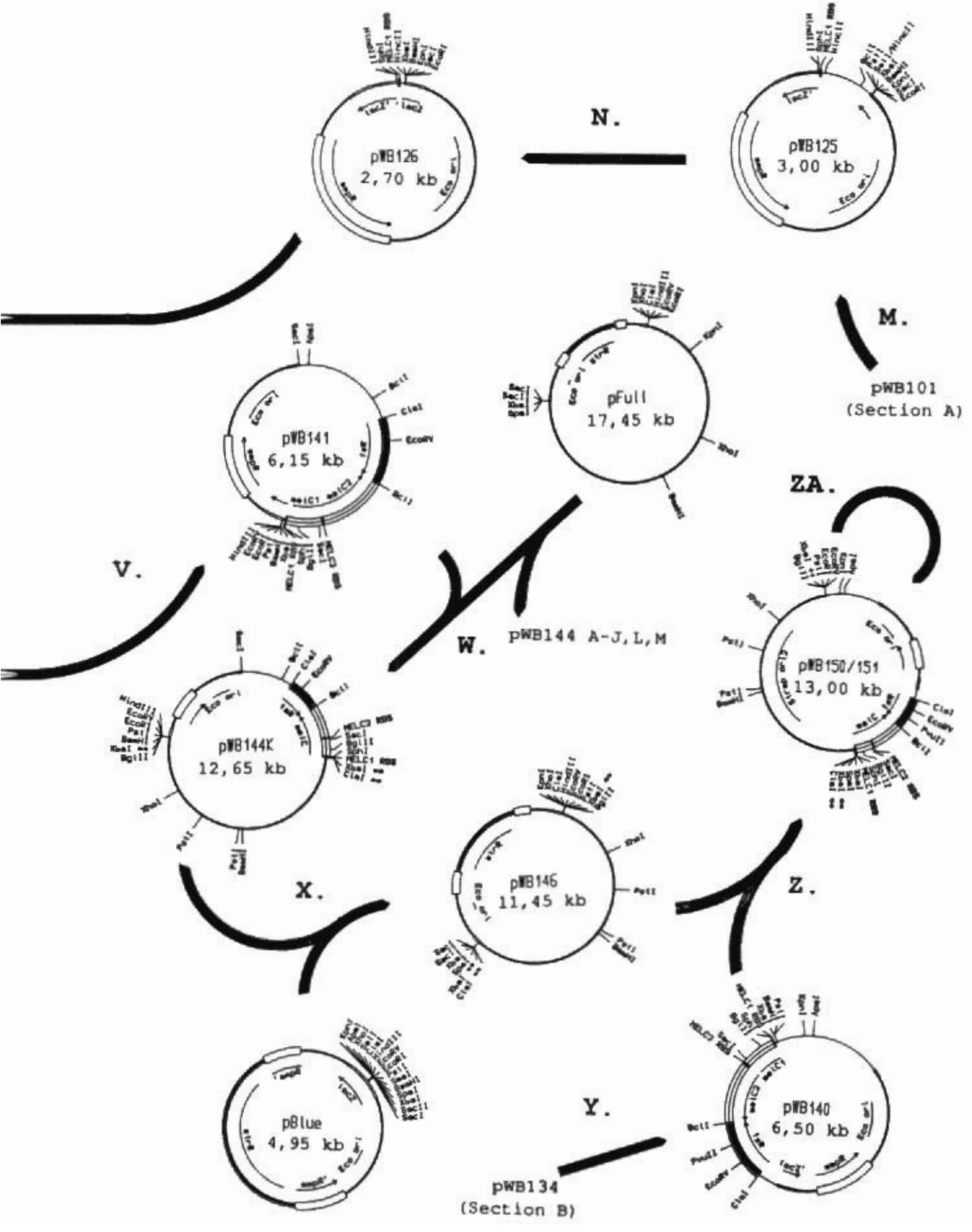


Figure 2.1 (continued) Section D



Legend to Figure 2.1

The following procedures utilised below are described in Appendix C: plasmid and chromosomal DNA isolation (C.3.1-5), partial and complete restriction endonuclease digestion (C.4.1), ligation (C.4.7), CIAP treatment (C.4.11), fill in of recessed 3' DNA ends (C.4.4 and C.4.5), removal of overhanging 3'DNA ends (C.4.5), partial fill in of recessed DNA ends (C.4.6), agarose gel electrophoresis (C.4.2), gel purification of DNA restriction fragments (C.4.3) and transformation of streptomycetes and *E.coli* (C.2.2, C.2.4 and C.2.5).

The plasmid DNA used in each cloning step was obtained by large scale isolation from the organism in which that particular plasmid was originally constructed, or in which it was originally maintained, unless otherwise specified. Large scale and small scale plasmid isolation are specified in brackets as "maxiprep" or "miniprep" respectively. For large scale plasmid isolation from streptomycetes, plasmids with the pIJ702 origin of replication were isolated by Method 1 (C.3.4) while those with the pSPN1 origin of replication were isolated by Method 2 (C.3.5). Antibiotic selective agents are specified in brackets at concentrations specified in Appendix B. The structure of each plasmid constructed was confirmed by restriction mapping. Growth media are described in Appendix D or Fig. 2.1.

A). pIJ702 was digested with *Bgl*II and the DNA ends were filled in, followed by recircularisation ligation and transformation of *S.lividans* TK24. Selection for pMELC1 was on M3 medium (thiostrepton). pMELC1 was identified by isolation (maxiprep) and restriction mapping. Four apparently identical plasmids were independently isolated.

B). *Pst*I digested pIJ702 and pUC18 were ligated. *E.coli* JM105 was transformed with the ligation mixture and white colonies were selected from XGal (ampicillin) plates. pWB101 was identified by isolation (miniprep) and restriction mapping.

C). *Sph*I digested pWB101 was recircularised by ligation and *S.lividans* TK24 was transformed with the ligation mix. Selection for pWB102 was on M3 (thiostrepton). pWB102 was identified by isolation (maxiprep) and restriction mapping.

D). *S.coelicolor* M130 chromosomal DNA and pWB102 were each digested with *Pst*I and *Sph*I and ligated. *S.lividans* TK24 was transformed with the mix. Selection was on MMTCu (thiostrepton). pPS1 was chosen for the ability to direct strong early melanin production on MMTCu.

E). As for D except that selection was on M3T/6.5 (thiostrepton). pPS9 was chosen for the ability to direct late melanin production.

F). *Sac*I digested pWB102 and pUC19 were ligated. *E.coli* JM105 was transformed with the ligation mixture and white colonies were selected from XGal (ampicillin) plates. pWB104 was identified by isolation (miniprep) and restriction mapping.

Legend to Figure 2.1 (continued)

G). *Pst*I digested pWB104 was recircularised by ligation. Selection was on M3 (thiostrepton) after *S.lividans* TK24 transformation, followed isolation (maxiprep) of pWB105 and restriction mapping.

H). As for G, except that the initial restriction endonuclease digestion was with *Kpn*I.

I). pWB102 was digested with *Kpn*I while pUC19 was digested with *Eco*RI and *Hin*dIII, following which the recessed DNA ends were converted to blunt ends in both cases. The pUC19 DNA was treated with CIAP and a mixed ligation performed. *E.coli* LKIII was transformed with the ligation mixture with ampicillin selection. Isolation (miniprep) and restriction mapping confirmed the structure pWB117.

J). *Sph*I digested pWB100 DNA was recircularised by ligation and used to transform *E.coli* LKIII. The resulting plasmid was isolated, cut with *Kpn*I, recircularised and used to transform *E.coli* LKIII. The resulting construct was isolated, cut with *Eco*RI and the recessed DNA ends were filled in before recircularisation and transformation of *E.coli* LKIII. Again, the resulting construct was isolated, cut with *Kpn*I and the overhanging DNA ends were filled in before recircularisation and transformation of *E.coli* LKIII. All selection was with ampicillin and all plasmids were identified by restriction mapping.

JA). pBP223 was created by ligating *Hae*III digested *S.coelicolor* M130 chromosomal DNA with *Hin*dIII digested pWB113 in which the recessed DNA ends had been converted to blunt ends and treated with CIAP. *E.coli* LKIII was transformed with the ligation mixture using ampicillin selection. Plasmid DNA was isolated from individual transformants (miniprep) and used to transform *S.lividans* TK24 by the small scale transformation method. Over 200 clones were tested. Selection was on M3T/7.0 (thiostrepton). pBP223 was selected for its ability to promote strong melanin expression and the smallness of the chromosomal DNA insert, as judged by restriction mapping.

K). The phage fd transcription terminator bearing plasmid pIJ4642 and Bluescript were digested with *Hin*dIII and *Xho*I respectively and the recessed DNA ends were filled in. The Bluescript DNA was then treated with CIAP. A mixed ligation was performed and *E.coli* LKIII was transformed with the ligation mix with ampicillin selection. Isolation (miniprep) and restriction mapping confirmed the identity of pWB120.

L). The recessed ends of *Sac*I digested pWB120 DNA were converted to blunt ends and the phage fd transcription terminator bearing fragment was isolated by gel purification. *Hin*dIII digested pWB117 DNA with the recessed ends converted to blunt ends and treated with CIAP was ligated with the terminator bearing DNA fragment. Following transformation of *E.coli* LKIII with ampicillin selection, pWB122 was identified by isolation (miniprep) and restriction mapping.

Legend to Figure 2.1 (continued)

M). *Sph*I digested pWB101 was recircularised by ligation. pWB125 was isolated (miniprep) from *E.coli* LKIII transformants (ampicillin selection) and its structure was confirmed by restriction mapping.

N). *Hinc*II digested pWB125 was recircularised by ligation. pWB126 was isolated (miniprep) from *E.coli* LKIII transformants (ampicillin selection) and its structure was confirmed by restriction mapping.

O). pWB122 and pWB126 were both subjected to restriction digestion with *Xba*I and *Sph*I, followed by mixed ligation and *E.coli* LKIII transformation (ampicillin selection). pWB128 was identified by isolation (miniprep) and restriction mapping.

P). pWB113 and pWB128 were both subjected to restriction digestion with *Xba*I and *Shp*I, followed by mixed ligation and *E.coli* LKIII transformation with ampicillin selection. pWB129 was identified by isolation (miniprep) and restriction mapping.

Q). pWB105 was partially digested with *Bcl*I such that linearised plasmid was largely generated and the overhanging DNA ends were filled in. This was ligated with *Eco*R1 and *Hind*III digested pUC19 in which the overhanging DNA ends had been filled in and treated with CIAP. After *E.coli* LKII transformation (ampicillin selection) pWB133 was identified by isolation (miniprep) and restriction mapping.

R). pWB133 and pWB128 were both subjected to restriction digestion with *Pst*I and *Sac*I, followed by mixed ligation and *E.coli* LKIII transformation with ampicillin selection. pWB134 was identified by isolation (miniprep) and restriction mapping.

S). pWB133 was subjected to *Kpn*I digestion and removal of the overhanging DNA ends with T4 polymerase. pWB129 was digested with *Hind*III and the recessed DNA ends were filled in. Both DNA samples were digested with *Sac*I, followed by mixed ligation and *E.coli* LKIII transformation with ampicillin selection. pWB135 was identified by isolation (miniprep) and restriction mapping.

T). pWB106 DNA was digested with *Kpn*I and *Pst*I and the overhanging DNA ends removed with T4 polymerase. Recircularisation ligation was followed by transformation of *S.lividans* TK24 with thiostrepton selection. pKP1 was identified by isolation (maxiprep) and restriction mapping.

U). pWB122 was digested with *Hind*II and *Xba*I, the recessed DNA ends were filled in and the plasmid recircularised by ligation. pWB152 was isolated (miniprep) after transformation of *E.coli* LKIII (ampicillin selection) and its identity confirmed by restriction mapping.

V). *Apa*I digested pWB129 DNA was recircularised by ligation followed by *E.coli* LKIII transformation (ampicillin selection). pWB141 was identified by isolation (miniprep) and restriction mapping.

Legend to Figure 2.1 (continued)

W). *Xba*I digested pWB141 DNA was partially filled in using dGTP, dATP and dTTP. *Sau*3A partially digested pFull DNA was partially filled in using dCTP, dATP and dTTP. Mixed ligation was followed by transformation of *E.coli* LKIII (ampicillin selection). Transformants from each plate were pooled, and mixed plasmid was isolated (maxiprep). The mix was used to transform *S.lividans* TK24, with selection on R2YE (thiostrepton). Plasmid was isolated (miniprep) from the least pigmented transformant in each pool and used to transform *E.coli* GM41 (ampicillin selection). The novel plasmids were termed pWB144A-M.

X). *Bam*H1 digested pBlue and *Xba*I digested pWB144K (isolated from *E.coli* GM41) were partially filled in using dGTP, dATP and dTTP, and dCTP, dATP and dTTP, respectively. Mixed ligation, transformation of *E.coli* LKIII (streptomycin selection), plasmid isolation (miniprep) and restriction mapping enabled the identification of pWB146.

Y). pWB134 was digested with *Apa*I, recircularised by ligation and used to transform *E.coli* LKIII (ampicillin selection). pWB140 was identified by isolation (miniprep) and restriction mapping.

Z). The overhanging DNA ends of *Pst*I digested pWB140 were removed and the DNA was digested with *Xba*I. *Hin*dIII digested pWB146 (isolated from *E.coli* LKIII) with the recessed DNA ends filled in was digested with *Xba*I. Mixed ligation, transformation of *E.coli* LKIII (ampicillin selection), plasmid isolation (miniprep) and restriction mapping resulted in the isolation of pWB150.

ZA). pWB150 was used to transform *E.coli* GM41 (ampicillin selection). pWB150 was isolated from a transformant (maxiprep), subject to partial *Sau*3A digestion, the recessed DNA ends were filled in, and the linearised form of the plasmid was isolated by gel purification. Recircularisation by self ligation, transformation of *E.coli* GM41 (streptomycin selection), plasmid isolation (miniprep) and restriction mapping allowed the isolation of pWB151.

* The *Bgl*II site in pIJ702 is absent in pMELC1 while the *Cla*I site is absent in pIJ702 but present in pMELC1.

** Signifies that these restriction sites (*Cla*I and *Xba*I in pWB144, pWB146 and pWB150) are subject to *dam* methylation protection.

++ Signifies that this *Xba*I site (which is subject to *dam* methylation protection) in pWB150 is absent in pWB151.

Thirdly, it was important to determine if the presence of the *melC1* gene was crucial for the use of *melC2* as an indicator gene. Other workers have found that the *melC1* gene is essential for the export and activity of tyrosinase. It was also noted that induction by amino acids of the cloned *mel* operon did not occur in *S.lividans* (section 1.5). At the time this work was initiated neither of these observations had been reported and the following work was undertaken because of this.

2.3.3 Initial tests on growth media

Initial tests showed that using the MMT growth medium recommended for tyrosinase detection (Hopwood *et al.*, 1985a), with *S.lividans* TK24 (pIJ702) resulted in melanin pigment not being visible (Table 2.1, row A). Possibly this was due to the absence of copper ions in the medium. To determine the effects of copper ion concentration and amino acid induction on tyrosinase expression, different media (based on MMT) but with different concentrations of copper ions and casamino acids were used. *S.lividans* TK24 (pIJ702) spores were streaked on these media and tyrosinase expression and sporulation was recorded (Table 2.1, rows B-H). From these results it can be seen that increasing copper concentrations result in increased ability to detect tyrosinase activity. It was also noted that copper concentrations of up to 0.05% do not visibly affect cell growth. Furthermore, it appeared that the presence of casamino acids stimulated tyrosinase production. However, in none of these cases was sporulation or growth vigorous.

Because simplicity is a desirable feature of minimal media, the effect of the elimination of "Tiger Milk" was tested (Table 2.1, row H and I). No visible difference in sporulation or melanin production was recorded. This implied that "Tiger Milk" could be excluded, however the addition of some of its components (which is necessary when auxotrophic strains are used, see Chapter 4) should have little effect.

Sporulation and *mel* expression tests on complex medium were conducted on M3 medium, traditionally used in this

Table 2.1 Media effects on melanin production

	MEDIA NAME	MEDIUM BASE[a]	TIGER MILK %	CASAMINO ACIDS %	TYROSINE[b] %	$CuCl_2$ %	pH[c]	MELANIN PRODUCTION[d] pIJ702	pMELC1	pPS1	pWB102
A	MMT	MM	0.75	0.6	0.0375	0.00	+7.0 (NaOH)	-			
B		MM	0.75	0.6	0.0375	1×10^{-6}	+7.0 (NaOH)	+			
C		MM	0.75	0.6	0.0375	1×10^{-4}	+7.0 (NaOH)	++			
D		MM	0.75	0.6	0.0375	0.05	+7.0 (NaOH)	++			
E		MM	0.75	0.0	0.0375	0.01	+7.0 (NaOH)	+	-	++	
F		MM	0.75	0.1	0.0375	0.01	+7.0 (NaOH)	++			
G	MMTCu	MM	0.75	0.6	0.0375	0.01	+7.0 (NaOH)	++			
H		MM	0.75	2.4	0.0375	0.01	+7.0 (NaOH)	++	+	++	-
I		MM	0.00	2.4	0.0375	0.01	+7.0 (NaOH)	++	+	++	
J	M3/6.5	M3/6.5	0.00	0.0	0.0000	0.00	6.5 (PHOS)	-	-	-	
K	M3T/6.5	M3/6.5	0.00	2.4	0.0375	0.01	6.5 (PHOS)	++	+	++	-

[a] MM base, L-asparagine 0.05%, K_2HPO_4 0.05%, $MgSO_4.7H_20$ 0.02%, $FeSO_4$ 0.001%, glucose 1%, agar 2%, adjusted to pH 7.0 with NaOH. M3/6.5 base, Malt extract 2.4%, Yeast extract 0.5%, Na_2HPO_4 0.016 M, NaH_2PO_4 0.034 M. Other components added to base after autoclaving (Appendix D).

[b] Dissolved in NaOH (Appendix D).

[c] pH adjusted with NaOH or sodium phosphate buffer. +7.0 signifies that the pH is unknown but greater than pH 7 (see text, section 2.3.5).

[d] All plasmids were in *S.lividans* TK24. From examination of individual colonies after 4 days. -, white colonies; +, brown discoloured colonies; ++, black colonies.

laboratory. Previous work (Bourn, 1987) had shown that the pH of the growth medium was a crucial determinant in the ability of *S.lividans* and *S.coelicolor* to sporulate. It had been shown that sporulation was vigorous on media of pH between 6.5 and 7.5. *S.lividans* TK24 (pIJ702) spores were streaked on M3 medium buffered at pH 6.5 with phosphate buffer and other additives (Table. 2.1, rows J and K). Sporulation was vigorous and expression of tyrosinase was recorded when copper ions and tyrosine were present. This medium was chosen for use in routine tests and was termed M3T/6.5 (see Table 2.1, also Appendix D).

2.3.4 The suitability of the *mel* locus as an indicator

Two plasmids, termed pMELC1 and pPS1, were created. Plasmid pMELC1 was identical to pIJ702 except that the *BglII* site had been eliminated by restriction digestion, filling in of the recessed DNA ends and religation (Fig. 2.1, step A). This resulted in a frame shift within the *melC1* gene. Although the mutation was not confirmed by sequencing, four independently made constructs were isolated and each showed the expected loss of the *BglII* site and were identical in all tests conducted. Furthermore, the *BglII* site was replaced by a *ClaI* restriction site, as would be expected from ligation of filled in *BglII* digested DNA.

Construction and selection of pPS1 is described in Fig. 2.1, steps B-D. In order to construct pPS1 the promoter probe pWB102 was first made (Fig. 2.1, steps B and C). Elimination of the *mel* promoter was such that the start codon of *melC1* was approximately 400 bp downstream from two presumed transcription terminators found on the *KpnI-ApaI* region bordering the origin of replication (Kendall and Cohen, 1988; see Appendix F). Plasmid pPS1 was shown by restriction mapping to be identical to pIJ702 except that the *PstI-SphI* *mel* promoter containing fragment had been removed and replaced with a 1.2-kb *PstI-SphI* *S.coelicolor* M130 chromosomal DNA fragment upstream of the promoterless *mel* operon. The clone was selected for its ability to direct the strong and early production of melanin on MMTCu medium (Fig.

2.1, step D). The cloned insert thus appeared to carry a powerful, but uncharacterised promoter that directed transcription of the *mel* operon. It should be noted that in this construct the *melC1* ribosome binding site had been eliminated, and *melC1* was not necessarily expressed.

S.lividans TK24 (pMELC1), *S.lividans* TK24 (pPS1), *S.lividans* TK24 (pWB102) and *S.lividans* TK24 (pIJ702) spores were streaked on different solid media as before. The results were compared in order to determine if the *melC1* gene was essential for expression (Table 2.1). In every case the expression of tyrosinase was considerably weaker for pMELC1 than pIJ702, but was detectable to a degree that indicated that the *melC2* gene alone could be used as an indicator gene under the conditions utilised here. Furthermore, the media dependent pattern of tyrosinase expression was the same for both clones.

No melanin production was observed with *S.lividans* TK24 (pWB102) indicating that any transcription readthrough from unknown promoters upstream of the *mel* operon was beyond the limits of detection using the *melC2* gene alone (see also Fig. 2.2). It should again be noted that in pWB102 the *melC1* ribosome binding site was lost, so *melC1* would not be translated.

As discussed above, tests using *S.lividans* TK24 (pIJ702) showed increased expression of tyrosinase with the addition of casamino acids. This could be due to the induction of the *mel* operon, however it is also possible that the effect was due to the presence of higher amounts of the tyrosine substrate in the medium. To test this, melanin production by *S.lividans* TK24 (pPS1) colonies was compared with that of *S.lividans* TK24 (pIJ702) and *S.lividans* TK24 (pMELC1) (Table 2.1). Significantly, strong tyrosinase expression was recorded in the absence of casamino acids only in the case of *S.lividans* TK24 (pPS1), and the slight increase when casamino acids were added was likely to be due to slightly more vigorous growth and an increase of tyrosine in the medium. This indicates that casamino acids were not required for tyrosinase activity but were required for its expression from the *mel* promoter. This observation is in contradiction

to that of Hinterman *et al.* (1985) who had found that *mel* expression from pIJ702 is constitutive in *S.lividans*. While it is not clear if the *melC1* gene is expressed in *S.lividans* TK24 (pPS1), this is irrelevant to the above discussion. These results also indicated that it would be unnecessary to supplement media with any amino acids besides tyrosine when using the *mel* operon as a promoter probe.

The results above (notably those for pPS1) indicate that the *mel* operon is suitable for use in a promoter probe. There remained, however, a further question, namely, is the *mel* operon suitable for the study of developmentally regulated promoters? It is possible that there are factors on which tyrosinase expression depends which are unknown and which are themselves dependent upon developmental stage. It is therefore possible that melanin expression could be prevented in old colonies, whatever promoter is driving transcription. To test if this is the case 18 plasmids (pPS5 to pPS22) were isolated (Fig. 2.1, step E). These were created in the same fashion as pPS1, except that selection for melanin producing colonies was on M3T/6.5 medium (thiostrepton). The plasmids were shown by restriction mapping to carry DNA inserts of various sizes. Spores from each clone were spotted on M3T/6.5 medium and the production of melanin and sporulation monitored. Clear differences in the timing of melanin production were recorded (examples shown in Fig. 2.2 A), with 13 of the colonies turning black within 2 days. Only 1 clone (pPS9) produced melanin at the onset of sporulation and expression was weaker than in most other clones (day 5; Fig. 2.2 A). Spores were also streaked on M3T/6.5 medium and several of the late expressing clones showed earlier expression where the colonies were clustered as opposed to widely separated on the plate (pPS12 for example; see Fig. 2.2 C). This argues that the expression is regulated by factors such as starvation or diffusible factors. The temporal regulation was also maintained in liquid M3T/6.5 medium (Fig. 2.2, B). Such results indicate that the *mel* operon is suitable for the detection of at least some of the late expressing promoters in *Streptomyces*. It should be noted that this does not necessarily mean that

all late or developmentally regulated promoters can be monitored by this system.

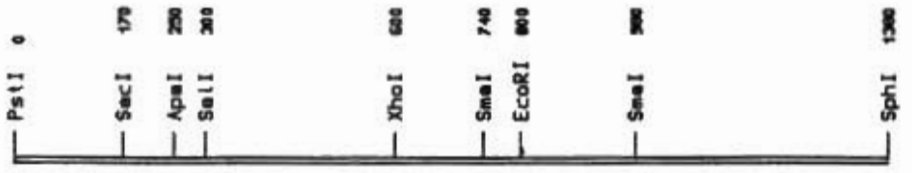
A restriction map of the chromosomal DNA insert in pPS9 was generated (Fig. 2.3) prior to further analysis of this clone (Chapter 4).

2.3.5 The effect of pH on melanin production

During the course of this work an attempt was made to develop a system whereby the tyrosine substrate was sprayed on the plates in a similar manner to the substrate for *xylE* based indicator systems (Ingram *et al.*, 1989). Such an approach eliminates the effects that the presence of the substrate may have on the growing colony. This would have allowed examination of genetic systems that may be so affected (eg. that of nitrogen metabolism) using *mel* based promoter probes. Because tyrosine is only soluble in basic solutions tyrosine dissolved in NaOH (Appendix D) was sprayed onto MMTCu(-) medium (which was essentially the same as MMTCu medium but without tyrosine or casamino acids; Appendix D) after colonies had appeared. *S.lividans* colonies carrying different plasmids which contained the *mel* operon were tested. In every case strong melanin production was recorded, even for *S.lividans* (pIJ702) in which tyrosinase expression had not been induced (notice there are no casamino acids in MMTCu(-) medium). The same effect was recorded if the tyrosine was incorporated in the solid medium and the colonies were sprayed with 0.1 M NaOH alone (data not shown). It therefore appeared that the pH of the medium affected melanin production.

To confirm this spores of plasmid bearing strains of *S.lividans* TK24 were spotted on a MM-like growth medium buffered at various pH values, with different concentrations of tyrosine, both with and without casamino acids. Initially pIJ702, pMELC1, pWB101 and pWB102 were tested (Table 2.2). Again, melanin production was inducible with casamino acids in the case of pIJ702, pMELC1 and pWB101 while no pigment was seen with pWB102 bearing patches. The pH of the medium clearly had a major effect on the level of melanin

Figure 2.3 Restriction map of pPS9 *S.coelicolor* DNA insert



Restriction sites tested but not present:

*Bcl*I, *Bgl*II, *Cla*I, *Eco*RV and *Pvu*II

Figure 2.4 Sequence of pBP223 *S.coelicolor* DNA insert

```
----------INSERT---------        SphI
CCACGGTGGAGCTGCTGTCGAGCCAGGGAGGTTGCATGC
                    RBS  RBS       START
```

Bold, *S.coelicolor* M130 chromosomal DNA insert.
RBS, **double underline**, potential ribosome binding site.
START, **underline**, Start codon of *mel*C1 gene.

Table 2.2 Melanin production of clones

A. Without casamino acids.

	PIGMENTATION OF GROWTH PATCH[a]														
	pH 6.0			pH 6.5			pH 7.0			pH 7.5			pH 8.0		
	TYROSINE CONCENTRATION[b]														
PLASMID	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
pIJ702	-	-	-	-	-	-	-	-	-	-	-	-	W	W	W
pMELC1	-	-	-	-	-	-	-	-	-	-	-	-	W	W	W
pWB101	-	-	-	-	-	-	-	-	-	-	-	-	W	W	W
pWB102	-	-	-	-	-	-	-	-	-	-	-	-	W	W	W
pWB105	-	-	-	-	-	-	-	-	-	-	-	-	W	W	W
pWB106	-	-	-	-	-	-	1	-	-	5	1	-	W	W	W
pWB113	-	-	-	-	-	-	-	-	-	-	-	-	W	W	W
pWB117	-	-	-	-	-	-	-	-	-	-	-	-	W	W	W
pWB122	-	-	-	-	-	-	-	-	-	-	-	-	W	W	W
pWB128	2	1	-	5	1	-	5	2	-	5	1	-	W	W	W
pWB129	2	1	-	5	1	-	5	2	-	5	1	-	W	W	W
pWB134	3	1	-	5	1	-	5	1	-	5	1	-	W	W	W
pWB135	3	1	-	5	1	-	5	1	-	5	1	-	W	W	W
pKP1	-	-	-	-	-	-	-	-	-	-	-	-	W	W	W
pWB150	-	-	-	-	-	-	-	-	-	-	-	-	W	W	W
pWB151	-	-	-	-	-	-	-	-	-	-	-	-	W	W	W
pWB152	-	-	-	-	-	-	-	-	-	-	-	-	W	W	W

B. With casamino acids.

	PIGMENTATION OF GROWTH PATCH[a]														
	pH 6.0			pH 6.5			pH 7.0			pH 7.5			pH 8.0		
	TYROSINE CONCENTRATION[b]														
PLASMID	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
pIJ702	3	2	-	5	3	1	5	4	1	5	4	1	5	3	1
pMELC1	-	-	-	-	-	-	1	-	-	2	-	-	3	-	-
pWB101	3	2	-	5	2	-	5	3	1	5	2	-	5	2	-
pWB102	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

A. Melanin production by *S.lividans* TK24 plasmid bearing strains. Growth was on Minimal Medium supplemented with $CuCl_2.2H_2O$ (0.01%), thiostrepton (50 $\mu g l^{-1}$), tyrosine (variable concentrations) and buffered with sodium phosphate buffer (0.05 M) to different pH values.

B. As for A except that growth medium was also supplemented with casamino acids (0.6%).

[a] W, weak growth of colony, - colourless, 1, faint brown discolouration, 2, dark brown discolouration, 3, black pigment in medium less than 1 mm from colony edge, 4 black pigment in medium between 1 mm and 5 mm from colony edge, 5 black pigment in medium more than 5 mm from colony edge.

[b] Tyrosine concentrations. A, 0.037%; B, 0.009%; C, 0.002%.

production, with higher levels produced at higher pH values. The reasons for this are unclear. Possibly it is due to the greater solubility of melanin in basic conditions. Alternatively the effect may arise from the greater stability of the tyrosinase, which has been reported to be most stable at pH 8 (with optimal activity at pH 6.8), in the case of *S.glaucescens* (Lerch and Ettlinger, 1972). A further possibility is that a high pH promotes the final, nonenzymatic, steps in the formation of melanin (V. Coyne, personal communication). It is unlikely that the effect of pH was due to increased tyrosinase expression levels as it was observed with all melanin producing clones, whatever the promoter that was driving transcription (see below).

It was also noted that sporulation and growth were more vigorous on buffered medium. Possibly this was because in the case of unbuffered medium tyrosine was added as a basic solution and presumably caused the pH to rise above the value of 7.0, to which the MM base was originally adjusted.

2.3.6 Construction and testing of *melC2* based promoter probes

As it appeared that the *melC2* gene alone could be used as a reporter gene to detect the activity of at least some promoters, potential promoter probes were constructed using only this gene. Two approaches were taken in parallel, each based on some of the principles adopted for the design of pWB102.

The loss of the *melC1* ribosome binding site with the retention of the *melC1* gene in pWB102 was a deliberate ploy based on the following logic. While it had been shown that *melC2* alone could be used as an indicator gene, it was obvious that this was less sensitive than when used in conjunction with *melC1* (sections 2.3.4 and 2.3.5). Weak promoters would therefore not necessarily be detected by transcriptional fusion. While the absence of the *melC1* ribosome binding site meant that this remained true for pWB102, it was thought that should in-frame translational fusions at the ATG start site of the *melC1* gene occur, weak

promoters could be detected. An additional advantage of this would be that translational regulation could also be detected. For example, if the fusion protein carried a TTA codon which was used for regulatory purposes (as has been speculated upon, see section 1.3.4.2), melanin production would reflect that regulation. Although it was then, and remains, unknown if the MelC1 protein would be functional when fused to another protein there is a further possibility. Inspection of known *Streptomyces* DNA sequence reveals that not only is ATG a rare codon but it is rare in every frame of coding sequence. Later studies involving computer assisted analysis of *Streptomyces* DNA sequence bias (Chapter 3) confirmed this (Appendix G). This implies that the *Sph*I restriction site (G/CATGC) will be commonly found at the start codon of proteins, as it is for the *mel* operon itself. It follows, therefore, that fusion of *Sph*I digested chromosomal DNA with the *Sph*I site of pWB102 would result in a relatively high proportion of clones in which the join was in frame between two ATG start codons. This sort of fusion would result in *mel*C1 expression that is largely dependent (excluding plasmid effects) on the factors that govern the original protein expression, including start codon mediated translational effects. Such effects may govern a number of developmentally regulated genes, including those of the *whiE* cluster (section 1.3.4.3) and transcribed by the *whiG* dependent promoters (section 1.3.4.5). Both of these have translation starts within other protein coding sequence.

Plasmids pWB105 and pWB106 were created with the purpose of extending the number of restriction sites that could be used for promoter cloning (Fig. 2.1., steps F, G and H). These plasmids were not designed for the detection of translational fusions, as for pWB102, but were based on the observation that melanin production was abolished in *S.lividans* TK24 (pWB102) colonies (section 2.3.5). In this case it is not crucial that no transcription of a promoter probe indicator gene occurs prior to the cloning of promoters, however this must be beyond the limits of detection of that probe in order for it to function. The absence of melanin production by pWB102 implies that this is

the case and that the transcription terminators upstream of the *mel* reporter function sufficiently well for the purpose. pWB105 and pWB106 were designed such that the relative positions between the terminators and reporter gene were retained, and in the case of pWB106, part of the intervening DNA was removed.

Both plasmids were tested for melanin production on minimal medium (Table 2.2). Although *S.lividans* TK24 (pWB105) remained unpigmented, unexpectedly *S.lividans* TK24 (pWB106) turned black. It was presumed (incorrectly, see section 2.3.7) that this was due to polar effects that had been eliminated by deletion of the 0.36 kb *Kpn*I-*Pst*I fragment and pWB106 was not used further.

In order to test whether pWB105 would function as a promoter probe *S.coelicolor* M130 *Pst*I-*Sph*I chromosomal DNA restriction fragments were cloned upstream of the *mel* indicator gene and melanin expression of these plasmids in *S.lividans* TK24 monitored on M3T/6.5 medium. Numerous black colonies were identified, some of which appeared to express late (data not shown). These were not investigated further because it was intended to develop a pWB105 based *E.coli/Streptomyces* shuttle vector to facilitate DNA manipulation (see below, plasmid pWB133).

The *E.coli/Streptomyces* shuttle vectors pWB113 and pWB117 were designed in such a manner that the potential to detect translational fusions as for pWB102 was retained. It was expected however that a transcription terminator would have to be inserted upstream of the indicator gene, as has been required for other promoter probes (eg. Ward *et al.*, 1986). pWB113 and pWB117 were constructed (Fig. 2.1, steps I and J) and *S.lividans* TK24 (pWB113) and *S.lividans* TK24 (pWB117) spores were used to test for melanin production. No melanin production was recorded after four days for either clone (Table 2.2) although at pH 8 discolouration was observed after periods of longer than this in the case of *S.lividans* TK24 (pWB117).

To test that pWB113 could function as a promoter probe *Hae*III digested *S.coelicolor* M130 chromosomal DNA fragments

were then inserted into the *Hin*dIII restriction site (Fig. 2.1, step JA). Several melanin producing colonies of *S.lividans* TK24 carrying pWB113 with such inserts were found. A single clone, pBP223, was selected for further testing on the basis of strong early melanin production on M3T/7.0 medium, and the fact that restriction mapping showed that the insert was extremely small. The putative promoter bearing insert was sequenced (section 2.2.3) and is shown in Fig. 2.4. From this it can be seen that the fragment is 28 bp long and carries no obvious *E.coli*-like promoter sequences. There are however, two ribosome binding site-like sequences positioned 8 and 13 base pairs from the *mel*C1 start codon. Although each site bears only limited similarity to the ribosome binding site consensus sequence ([a/g]GGAGG) such short sequences are known to be functional in streptomycetes (section 1.2.3.2). Furthermore an 8 base pair spacing between the binding site and the start codon is the most common for streptomycetes (Strohl, 1992) and within the effective range (5-13 bp) for prokaryotes (Kozak, 1983).

The results of this sequencing experiment had two possible interpretations. Firstly, it was possible that the insert carried a promoter. The fact that the fragment was only 28 base pairs in length did not exclude this possibility as it is known that such short stretches of DNA can act as promoters in *Streptomyces* (see section 1.2.3.1). The alternative was the previously unrecognised possibility that the insert could carry only a ribosomal binding site, or a sequence that resembled one closely enough to be active in the correct context. Inspection of the insert sequence revealed that such a correctly positioned (in relation to the *mel*C1 start codon) site was present. If the melanin production observed was due to such a feature then this implied that there was some degree of transcription of the *mel* operon occurring from an unknown upstream promoter. Such a promoter might be too weak to cause melanin production in pWB113 because the *mel*C1 gene was not translated. The correct placement of a ribosome binding site and subsequent translation of the *mel*C1 gene might then have sufficed to increase the level of tyrosinase activity such that melanin

was formed to a detectable level. Such an interpretation is consistent with the previous results.

Tests for transcription typically involve RNA hybridisation experiments of some form (Maniatis *et al.*, 1982). However, such tests have inherent weaknesses, in that they are only capable of proving that transcription is occurring, and not that it fails to occur. It can only be argued in such cases that transcription levels are beyond the limits of detection. Furthermore, if the transcription is transient, or limited to few cells, as may occur in a differentiating colony then levels of mRNA may be relatively low. Comparisons between the tyrosinase expression levels of pWB113 and both pWB106 and pMELC1 indicated that any putative promoter upstream of the *mel* operon of pWB113 must be relatively weak. These considerations mean that it is difficult to develop a reporter system designed to detect translational fusions as described here. Such a system must always suffer the possibility that it will mainly detect ribosome binding sites, particularly if very short fragments of DNA are inserted adjacent to the reporter gene. It was, therefore, crucial in this case that any transcription from upstream promoters be eliminated.

For this reason two plasmids, pWB122 and pWB128 were constructed (Fig. 2.1, steps K, L, M, N and O). These constructs were similar to pWB117 except for the region directly upstream of the *mel* operon. In addition to the multiple cloning site sequence on each plasmid, pWB122 and pWB128 carried the phage fd transcription terminator and pWB128 carried the ribosome binding site of *mel*C1 in its correct position. Such an approach was intended to achieve three objectives at once: to eliminate all possible transcription readthrough by promoters upstream of the *mel* indicator, to prove that this had been eliminated by using the complete *mel* operon as an indicator (including the *mel*C1 ribosome binding site) and to create a sensitive promoter probe. Patches of *S.lividans* (pWB122) and *S.lividans* (pWB128) were tested for melanin production as before. It was expected that colonies containing pWB122 would not produce melanin, and this was the case (Table 2.2). However

colonies containing pWB128 which were expected to be colourless or only weakly pigmented produced melanin, at levels roughly equivalent to that of colonies of *S.lividans* TK24 (pIJ702) in which melanin production was induced (Table 2.2). It was initially thought that this indicated that the fd transcription terminator was not effective in *S.lividans* under the growth conditions used here. It had previously been shown that the terminator was 97% effective in *S.lividans* but this work was conducted using stationary phase cultures grown in liquid medium (Ward *et al.*, 1986). Thus it could only strictly be said that the terminator was effective under those conditions, and not necessarily in differentiating colonies on solid medium.

A standard solution to the elimination of the effects of upstream promoters in antibiotic resistance based promoter probes involves selection for cloned promoters at higher levels of antibiotic. The equivalent approach using chromogenic indicators would be to use lower levels of indicator substrate. Such a solution is unsatisfactory, because unlike lethal indicators, there is no precise cut off point for chromogenic indicators. This approach does not affect the sensitivity of lethal indicator genes, however the reduction in sensitivity is the basis of the solution to the problem for chromogenic indicators. This pitfall is compounded when the promoter probe is specifically designed for the isolation of late expressed genes because even low levels of expression are unacceptable as the pigment can accumulate over the several days prior to *Streptomyces* sporulation.

As it appeared impossible to prevent transcription from upstream promoters in pWB117 using the fd transcription terminator, plasmid pWB129 was constructed (Fig. 2.1, step P). This construct was similar to pWB113 except that the *melC1* ribosome binding site was included. It should be noted that compared to pWB117 and pWB128 the pUC18 derived part of the plasmid was reversed in relation to the *mel* reporter. This meant that the *lac* promoter was no longer in a position to drive transcription of the *mel* operon. This approach was based on the observation that it is common practice, when

using pUC18 or pUC19 for cloning purposes, to compare expression of the cloned gene in both orientations with respect to the *lac* promoter to determine if expression is from the *lac* or a cloned promoter (R. Kirby, personal communication). That such an approach can be taken implies that only weak or no transcription of the *lac* antisense DNA strand occurs in *E.coli*. Expression tests with *S.lividans* TK24 (pWB129) showed strong melanin production (Table 2.2) and this approach was abandoned.

Further attempts to eliminate readthrough transcription were made. The plasmid pWB134 was created (Fig. 2.1, steps Q and R) in order to place the complete promoterless *mel* operon downstream of the two *rep* transcription terminators on an *E.coli/Streptomyces* shuttle vector. Tests with *S.lividans* TK24 (pWB105) and *S.lividans* TK24 pWB102 (section 2.3.6) had indicated that any transcription through these terminators was too weak to be detected using the *Mel*C2 gene alone. Again, expression testing showed strong melanin production (Table 2.2) indicating that some transcription was occurring. This transcription might not have been as a result of promoters beyond the transcription terminator but from the region between the terminator and the multiple cloning site. For this reason pWB135 was created (Fig. 2.1, step S) and *S.lividans* TK24 (pWB135) was tested for melanin production. Again strong expression was recorded (Table 2.2).

2.3.7 The Bluescript multiple cloning site can act as a promoter in *Streptomyces*

Due to the results described in section 2.3.6, previous results were re-examined. The difference in tyrosinase expression between pWB105 and pWB106 was previously attributed to polar effects (section 2.3.6). However, examination of the sequence of the Bluescript multiple cloning site revealed a possible promoter-like sequence. If this, or another sequence was operating as a promoter in *Streptomyces* then this would explain the tyrosinase expression patterns of colonies containing pWB105, pWB106,

pWB128, pWB129, pWB134 and pWB135. Subsequent computer analysis of the multiple cloning site sequence supported the plausibility of this theory (see Chapter 3).

To test the possibility that the Bluescript multiple cloning site sequence was acting as a promoter in *S.lividans* the plasmid pKP1 was created (Fig. 2.1, step T). Expression of tyrosinase in *Streptomyces* TK24 (pKP1) was monitored as before and it was found that no melanin was produced (Table 2.2). As the only difference between pWB106 and pKP1 was the presence of part of the Bluescript multiple cloning site, this indicated that this DNA fragment (as a whole or in part) was itself acting as a promoter. Other workers have also found that a short fragment of artificial DNA sequence could act as a promoter in *Streptomyces* (Brasch *et al.*, 1993).

2.3.8 Construction and preliminary testing of a high copy number promoter probe

To confirm that the Bluescript multiple cloning site was acting as a promoter in pWB128, and at the same time create a useful promoter probe, plasmid pWB152 was constructed (Fig. 2.1, step U). Tyrosinase expression of *S.lividans* TK24 (pWB152) was tested and no detectable expression was observed, confirming that the transcription recorded previously was due to the multiple cloning site Table 2.2).

*Sau*3A restriction endonuclease digested *S.coelicolor* M130 chromosomal DNA fragments were cloned in the *Xba*I restriction site of pWB152. Cloning utilised the partial fill in procedure (C.4.6) for both the chromosomal and pWB152 DNA (resulting in single base overhanging DNA ends) followed by mixed ligation (C.4.7). Transformation of *S.lividans* TK24 with the ligation mix and selection on M3T/7.0 resulted in the isolation of melanin producing colonies (data not shown). This indicated that pWB152 could be utilised as a promoter probe, however the clones were not investigated further as an alternative strategy using a low copy number promoter probe had proved successful. It should be noted that pWB152 is very similar to the promoter probe

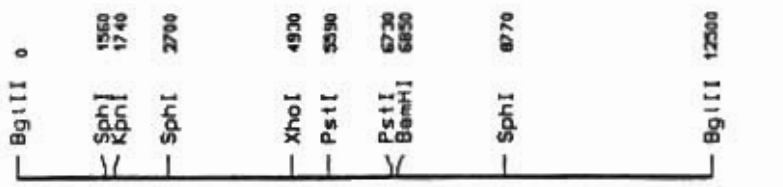
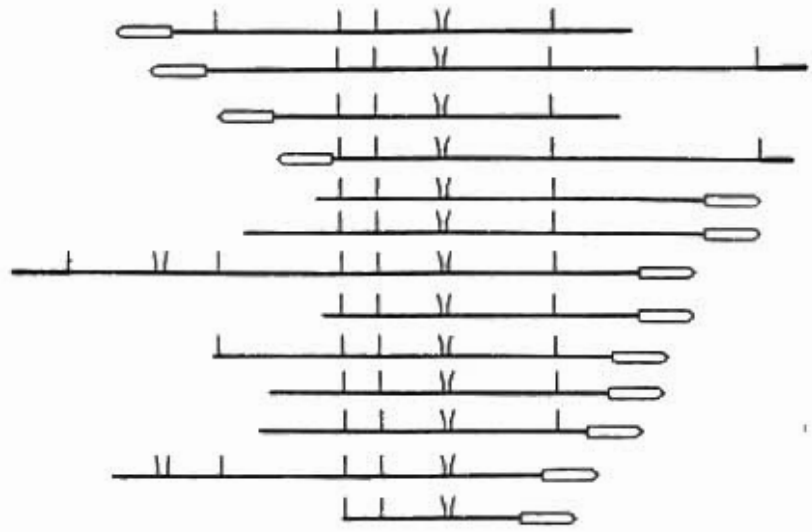
pMT3000 described in the recently published work of Paget *et al.* (1994).

2.3.9 Isolation of minimal replicon of pSPN1

As the activity of a reporter gene is potentially more faithful at a low copy number (section 1.4), and because of the difficulty experienced in the construction of a high copy number promoter probe (section 2.3.6), the first steps towards the creation of a low copy number promoter probe were taken. This involved the isolation of the minimal replicon for the plasmid pSPN1. This plasmid had been isolated from *Streptomyces penemafaciens* by Smith (1991), who reported that it occurred at a copy number of 0.2/chromosome. This figure was based on the assumption that the *S.lividans* chromosome size was 5 mb. It is now known that the chromosome is 8 mb (section 1.2.3), so the estimation of the copy number from the original data should correctly be 0.13/chromosome.

At the time of construction it was thought that the phage fd terminator was not efficient enough to be used in a promoter probe to prevent transcription readthrough from promoters upstream of the indicator gene (section 2.3.6). A cloning strategy by which the replication origin of pSPN1 would be used to ensure that no transcription readthrough into the indicator gene occurred was therefore devised. The source of the pSPN1 origin of replication was the plasmid pFu11 (Smith, 1991; Appendix F). Initially the limited restriction map of the pSPN1 origin of replication was extended (Fig. 2.5). Plasmid pWB141 was created for use as a replication origin trap, as it had no origin of replication that was functional in *Streptomyces* (Fig. 2.1, step V). Following the cloning strategy outlined (Fig. 2.1, step W), 13 different plasmids carrying *Sau*3A partial digest generated fragments of the pSPN1 origin of replication were isolated. These were subject to restriction mapping and the results are shown in Fig. 2.5. From this it can be seen that the smallest fragment isolated in this experiment that can support replication is 3.25 kb in length.

Figure 2.5 Replication origin fragments in pFull and pWB144 plasmid series and melanin expression

FRAGMENT FROM PLASMID	MELANIN PRODUCTION ON R2YE	MMTCu/7.0	DNA FRAGMENT AND RESTRICTION MAP
pFULL	-	-	
pWB144A	NO	YES	
pWB144B	NO	YES	
pWB144C	NO	YES	
pWB144D	WEAK	YES	
pWB144E	NO	YES	
pWB144F	NO	YES	
pWB144G	WEAK	YES	
pWB144H	NO	WEAK	
pWB144I	NO	YES	
pWB144J	NO	YES	
pWB144K	NO	NO	
pWB144L	YES	YES	
pWB144M	YES	YES	

The 13 plasmids were then tested for their ability to direct melanin production. Spores of *S.lividans* TK24 strains carrying the plasmids were spotted on MMTCu/7.0 medium and the results are shown in Fig. 2.5. Only one clone showed no melanin production. This clone was termed pWB144K and was used for the construction of the promoter probes pWB150 and pWB151.

2.3.10 Construction of low copy number promoter probes

The plasmid pWB144K had all the characteristics necessary for a promoter probe except for any useful promoter cloning sites. The promoter probe plasmid pWB150 was therefore constructed (Fig. 2.1, step Z). pWB146 (Fig. 2.1, step X) was used as the source of the origin of replication and the construct pWB140 (Fig. 2.1, step Y) was used as a replication origin trap. It should be noted that the partial fill in and ligation of *Xba*I and *Sau*3A digested DNA always results in the recreation of an *Xba*I site which is subject to methylation protection by *dam* methylation systems. Thus pWB150 could only be used for cloning promoters if the plasmid DNA was extracted from bacteria that have the *dam* methylation system. The streptomycetes do not have such a methylation system, as proven by the susceptibility of *Streptomyces* DNA to *Bcl*I restriction enzyme digestion (for example see Katz *et al.*, 1983). It is preferable to use vector DNA derived from the host organism when cloning in the streptomycetes, in order to achieve a high transformation rate. Likewise, vector DNA derived from *E.coli dam*$^-$ strains generates high numbers of transformants (Neeson and Volckaert, 1989). Plasmid pWB151, which can be used when isolated from *dam*$^-$ strains was therefore constructed (Fig. 2.1, step ZA).

Both of these constructs were tested for melanin production, which was shown not to occur (Table 2.2, see also Chapter 4), even after extended incubation. This indicates that any transcription readthrough from upstream of the *mel* reporter operon is below the level of detection of the system under

the conditions used here. The plasmids pWB150 and pWB151 could therefore be used as promoter probes.

2.3.11 The stability of pWB150

The concept of plasmid stability, when applied to *Streptomyces*, is fundamentally different from the concept as it is applied to unicellular bacteria. Because the streptomycetes have a true mycelium (section 1.3.2) the stability of the plasmid cannot be measured in terms of individual cells and can only be quantified from one round of sporulation to the next. Thus the concept of a plasmid with a copy number of less than one/chromosome proposed by Smith (1991) is plausible, but does have implications with regard to whether it is possible that each spore will result in a plasmid bearing colony, given that the chromosomes outnumber the plasmids. The production of plasmid free spores would not preclude the use of the pSPN1 origin of replication in a promoter probe, and indeed this characteristic could prove useful. Neither would any instability of the plasmid in the mycelium affect the function of the promoter probe, provided it was always used in conjunction with antibiotic selection. This selection mechanism must, however, operate within the cell (as does the ribosome mediated thiostrepton resistance used here; (Holmes *et al.*, 1993) and not extracellularly (for example by enzyme inactivation of the antibiotic). However, it might be desirable to use the promoter probe in the absence of antibiotic selection because the presence of antibiotics can affect growth rate (personal observation, data not shown) and certain promoters are known to be regulated by thiostrepton (Holmes *et al.*, 1993).

The plasmid stability test was carried out as described (section 2.2.4). For *S.lividans* TK24 (pWB150) and *S.coelicolor* J1501 (pWB150) the percentage of colony forming units that were thiostrepton resistant after one round of sporulation were 66.9% and 69.2% respectively. These data indicate that the plasmid is unstable. However, it should be noted that due to the relatively low numbers of colonies

counted (6 plates witha an average of 117.6 colonies /plate was the lowest) this is a very rough estimation. Furthermore, caution must be used in the interpretation of these results. The presence of a fibrous sheath around the spores (section 1.3.2) could result in the spore suspension containing long chains of spores. Such a feature could obscure the fact that the chains might carry both plasmid free and plasmid bearing spores. Thus the instability of pWB150 may be greater than is indicated in the above experiment. It must also be noted that these results are true only for the growth conditions used here. As noted previously (section 1.3.2), within a colony there may be several rounds of sporulation and germination, and the degree of this is presumably media and inoculum dependent. These factors may, therefore, also determine the apparent stability of the plasmid. The possibility that the plasmid is mobilisable could also affect the results.

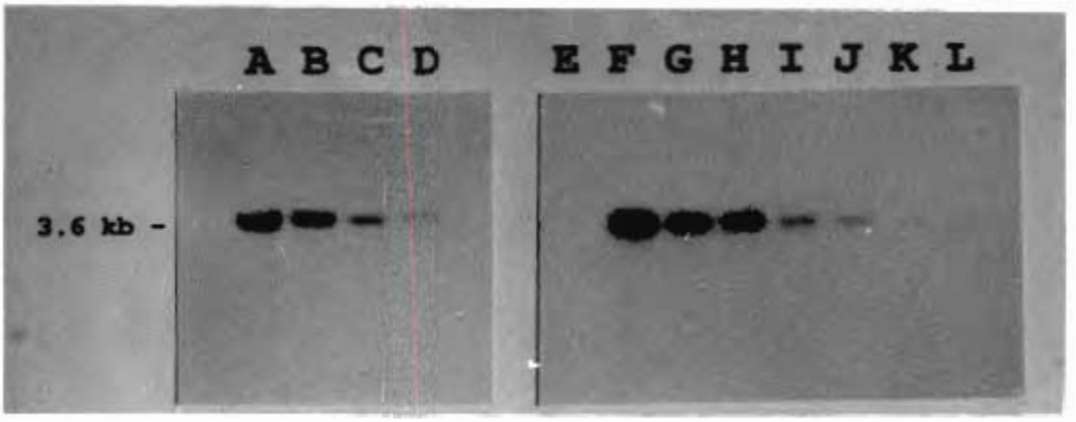
2.3.12 The mobility of pWB150

The fragment of the pSPN1 origin of replication used in pWB150 was approximately 3 kb larger than the minimal replicon (sections 2.3.9 and 2.3.10), and thus may encode functions such as mobilisation ability. As this could have a bearing on the function of the promoter probe the mobilisation ability of pWB150 was tested (section 2.2.5). The results are shown in Table 2.3. No colonies that were resistant to both thiostrepton and streptomycin were found. This indicated that pWB150 was not mobilisable under these conditions and within the limits of the numbers tested.

2.3.13 Estimation of the copy number of pWB150

Although the instability of pWB150 was high, it was not in the order expected from the results of copy number experiments by Smith (1991) if it were assumed that no spore specific partitioning mechanism existed. Because the fragment of the pSPN1 origin of replication used here was shorter than that isolated by Smith (1991), the possibility existed that the plasmid copy number control

Figure 2.6 Test for copy number of pWB150



A-D, Total DNA from *S.coelicolor* J1501 (pWB150). **A**, 2.5 µg; **B**, 1.25 µg; **C**, 0.65 µg; **D**, 0.31 µg. **E**, 5.0 µg chromosomal DNA from *S.coelicolor* J1501. **F-L**, pWB150 DNA. **F**, 50 ng; **G**, 25 ng; **H**, 12.5 ng; **I**, 6.25 ng; **J**, 3.1 ng; **K**, 1.5 ng; **L**, 0.75 ng.

Figure 2.7 Sequence of border regions of pSPN1 derived fragment of pWB146

Sequence data that were in the slightest doubt are shown in lower case. Reverse complementary sequence to the ORI primer (used later, Chapter 4) is underlined.

A. Sequence using the 1212 primer.

```
    ClaI
  Sau3A
1  GATCGATTCA GGCGACCTCA AGGCGGACGG CAAACTTCCG TCCACGCGGG

51 TCCTGATGGA GACCTACGG
```

B. Sequence using the 1233 primer.

```
    Sau3A
1  GATCTCGACA TCGAGGGCGA CGGCcAGGCG GTGCTCGCcG CCCTCGCCGT

51 CCGcCTCGGT CAGCCGTGgc CGGAGACGCT GACCGTGGCC ACCC
```

Table 2.3 Plasmid pWB150 mobility test

	GROWTH[a]							
	WITHOUT THIOSTREPTON				WITH THIOSTREPTON			
INOCULUM SOURCE	STREPTOMYCIN CONCENTRATION ($\mu g ml^{-1}$) 0	4	16	128	0	4	16	128
S.lividans TK21	G	G	WG	-	-	-	-	-
S.lividans TK24	G	G	G	G	-	-	-	-
S.lividans TK24 (pWB150)	G	G	G	G	G	G	G	G
S.lividans TK21 (pWB150)	G	G	WG	-	G	G	WG	-
S.coelicolor J1501	G	G	G	G	-	-	-	-
S.coelicolor J1501 (pWB150)	G	G	G	G	G	G	G	G
S.lividans TK21 (pWB150) +*S.lividans* TK24	G	G	G	G	G	G	WG	-
S.coelicolor J1501 (pWB150) +*S.lividans* TK24	G	G	G	G	G	G	WG	-

[a]Spores from each inoculum source were germinated on M3/6.5 medium both in the presence and absence of thiostrepton (50 $\mu g ml^{-1}$) and with different concentrations of streptomycin present. Growth was recorded. G, growth; WG weak growth; -, no growth.

mechanism had been affected. Furthermore, it was possible that the results of Smith (1991) were erroneous, as it has since been found that the use of the streptomycin resistance marker carried by pFull for plasmid selection when low copy number plasmids are used is problematic. This is because the antibiotic resistance generated is relatively weak, and consequently a low background growth of plasmid free mycelia results (A. Smith, personal communication).

The copy number of pWB150 in *S.coelicolor* J1501 was examined (section 2.2.6). The results are shown in Fig. 2.6. From this it can be seen that the intensity of the band in lane C (0.65 µg loaded) is roughly equivalent to that of the pWB150 band in lane I (6.25 ng loaded). For bands of equivalent intensity,

$$\frac{(\text{Plasmid size})(\text{copy number})}{\text{Chromosome size}} = \frac{\text{Plasmid DNA loaded}}{\text{Total DNA loaded}}$$

From the sizes of the chromosome (8000 kb) and plasmid (13 kb),

$$\text{copy number} = \frac{(\text{Plasmid DNA loaded})(8000)}{(\text{Total DNA loaded})(13.0)}$$

Therefore in *S.coelicolor* J1501 plasmid pWB150 appears to have a copy number of approximately 6/chromosome. Because the band in lane C is lighter than that in lane H but darker than that in lane J, it can be said with certainty that the plasmid has a copy number of less than 12 but more than 3/chromosome.

2.3.14 Sequencing of the borders of the pWB150 origin of replication

A short section from each of the outer limits of the pSPN1 origin of replication bearing DNA fragment used in the construction of pWB150 and pWB151 was sequenced. This was for two reasons. Firstly it enabled DNA sequence database searching to determine if similar plasmids had been isolated. Secondly it allowed the sequence of the region upstream of the *mel* reporter operon in pWB150 and pWB151 to be determined. Subsequently a DNA primer to facilitate the sequencing of cloned promoters could be created (Chapter 4). It should be noted that the sequence was determined from one DNA strand only, so these data must be viewed with caution, especially as G+C rich sequence is prone to generate spurious data (USB Sequencing manual). The results are shown in Fig. 2.7. A database search using the FASTA program (Genetics Computer Group, Wisconsin, USA; Pearson and Lipman, 1985) to search the GenBank and EMBL databases (release numbers 84 and 39 respectively) was conducted. No sequence with extensive homology to either of the two sequenced fragments of the pSPN1 origin of replication was identified.

2.4 Conclusion

In retrospect, it is clear that the decision to utilise only the *melC2* gene as a reporter gene was not the correct one. The low levels of melanin production achieved meant that only strong promoters could be detected. However in defense of this approach it must be pointed out that at the time of creation of these plasmids the function of the *melC1* gene was poorly understood, and it was felt that, as far as possible, a promoter probe should not incorporate factors that may have unknown effects. However, both pWB105 and pWB133 function as promoter probes, although they were not, in the end, used as such in the work reported here.

The attempt to create a reliable reporter system that operated by translational fusion was unsuccessful as it was never shown that transcription from upstream promoters was

eliminated. It was therefore possible that the cloning of ribosome binding sites alone would result in melanin expression. However, the plasmid pPS9 showed clear temporal regulation and was retained for further analysis (Chapter 4).

The promoter probes pWB150, pWB151 and pWB152 should prove to be useful additions to the tools available for *Streptomyces* genetic research. They do not suffer the disadvantage of containing TTA codons, and the substrate is cheap and freely available. They do, however, have some disadvantages. The fact that the substrate is a useful metabolite precludes the use of the *mel* reporter operon for study of certain regulatory pathways such as nitrogen metabolism. Secondly, many streptomycete species produce melanin naturally. Again, this prevents the use of the promoter probes, however, *melC* mutants can be isolated (Hinterman *et al.*, 1989). Finally, it must be pointed out that because the natural function of the *mel* operon is unknown the possibility that the activity of the promoter probe is affecting the system being examined cannot be excluded.

CHAPTER THREE

COMPUTER ASSISTED ANALYSIS OF REGULATORY AND PROTEIN CODING DNA SEQUENCES

Abstract

The possibility that DNA sequences recognised by DNA binding proteins are avoided in protein coding regions of the bacterial genome is discussed. On the basis of this premise an attempt to identify DNA-protein recognition sites, using *Escherichia coli* as a model, was made. The frequency of occurrence of all possible octamers of various classes of DNA sequence data was catalogued. From these data the frequencies of occurrence of monomers to heptamers was determined. The data from the six different frames of protein coding DNA were compared and the correlation coefficients derived. In conjunction with scatter graphs, this information showed that unexplained relationships exist between the various frames. Similar studies were conducted using *Bacillus subtilis* and *Streptomyces* species sequence data. The results indicated that any possible sequence bias due to the avoidance of protein binding sites would be difficult to identify. In order to classify and identify *Streptomyces* promoter sequences, short oligomers that were over represented in a database of *Streptomyces* promoter regions were identified. These sequences, and others that were selected on the basis of the characteristics of known promoters, were tested to determine if they were found predominantly at particular distances from the transcription start site. In several cases positional preference was recorded. This allowed the identification of potential promoter core sequences. In some cases these may define novel promoter classes. One hundred and fifty three *Streptomyces* promoters were listed and grouped on this basis. A new and extended consensus sequence for the *Streptomyces E.coli* σ^{70}-like promoters was determined. It showed differences from that of *E.coli*, both in sequence and in the spacing between the -35 and -10 regions. A novel method of identifying the protein coding regions of *Streptomyces* DNA was developed.

3.1 Introduction and rationale of approach

3.1.1 Introduction

A preliminary objective of the work described here was the isolation of developmentally regulated promoters and determination of the sequences (possibly novel in type) that direct their transcription. Such studies typically involve some form of transcript mapping. However, since this work was initiated, many other researchers have isolated developmentally regulated promoters, and yet no obvious consensus sequences, or novel promoter classes have emerged (section 1.2.3.1). To simply add to that list would be a costly and largely futile exercise. Furthermore, the large numbers of promoters needed for any realistic attempt to determine a consensus sequence (section 1.7) were difficult to isolate (Chapter 4).

A more useful and generally applicable approach involves the classification of *Streptomyces* promoters, with particular emphasis on determining the sequences that define them. Such an analysis, if successful, could then be applied to both the newly and previously isolated promoters. There have been no reports of any formal attempts to classify *Streptomyces* promoters, based on anything other than a loose (but undefined) similarity to that of the *E.coli* major class of promoters (Strohl, 1992; section 1.2.3.1).

3.1.2 Rationale for the use of a protein-coding DNA sequence database

The interaction between nucleic acids and proteins must be one of the most ancient and fundamental features of living systems. However, the study of this interaction has concentrated mainly on physical structure and gene regulation. Little consideration has been given to the evolutionary constraints imposed on the primary structure of DNA by DNA-protein interactions. It is possible that a case for the existence of such constraints can be made, with the argument as follows.

It is commonly accepted that a DNA-binding protein can, in many cases, act at a number of different sites where it displays

different activities. Part of the difference in activity is attributable to the fact that the sequence of the sites themselves show differences from the consensus, and from each other. Furthermore, different regulatory proteins often act at the same locus and so must be able to accommodate the overlapping recognition sites of other DNA-binding proteins (Collado-Vides *et al.*, 1991). These two factors dictate, therefore, that DNA-binding proteins must have a certain degree of flexibility in their recognition of binding sites, without which they would be unable to perform their function. Such flexibility ensures that there are a large number of potential subtypes of binding sites for any particular DNA-binding protein.

It is probable that non-functional DNA-binding sites (pseudosites) could appear by chance within the genome of an organism during evolution (Berg and von Hipple, 1987). The natural flexibility of sequence-specific DNA-binding protein interaction can only serve to increase this possibility. The chances of pseudosites occurring would also be higher for DNA-binding proteins that require only short DNA recognition sequences. The *E.coli* RNA polymerase, which has been shown to be functional with only an 8-base (and possibly shorter) consensus -10 sequence (with no associated -35 region) is an example of this (Kumar *et al.*, 1993). Furthermore, the fact that the spacing between the -35 and -10 regions of promoters is variable will also increase the possibility that promoters will appear randomly throughout the genome.

It is logical to argue that pseudosites may be detrimental to the cell by interfering with the transcription or replication apparatus or by interfering with true binding sites. Pseudosites may also place a metabolic load on the cell, either by causing transcription from functionless promoters, or by titration of regulatory proteins, both of which will result in an unnecessary expenditure of the organism's resources. It is therefore likely that such inappropriately placed DNA-binding sites would be eliminated by natural selection. Given the codon redundancy of the translation apparatus and the flexibility of DNA recognition sites, pseudosites could be eliminated without

having major consequences for the other functions of the locus in which they were found.

Support for the preceding postulation comes from the work of Alvarez *et al.* (1994). These authors have inserted random genomic fragments from the yeast *Sachcaromyces cerevisiae* into a promoter probe vector and tested for expression of the reporter gene in *S.lividans* and *E.coli*. Remarkably, 10% of all the inserts were able to drive transcription in these bacteria, implying that bacterial promoter-like sequences are found in yeast. Sequencing of the certain of the genomic fragments revealed that they consisted of pieces of protein encoding DNA. There is no known reason why yeast DNA should carry *E.coli* promoters, but, equally well, there is no reason why it should not. This work indicates that not only is it possible for active promoter sequences to appear at random in the chromosome during evolution, but that these are not eliminated by selection in organisms where they confer no disadvantage.

Such selection would ensure that the occurrence of protein-DNA recognition sites is biased, and subsequently the occurrence of those short overlapping sequences that make up binding sites would also be biased. This means that it may be possible to analyse a genomic database in such a manner that sequences that have meaning in terms of DNA-protein interaction can be recognised.

To create a database which will enable sequence bias to be analysed involves the cataloging of oligonucleotides and recording the frequency of occurrence of each of these.

The length of the oligomer to be considered is limited by the size of the database that is to be searched. It has been calculated by Von Hipple and Berg (1986) that for an organism with the genome size of *E.coli*, a 12 base sequence is likely to appear only once, if the order of the bases in the genome was strictly random. This means that even if the entire genome of *E.coli* had been sequenced, a database search for 12-mers would be statistically meaningless. DNA binding sites however, are in many cases, much longer than this, so a search for the complete binding sequence would be impossible. However, it is commonly accepted that most protein-DNA binding sites consist of a

highly conserved core sequence in association with more variable and less conserved additional sequence. The consensus core consists in most cases, including that of promoters (Hareley and Reynolds, 1987), of just a few nucleotides, and thus database searches for this type of core sequence could be conducted.

The first question to be resolved for this approach is which sequences are to be included in the database. Clearly, only sequence data for a single species or very closely related group of organisms should be included. DNA sequences that encode proteins but not the regulatory regions that govern their expression would be ideal candidates for inclusion in the database, as these regions contribute the majority of sequence data in existing databases and are likely to contain relatively few DNA-protein recognition sites. Thus sequences that can act as protein binding sites should be under represented. The use of protein coding sequences will, however, have disadvantages. There are constraints on the sequence that may have little to do with DNA-protein interaction, but are as a result of the translation machinery. These include codon bias (Andersson and Kurland, 1990), amino acid composition or protein structure. However, a sequence that shows a bias in occurrence because of these factors is only likely to be biased in a frame specific manner, and will not necessarily be biased in the other frames or on the opposite DNA strand. The selective pressure that acts to eliminate pseudosites would, in theory, exert its influence on the DNA regardless of coding frame or DNA strand. This in turn implies that there will be some correlation in the frequency of occurrence of certain oligomers between different frames.

It has been noted by Rogerson (1989) that in *E.coli* protein coding DNA, the frequencies of tetramers in the sense and antisense strands show statistically similar distributions. This indicates that if a particular tetramer is rare in the protein coding sense strand it is likely to be rare on the opposite strand. Recent work by Yomo and Urabe (1994), using all sequenced bacterial genes and testing trimers, demonstrated that this correlation existed only between the sense coding frame and the antisense to that frame. This means that the

frame that is antisense to the coding frame displays a triplet bias similar to that of the coding frame. An obvious bias in the distribution of stop codons was also observed. The authors have shown that these factors result in long open reading frames (ORFs) existing on the antisense strand, and that these ORFs have a similar bias to the codon bias of the coding frame. They postulate that the antisense ORFs are the result of selection and enable the creation of ancestral genes during the course of evolution.

However this pattern arose, any bias that is as a result of the avoidance of inappropriate DNA binding sites would, in theory, be superimposed upon it. Unfortunately, Rogerson (1989) and Yomo and Urabe (1994) tested only tetramers and trimers respectively. Furthermore, in the first case a small (343267 base) *E.coli* database was used, while in the second sequence data from very different species were combined.

It is probable that if the avoidance of certain protein-DNA binding sites affects the distribution of oligomers, this would only be obvious for longer oligonucleotides. Furthermore, it should manifest in the form of some degree of correlation between the frames that do not show correlation when shorter oligomers are compared.

Of all the prokaryotes, most is known about *E.coli*, with regard to total DNA sequence and DNA-protein binding sites. Thus the above theory was tested in this organism.

A second use for a protein coding sequence database could be the determination of protein coding regions in stretches of DNA sequence (see section 3.1.4).

3.1.3 Rationale for the use of a promoter region DNA sequence database

The argument made in section 3.1.2 can be reversed if the database to be used consists of non-coding promoter regions only. In this case it would be expected that certain oligomers would be over represented because they define DNA-protein binding sites. In the case of those sites that represent promoters, these short sequences would be predominantly found

clustered at specific distances from the site of transcript initiation. By searching for short sequences that display these characteristics, it may be possible to define different classes of promoters without performing complex biochemical tests which typically involve site directed mutagenesis and/or *in vitro* transcription.

Because the streptomycetes are known to have different classes of promoters which have yet to be classified in any meaningful way (section 1.2.3.1), that genus was selected to test this theory.

3.1.4 The possible use of an oligomer database for protein coding sequence recognition

There have been several approaches to determining, on the basis of sequence data alone, the regions of DNA that encode proteins. One method uses codon bias (Andersson and Kurland, 1990; Griboskov *et al.*, 1984). Another is the FRAME analysis described by Bibb *et al.*(1984), which is based on the third position periodicity of nucleotides found in the protein coding regions of organisms whose genomes are either A+T or G+C rich. Both of these methods are used in conjunction with analysis of potential start and stop codons to determine possible ORFs.

In a G+C rich organism these approaches can have weaknesses. Codon bias based searches do not always give obvious results as the codon bias tends to reflect G+C bias. Furthermore, any correlation pattern similar to that discussed above (Yomo and Urabe, 1994; section 1.3.2) must exacerbate this. Again, this can make the correct DNA strand difficult to determine. In both cases search windows of over 50 bp are necessary and so very short sequences are difficult to analyse and small protein coding ORFs can be overlooked. Finally, The G+C richness of streptomycetes dictates that stop codons are rare, so all frames tend to have large potential ORFs, particularly if GTG is considered as a potential start codon as is often the case in streptomycetes (Chater and Hopwood, 1993).

It should be possible to use the oligomer frequency catalogue described above (section 3.1.2) for coding sequence limit studies. This would involve tabulating the data in a frame

specific manner and plotting the frequencies of occurrence of oligomers along the DNA sequence. Oligomer bias is a consequence of a number of factors, including codon bias, G+C ratio, amino acid preference and the likelihood that certain amino acids are found adjacent to each other. Thus all these factors would be used to determine the range of the protein coding sequence. Such an approach may be more accurate than standard methods, and allow the correct strand to be identified, using smaller search windows (which will facilitate analysis of short sequences).

3.2 Methods

The computer programs described below were kindly written by Brendan Babb (Department of Medical Microbiology, University of Cape Town) to specifications supplied by the author. All data analysis systems were tested with known data to confirm the veracity of the output. Programs were written in Borland C++ and run under MS DOS. The programs used, and information about them, are supplied on floppy disc (Appendix G). The programs are executable under MS DOS. All data used are also supplied on the disc. Instructions for use are under the MS Word5 file README.DOC.

3.2.1 Creation of catalogues of oligomer frequencies

DNA sequence data of various classes, from different sources (see below) were assembled. From these sequence data, the frequencies of occurrence of each of the 65536 different possible overlapping octanucleotides was catalogued. An eight base window was moved, one base at a time, down the sequences, and at each position the octanucleotide sequence was determined. A tally of each of the different octanucleotides was generated. For each possible octamer the frequency of occurrence (F) was calculated in normalized form, by dividing the actual number of occurrences by the expected number of occurrences when assuming completely random sequence. The number of occurrences and frequencies of monomers to heptamers were determined from the octamer data.

$$F = \frac{(\text{number of occurrences})(4^{\text{oligonucleotide length}})}{\text{total number of nucleotides in database}}$$

The fraction of the total number of nucleotides in the database (or subset of the database, see below) which consisted of each of A, G, C and T residues was calculated from the F values of the monomers (%A, %G, %C and %T respectively). The frequency of each oligomer, corrected for the nucleotide ratio (NCF) was then calculated using the number of times each base was found in the oligonucleotide under consideration (XA, XG, XC and XT).

$$NCF = \frac{F(XA)(\%A)(XG)(\%G)(XC)(\%C)(XT)(\%T)}{\text{oligonucleotide length}}$$

Five different databases were created, each of which consisted of a catalogue of octamer frequencies found in five different collections of sequences. These were termed Eco, Bsub, Strep, Spro#1 and Spro#2 (created from *E.coli*, *B.subtilis*, *Streptomyces* species, *Streptomyces* promoters and a subgroup of *Streptomyces* promoters respectively). A summary of information regarding each database is given in Table 3.1. It should be noted that the figures for the number of nucleotides in each database do not include the last seven bases of each sequence entry and are equivalent to the total number of octamers counted. This means that the frequency calculations assume that the nucleotide ratio in those last seven bases are similar to the ratio found in the rest of the database. Even if this were not so the effect is likely to be small and the databases were only used to identify gross biases.

Only one strand of the DNA of each sequence was entered in the database. In the case of Genbank sequence data the strand used was that recorded in the Genbank database. The descriptive title for each Genbank sequence entry was used to eliminate repeats, as far as possible. Each sequence entry was analysed individually, without concatenation. The number of octamer

Table 3.1 Name and details of sequence databases and subsets

NAME OF DATABASE	Eco	Bsub	Strep	Spro#1	Spro#2
ORGANISM	*E.coli*	*B.subtilis*	*Streptomyces* species	*Streptomyces* species	*Streptomyces* species
NUMBER OF SEQUENCES ENTERED	608	224	286	106	37
SOURCE OF SEQUENCE DATA	Genbank[a]	Genbank[a]	Genbank[a]	Various[b]	Various[b]
NUMBER OF NUCLEOTIDES					
TOTAL	1315773	514872	343313	3959	1567
NON CODING	345759	130672	102546		
CODING, TOTAL	885157	347954	215006		
CODING, FRAME 1	295406	116129	71758		
CODING, FRAME 2	295263	116076	71714		
CODING, FRAME 3	294488	115749	71534		

[a]Release 73, Wisconsin, USA.

[b]See Table 3.3.

occurrences for each database was stored in five subsets, each of which represents a different type of sequence data that is found within each database. The names, abbreviations and descriptions of the subsets are shown in Table 3.2.

The databases Spro#1 and Spro#2 were created using *Streptomyces* promoters that have been identified by transcription mapping (Table 3.3). Promoters number 154 and 155 are from *B.subtilis* and were not included. Where there existed a number of different overlapping promoters for a given gene eg. *amy*-p1, *amy*-p2 and *amy*-p3 (Table 3.3, promoters number 7, 142, and 134), these were entered in the database as a single contiguous sequence. As only one strand of the DNA was entered, in the case of overlapping divergent promoters one of them was entered as reverse complementary sequence. Only sequences between positions -53 and +1 as defined in Table 3.3 were used. No known protein coding sequence was entered. The Spro#1 database consisted of all promoters listed in Table 3.3 which, as far as could be established at the time of the creation of the database, fulfilled these criteria. The promoters included in the Spro#2 database fulfilled both these and other criteria (see Results and Discussion, section 3.3.6).

3.2.2 Oligomer frequency distribution

A catalogue of oligomers (in a specified database subset), grouped by their frequency of occurrence, was created. Each group included all the oligomers that displayed an F value that was within a particular 0.1 range of frequency. Groups were from the value 0 (inclusive) to 10 (exclusive). The total number of oligomers in each group was plotted against the frequency as a bar chart.

3.2.3 Scatter graphs and correlation coefficient

The frequency of each oligomer in one database subset was plotted against the frequency of the same oligomer in a different subset to generate a scatter graph.

The correlation coefficient (r) was determined by the following formula.

Table 3.2 Database subset names and descriptions

Database subset	Name	Description
Total	T	Oligomer frequencies were derived from all DNA sequence used in the database in question.
Non-coding	NC	Oligomer frequencies were derived from all DNA sequences not identified (on either DNA strand) as protein coding sequences.
Total coding	C	Oligomer frequencies were derived from all DNA sequences identified as protein coding sequences. Protein coding regions that overlapped or were subject to frame-shifting were excluded totally, as were sequences that were incomplete at the N-terminal end (ie lacking a start codon) or in which the coding region sense strand was not that strand entered in the database. Start codons but not stop codons were included in this subset.
Total coding (reverse)	CR	As for subset C except that each oligomer was assigned the frequency of the reverse complementary oligomer in subset C.
Coding, Frame 1	F1	As for the total coding (C) subset, except that only oligomers that initiate at the first nucleotide of a codon were used. This is always the coding frame data set whatever the oligomer length because it is defined by the position of the first base of the oligomer.
Coding, Frame 2	F2	As for the F1 subset, except that only oligomers that initiate at the second nucleotide of a codon were used. This is always the coding frame +1 subset.
Coding, Frame 3	F3	As for the F1 subset, except that only oligomers that initiate at the third nucleotide of a codon were used. This is always the coding frame +2 subset.
Coding, reverse complement F1	R1	Each oligomer frequency was taken as the frequency of the reverse complimentary oligomer in subset F1. This is equivalent to using data from the DNA strand that was not entered in the database. Notice that here the frame is defined by the position of the last base of the oligomer when considering the DNA strand that was entered. This means that the coding frames of the F type and R type subsets are not always related in the same way, but in a manner dependent upon oligomer length.
Coding, reverse complement F2	R2	Each oligomer frequency was taken as the frequency of the reverse complementary oligomer in subset F2. See comments for R1 subset.
Coding, reverse complement F3	R3	Each oligomer frequency was taken as the frequency of the reverse complementary oligomer in subset F3. See comments for R1 subset.

Table 3.3 *Streptomyces* promoters used in cluster analysis

	NAME OF PROMOTER	INCLUDED IN SPRO[a] #1	#2	SEQUENCE OF PROMOTER INCLUDED IN SPRO#1 AND #2 AND/OR TESTED FOR OLIGOMER DISTRIBUTION[b]	PPCS TYPES FOUND[c]	REFERENCE
				-50 -40 -30 -20 -10 +1		
1	XP55-P	I	N	CTCCCCACCTGGC**TTGAC**GCTTTATTGCGAGTGATG**TGCAAT**AGCTGC	C, G(18)	11
2	mmr-p	I	N	CGGCCATCAAAGT**TTGAC**AGCCGTCGTCATATGAGC**TTCAGT**GAGAACG	C, G(18)	11
3	gyl-p1	I	N	GATAACACAGCTC**TTGAC**GCGCGCGTGACGTCGAAC**GAGACT**CGCGTCCAT	C, G(18)	11
4	endoH-p	I	N	ATTGACTGA**TTGAC**GCGCTTCCGGCGGGCAGGGGAGGCACGG	C?	11
5	brpA-p1	I	N	CCTTGCCTCGACTGG**TTGAC**TCTCTACACTGAAGATTTACATCTGATT	C?	11
6	srmX-p	N	N	CTGACCACCCCTCGTTCC**TTGAC**CCCCGGCTCGGCGTGAACTGTTGGGTGTCC	C	3
7	amy-p1	I	P	TCTTGCAGAAAGCC**TTGAC**GCTGTTAACGACATGAACGGCAGGCTCCGGTA	C	11
8	actII2,3-p	I	N	CCACCGTTTC**TTGAC**GCGACCACCGTTCCATTGTAGAACGGTGGTCGT	C ,D	11
9	whiB-p2	N	N	CAGATCTCGGC**TTGAC**TCGCCCGGAGCAGCACACT**TGTAATT**TCACT	C, A, G(19)	10
10	amy-p(Stt)	N	N	TGACAGGTTCATGTCCAGAGTA**TTGAC**CGCCCGCAACGGCGCTCGT**ACGGT**CGGC	C?, A? G?(18)	7
11	rrnD-p2	I	N	CCCGCAAGAGCCG**TTGAC**ACGGAGCGAGCGGGGAGG**TAGATT**CGAACA	C, A, G(18)	11
12	choP-p	I	N	GGAACGATCTCG**TTGAC**AGCCTTCACATCGCCTCCA**TACGGT**CATTTC	C, A, G(19)	11
13	amlV-p	I	N	GGCGTCCGGAGGG**TTGAC**CGGGCGTCGGGGCACTCG**TACGGT**CACGGCT**GAAAA**	C, A, G(18)	11
14	aml-p	I	N	CGTGTCCAAAGGG**TTGAC**CGCGGGTACCGCTCGCTC**TACGGT**CTGCTTCGC	C, A, G(18)	11
15	ermSF-p	I	N	GAGGAAAAGGAG**TTG**CGGGCGAGCCGTGCCCCTGCT**TACCGT**TCCCG	A, G(19)	11
16	npr-p	I	N	TGCATGCCGAATG**TGAC**ATGCGCAATCCATGTGGCG**TAAAGT**CCCGGTG	A, G(19)	11
17	ßgal-p	I	N	TCCGACGGGGTAA**TTGAT**TCGGTTGTGTTCGGGTTC**TAGGGT**GACCCGT	A, G(18)	11
18	actII(1)-p	R	N	GGTCCTCGACTA**TTG**GTCCACGAACGACCACCGTTC**TACAAT**GGAACG	A, G(19)	11
19	pARC-p	I	N	GACACGCTCCCC**TTG**CCGGTGCTAGCGCGACCGCGC**TAGCGT**GGTCGG	A, G(19)	11
20	orfRP-p	I	N	**TCGAG**GCGGTCGAAGACCTCGGCG**TACATT**CCGTGA	A, G(19)	11
21	dagA-p4	I	N	CAGCCGTACCGA**TTGT**CACCCTGCGACACTCCGCTG**TAGCATT**CGGGAAA	A, G(19)	11
22	dnaA-p2	N	N	GGCTGGGGACAACAAC**TTGA**ACCCTACCCGTCGCGCTGAC**TACCGT**GACTTGGA	A, G(19)	14
23	aacC7-p1	P	N	CAGAGAAATACGGTGCCGG**TGAC**CGTGAGCGAC**GGATACCTT**CCCGTCCA	A, D, G?(18)	11

[a] Symbols defining whether the sequence was included in the Spro#1 or Spro#2 database; I, included; P, partially protein coding sequence, non-coding parts included; C, completely protein coding sequence not included; R, included as reverse sequence; N, not included for other reasons.

[b] Bold sequence symbols, nucleotides that define PPCS sequence; underlined, potential -35, -10 and extended -10 regions; double underlined, transcription start point.

[c] Symbols defining PPCS classes; A-H, PPCS Class; (), encloses spacing in number of bases between the -35 and -10 regions; ?, PPCS is close to, but not within the main body of the cluster.

[d] References. 1, Agnell *et al.*, 1992; 2, Forsman and Granstrom, 1992; 3, Geistlich *et al.*, 1992; 4, Guilfoile & Hutchinson, 1992; 5, Gunter *et al.*, 1993; 6, Ishikawa & Hotta, 1991; 7, Ogawara *et al.*, 1993; 8, Paradkar *et al.*, 1994; 9, Perez *et al.*, 1993; 10, Soliveri *et al.*, 1992; 11, Strohl, 1992 (and references therein); 12, Tan & Chater, 1993; 13, Vujaklija *et al.*, 1993; 14, Zakarzewska-Czerwinska *et al.*, 1994; 15, Leskiw *et al.*, 1993; 16, Chater, personal communication; 17, Westpheling *et al.*, 1985.

Table 3.3 (continued)

24	otrA-p	P	N	GTCGCCGTCAAGCCCTGACCCTGCGTGGCCGCCCTTGCTACCGTGATCA	A, D, G(19)	11
25	tcmA-p	N	N	TGACACCTGCATGCGATTGTCAAATGTCACTGACTGCTGTTAGCTTCGGGGCC	A, D, G(19)	4
26	srmB-p	N	N	TGCGACACGCATACCGTTGCGCGGCACCGGGTAGCACGGCTATAGTTCTTCGC	A, D, G(19)	3
27	des-p	N	N	AGTTGACGGGAATCACCCCCTGTGGACTTCCTCAGGTCGACATTAGGTTAGGCTC	A	5
28	kgmB-p	I	N	CCACGCGGGGACGTCCGAGGGAGCGGGCGAGCCTCGCTAGGCTGGACG	A	11
29	orfI-p1	I	N	CCGTCCGTGGACGGTGGTGGGGCCGAGGACCGCGGGATACGGTCGCCGGCC	A	11
30	rep-p2	I	R	TAAAGCTTTGGTAACGCACCCAGCCTACTCACGTGAGTAGCTTGGAGCG	A	11
31	tsr-p1	I	N	TGATTGCCGGTCAGGGCAGCCATCCGCCATCGTCGCGTAGGGTGTCACA	A	11
32	bar-p2	I	PR	ATTCCCAACCTGCCCCTCGATTTTTCTGATCATGCAGTACCCTGTGCCGCCAC	A	11
33	dagA-p1	I	N	TTTGGCGCCCAGGGTCTGCGGAAGTCATTGCCAAATATAAGATTCTTCA	A	11
34	nshR-p	P	N	CCGCTCTGGTGGCCGGGGGCGCAGGCTCCCGGCCACTAGACTGCGCGCA	A	11
35	drrAB-p	I	N	GAAGCCATCGCGGCCATGAAGTGTCCTCATTGGGGGCTACGGTACTCAAC	A	11
36	glnA-p	I	N	GAAACAAATGGGTCACGCCCGAGAAATCACCCGTCCCTAGGGTCGAGGAA	A	11
37	sta-p	I	N	CCGGCTGAAACCAGACCTCACCGGGGCAGGCCGGGCATAGCCTCGGGTCATG	A	11
38	korA-p	C	N	GTCCATACTGTGGTCGCACAGTTGCTGTGTCAAGGCATACACTGTGCTAG	A	11
39	gyl-p2	I	N	CGGTCGGCATTGTCGAACACCTACCGGCAATACGCGTTAGAGTGTCCACAGTG	A	11
40	whiG-p	N	N	CCCACCCCCTCCTTCGCACACCGCGACAGTCCAGTCACGCTACGCTCACGATG	A	16
41	cp2	I	N	TTTGCCGAAATGGCTCAGCCTACTCACGTGAGTAGTGTAGAGTCGGCTACG	A	11
42	orf-p1	I	N	ACCGCGTGCACCTGCGATCGCCGATCAACCGCGACTAGCATCGGGCGCA	A	11
43	bldA-p	N	N	GCCTGACCCGTAACTGATGCACCCTCAGGCGCCAATAAAATAAGACAACATCGG	A?	15
44	orfJ12-p	I	N	AACTTTGATGGCCGTCAACATTTGATGGCTGTCGTACCATGGTGGAAC	A, D	11
45	gal-p1	I	N	TTGTGATGTGACAGGGGGGTGGTGGGTTGTGATGTGTTATGTTTGATTG	A, D	11
46	ermE-p1	I	N	GCGATGCTGTTGTGGGCTGGACAATCGTGCCGGTTGGTAGGATCCAGCG	A, D	11
47	aphD-p1	I	N	ACAGCTTTACTTGGCCGTTGCCCGGATGTCCGGGTGCTACTATTCGCGAA	A, D	11
48	aphD-p2	C	N	CCGTCCGTGCGGCTCACCGAGGACGGTCGGGCGGTGCTACGTTGGCTGAA	A, D	11
49	rrnD-p1	P	N	ACAGCCGCCTGATGTGCATCCACCCCTGCGAGCTGCTAGTGTCCTCTTC	A, D	11
50	strb-p	I	N	GCCGCGGTGGACATATGCCCGAGCGAAGCGGCGCTGCTAGCCTGCGATGA	A, D	11
51	pCAT-p	I	N	GGCGGAAAATCGCTACGGCCCGCACACCGGCGGCTGATATGCTGAGCCGA	A, D	11
52	rrnD-p3	I	N	AAACGAAGGCCGGTAAGACCGGCTCGAAAGTTCTGATAAAGTCGGAGCC	A, D	11
53	rrnD-p4	I	N	GGAAAGCGCCGAGGAAATCGGATCGGAAAGATCTGATAGAGTCGGAAAC	A, D	11
54	casA-p	I	N	AGCGTGCGCGAGGGCTTCACCGCATGGCGCCCCGCCGTTACCGTGCCCT	A, D	11
55	sphP-p	R	PR	CCGTCCACGGACGGCGGGAGCCTGTACCGATCTCCGATAGGTTGCGCGA	A, D	11
56	rep-p1	I	R	ACCCAGCCTACTCACGTGAGTAGCTTGGAGCGTGGGCTAGGGTGAGCGA	A ,D	11
57	srmR-p	N	N	TCGTTCCGCCGGAAATCACGGTGTGGCCCCGGGCCACCGGGTAGCTTATGCCT	A, D	3
58	kan-p	N	N	CGCCGCCCAGCAGCGGAGGCGCGATCCGTTCGTCGTCCGTTATGATGATGGTC	A, D	6
59	ORF2+3-p23	N	N	GCGATGACCCGCGGAGCCCCGGAGCGCGGGCCCGTTCGGGTAACGTCGTAAGGACG	A, D	1
60	gylR-p	I	N	CCCGCCGACCATTGGCGGGAGGTCGGCATGGACCGGTAGTGTTCGGCATT	A, D	11
61	tylF-p	I	N	CGGGCTCCTCGTTCCGGCGCGCGCCGATAGCGTCCGTCCTC	A, D	11
62	orf3-p3	N	N	CGGGGCAGCGCATCGACCTGGACTGAGTCCTGCCTCGGGTACGGTTGCCCTTAG	A, D	1
63	tra-p	I	N	CACAGTGTATGCCTTGACACAGCAACTGTGCGACCACAGTATGGACCTTG	C, B, G(18)	11
64	pabS-p	I	N	GGGGCATGGTACCCCCACATCTATTGAATCCGCAACGCGCAGTATCATG	B	11
65	pA2	C	N	GGTACTTCCGGCGCAACGTGCTGTCGTCCATGGGCGGCATCATGGCAGA	B	11
66	tcmR-p1	N	N	GTGACATTTGACAATCGCATGCAGGTGTCACCAGGTGGCGCAAGATAGGGTTG	B	4
67	whiE-p	N	N	CGCTTAACCCTCCTGGTTCGCGGTTCCCCGCGGCACGGCCCACGATCGTGCTT	B	16
68	cefD-p	I	N	ACGACAGGACTCTTGAAGTGCTCTTCGGCTGGTCTTCAGAATCTCTTCGCTATTTT	B?, G?(19)	11
69	korB-p	P	N	CAGCCTGAACTAGTTGCGCAGACTGACACAGTCGGTCAGGATGACTTCA	B, D, G(18)	11
70	redD-pr4	I	N	GAAGCACCCCACCGGGGTGCGACGGCGGTGGACCGGCATGATGGTGGC	B, D	11
71	octD-p1	R	N	CGCTGTGGGATCCGCCCGGCGGGAACACACCCGCGGGCATGATCACGGT	B, D	11
72	orf-p2	N	N	CGGTCAGATCCTCCCCGCACCTCTCGCCAGCCGTCAAGATCGACCGCG	B, D	11
73	aph-p1	N	N	GACGAAAGGCGCGGAACGGCGTCTCCGCCTCTGCCATGATGCCGCCCA	B, D	11
74	rph-p	I	N	GTCAAATCACCTAGGGAGAAGGTGCCTTCTCTGCCATGATGCCGACCA	B, D,	11
75	pac-p	I	N	CTCGCGCCCCCGCGGCCGCCAGCACCGGTGCCGCCACCATCCCCTGACCC	B?, D?	11
76	ermE-p2	I	N	ACGCGGTCGATCTTGACGGCTGGCGAGAGGTGCGGGGAGGATCTGACC	C, D, G(19)	11
77	bli-p	N	N	GCGTCCGCCGTGATGAGGAGTGCGCGGAGATCCATCGATGCCAGGGTTGAGCA	D	8
78	nsha-p	I	N	TCGGCCCGGGACGGGGGCGCGCGGGGCCGATCCGTGGCATTGTCGACGAC	D	11
79	dagA-p3	I	N	TACCTCCTGGAGCCTAGCTCCTCCTGCGCCGTGGAATGATCGTGCCAC	D	11
80	lip-p	N	N	AAGACCGGTGATACCGCCGATGCGCGTCCACGGGCCGCACCGTGAGAGTGGCGAC	D	9
81	pA1	N	N	TCCGCGCCTTTCGTCGGGGCCCGTAGGGGTTTCGGACATTCTTGTGCGGGG	D	11
82	cp3	I	P	AGCTGTGCGCCGGGTGCTCGAAGGTCGTGCCGACCGGTGAGGTTTCCGTCTC	D	11
83	kamB-p	I	N	ACGACGTGGTGCTCGACGTCGGCACCGGCGACGGGAAGCATCCGTACAA	D, G(18)	11

Table 3.3 (continued)

84	cdh-p	I	N	GACAATCGGCCTCGAAACTGGAACCTGTTTCAGTTAAGCTGCCCGTCA	A, D, E, G(18)	11
85	hrdB-p	I	N	CCACGCGGATTGGGCGTAACGCTCTTGGGAACAACACGATGACCTAAGA	B?, E, H(19)	11
86	kilB-p1	I	P	CGGGGCCTACCATTGCGCAACCTGACACGTGCGTGTCAGACTGATACACG	D, E, G(19), H(20)	11
87	sapA-p	I	N	ACATCTGCCAACGACGTAACAACCCCCGAAGGTGCAAAGGTCTCAACTGG	D, E, H(20)	11
88	dac-p	I	N	TCCTTCGCGTGACATGCAACCCATCTGCCCCTCCTGCGCGTAGAGATGGTGCC	E, H(20)	11
89	hrdD-p1	C	N	ACTCCGCGTCCGTGGCAACCCTCAGGCGGTACGGGCCGTCTTCAGGGT	E, H(19)	11
90	whiB-p1	N	N	TCGGCGCGTGTATCGGAACGGGATCGATCGCCGGGGCGTCCTTCCTGGT	E, H(19)	10
91	dagA-p2	I	N	CACGTGGGCGTTCCGGAACTTTTTGCACGCACGCGAGCTCTCGAATTTT	E, H(19)	11
92	blaF-p	N	N	GACCGAGTGACAGATGAAACCGGTGGGACGGGAGGGACCGTCGTCGTGGATGT	E?, H(20)?	2
93	ssi-p2	I	I	GCATGGGGTCGAGGTGAGTTTCCGGCCGGGGACTCGGCAGACTCCGGCACGCCG	F, G(18)	11
94	redD-pr3	I	I	GGTGGGGGCCGCAGATGTTTGCCCGTCGAGCCGAAAGAGGAAGATGAAC	F	11
95	orf1590-p1	I	I	GGAGAACTACGCTGTGTTTACTGGTGTTCTCGACAGGGGGGCATATTCCT	F	11
96	tcmR-p2	N	P	GGACTGCCGTCGTCGTGCGTTTCCGTGCTCATTTCAGCGTCCCCCTAACGATT	F	4
97	cp3'	I	I	TGCCGACCGGTGAGGTTTCCGTCTCCGAAGAGTCCGTTGAGGTTCCGGGC	F	11
98	saf-p	I	I	GGAAACGGTGGTCCGTTTCCCGCCCCTGCCCGTAGGCCGTGCGCGTCCCGC	F	11
99	cp3"	I	I	AAGAGTCCGTTGAGGTTCCGGGCCTGAGTATGACGGTCAGTCAGAAGAGC	F	11
100	mel-p1	I	I	CGGTCCGGGCCGATTTCTCCCCTTCTCCTCCGGTCGATAGGTATGCGGGG	F	11
101	aph-p2	N	N	GCGCGGTGGGGGATTCCGGCCGAACGCGCCGACGCCCATGTGACCGCCT	F	11
102	aacC9-p	I	I	GAAAAATTACTCGGTTACCTGACGCCCCGGCTCAGGAGAGCCTGCTAGCTA	F	11
103	xylA-p	F	PR	GCTGCTGACATCGGTTTCTCCCTCTTCCGCGGCTCAGCGGGGTGTGTCTCTT	F	11
104	tipA-p	I	I	CGTCCGGGCTTGCACCTCACGTCACGTGAGGAGGCAGCGTGGACGGCG	F	11
105	nshA-p2	I	I	CGGCTCCGGTGCAGGCATGATCCGCAAAAGGGTGGAAAGTCGAATTAC	G?(18)	11
106	brpA-p2	I	N	CGGATGACCTTGCCTCGACTGGTTGACTCTCTACACTGAAGATTTACA	G(19)	11
107	bla-p	I	I	GCGCACTCGGGTCTTCGACCGTATGTCCGCACCGGGACGAGAGTGTTCTCG	G(19)	11
108	afsR-p1	I	I	GCGTCTCCCACGGCTGACGTGGTCGGCATGAACAAGGCAAACTGACGTG	G(19)	11
109	pA5	C	C	CCGGTGAGCTGGTAGACGAAGGCGCCCGAGTCTCCTTCGTTCACTGCGT	G(18)	11
110	P_{TH270}	C	C	TGTCGGGGTTGTCGGGCTACGCGACGGGGGGCGGCGCCCGATGACGGCACCGGT		12
111	P_{TH4}	C	C	CGCCCTGGCGCTGCTGGTGGACTGGCTGGCGTCGCTGGCCGAACTGCTGCTGC		12
112	aacC7-p2	I	I	CCCGTGCGGCCCTGTCAGTTGCGCCTCCGCGATCGGTTACGCCCTGACC		11
113	tlrC-p	I	I	GCACCGGGCCTTGGTGATCACGTACGTTGACCTTCATGGGAACACTG		11
114	hrdD-p2	R	I	CGAGGGGCAGGTTGGGAATTCTGTCCGGATTCCAGTCGTTGTTTCCAT		11
115	pIJ101-pc	I	I	AGCTGGACGAGATTGAGAAGGAGGCTGCGCCCGCTGATCGCGGAGGCGG		11
116	kilB-p2	I	I	CGACTCAGCACCGCCGCATGACGGATGGCCCCGCCGGAAACCGGCGGGGC		11
117	actIII-p	I	I	CCCATCTCCCTTCGACCGCCGCTCGAGCCGGGCGCCCAAAGCTGAGCGCG		11
118	pA3	C	C	GCGTGCCACTCGTGGTGCGGGTACTTCCGGCGCAACGTGCTGTCGTCCA		11
119	dnaA-p1	N	I	GGACGTGACCTGTGCCCCTCGGAGGCAGGCGGCAGCAGCCATCGACAACTCGACC		14
120	tsr-p2	I	I	TTGGCTCGACGCAGCCCAGAAATGTATGATCAAGGCGAATACTTCATAT		11
121	est-p	I	I	TCACCTCCATTGGATTATCACGCTCCGTATGTGACTGAATGCAGAGCG		11
122	octD-p4	R	I	CGCATTTCGTTCCGCCCTCTCTGACGTCGCTGTGGGATCC		11
123	melC1-p	I	I	CGGAGCCAACCGTTCGCGGGAGTGCCCCACGGGCCGGTAGGCATGCGGAG		11
124	orf1590-p2	C	C	GGGTCAACGACCTCGGAGCACAGCGCGGTCTGACACTTCGCTACGACAAG		11
125	actI-p	C	C	CGTCACCAGTGCGACTTCGGAGTCCTGCGTGGCCATGTGTTCCCCTCCCT		11
126	gal-p2	I	I	CCTGGAACTTTTCACTTCCGCCGTACGTCCGGCAAGCTGAAGTTCCTCG		11
127	cp3"'	I	I	GGGCCTGAGTATGACGGTCAGTCAGAAGAGCTACACGCCCGTTGAGGGC		11
128	rep-pA2	R	I	GTGCTTCGCTCACCCTAGCCCACGCTCCAAGCTACTCACGTGAGTAGGC		11
129	xylB-p2	I	I	GTAGTTCATGCCGCGGCTCCTTGCTCGCTGAGGCTATTTCGTCATGGCC		11
130	pA4	C	C	GTTCACTGCGTGCCACTCGTGGTGCGGGTACTTCCGGCGCAACGTGCTG		11
131	afsA-p	I	I	GGGGAGTTATGCCCGAAGCAGCAGTCTTGATCGATCCGGTGCCGACTA		11
132	octD-p2,3	R	I	CCGCCCTCTCTGACGTCGCTGTGGGATCCGCCCGGCGGGAACACACCCGCGGGC		11
133	rep-pA1	R	I	GTGAGTAGGCTGGGTGCGTTACCAAAGCTTTACCTTCGGAACACAAGAA		11
134	amy-p3	I	I	CCTCTTTCGCAACTCACCGAAATGTCTTGCAGAAAGCCTTGACGCTGTT		11
135	orfI-p2	I	I	CCTCGGGAGGGAGCGGGTGTTTCTCCCGCAACCTATCGGAGATCGGTACA		11
136	octD-p5,6	R	I	CGCATTTCGTTCCGCCCTCTCTGACGTCGCTGTGGGATCCGCCCGGCGG		11
137	vph-pa1	N	N	CACTGGAATGCCCCTACCACGGTTGGTTGTTCGAAACGGGC		11
138	afsB-p	I	I	GCCGAGCGAGCGGGGGGCGGCGCTGTCGCGGCTGCTGGACTTCTACCTG		11
139	redD-PL1	C	C	GATCACGGCGACCGTGTGCCCCTGCTGCTCCAGGGCCTGGGCGAGCGC		11
140	bar-p1	I	I	TTGTCTCCGTGTCTCCAGCCCCTGTCGGGACTCTCCTCGGGTGCCCTGCAT		11
141	brpa-p3	C	C	GACGGACCGGCGTCGCCGGCGCAGCATTTGCCTCCGGACTTCGAGCGGCAT		11

142 amy-p2	I	I	TCGCAACTCACCGAAATGTCTTGCAGAAAGCCTTGACGCTGTTAACGACA		11
143 vph-p2	N	N	GGGAGCGACGGAATGCGCGTGCACCGCCCGCTGCCCGTTTC		11
144 kmr-p	N	N	CTCGGCGCCCACTCCTCCACCAGGGAGCGGCTCCCCATC		11
145 redD-pr2	I	I	AAGAGGAAGATGAACGGGATGTTCACCGCGGTCAGAGGGCAGAGAAGAAG		11
146 vph-p1	N	N	GCAGCGCCGTGTGCGGCCTGCCCCGGCCCGGCGGGAGCGACGGAATG		11
147 A1-p	I	I	TGCAGGCGGGGAACTCCTATGCCGACCTCACGAGCCACGGCGGCAGCGG		11
148 orf-p3	C	C	ATCCTGGTTGCGGCGACGCGGGCGCGGCTGCTCGTCCGAACTGCTCACC		11
149 xylB-p1	R	I	AAATTAGTATGCAGGAGCACTTCTGGGAAGAGACACACCCCGCTGAGC		11
150 ssi-p1	I	I	TCAGGCCGATTAAGAGGCGGCGGATATTCGGCCATCTGGCCACTTCGCT		11
151 strD-p	N	I	CATCCTCCAGGGCGGAGAGGACCGGAACGGCCCCCTCCAGCCAGCGCCTGCGC		13
152 afsR-p2	I	I	TGCTGACGGTACGGGTGGAGTTGGAGGGGCAAAGCCGAAGAAGCTGATGG		11
153 hyg-p	I	I	GAACGTCCCCGACGTGGCCGACCAGCCCGTCATCGTCAACGCCTGACCG		11
154 cyt-p			TTTCAGGTTTAAATCCTTATCGTTATGGGTATTGTTTGTAATA	A	17
155 veg			ATATAATTAAATTTTATTTGACAAAAATGGGCTCGTGTTGTACAATAAATGTA	A, C, D, G(18)	17

$$r = \frac{\Sigma d_x d_y}{\sqrt{(\Sigma d_x^2)(\Sigma d_y^2)}} \quad \text{.....(Clarke and Cooke, 1983)}$$

A more detailed explanation of correlation coefficient is given in Appendix E

3.2.4 Promoter region sequence bias cluster analysis

A particular oligonucleotide was selected for testing if it had a high NCF value and was thus greatly over represented in the database being examined. Other sequences were selected for testing using different criteria (see section 3.3.6). In the search for the former, gaps (positions that could represent any base) were inserted in the oligomer sequence. The length of oligomer examined and the gap pattern used are shown in Table 3.4.

A number of *Streptomyces* promoters were aligned at the point of transcription initiation. Only sequences between and including positions -47 and +1 (as defined in Table 3.3) of the promoters were utilized. Where there were multiple transcription initiation points the nearest to the promoter was considered as position +1. Two groups of promoters were used on separate occasions (see section 3.3.6). Group 1 consisted of all the *Streptomyces* promoters listed in Table 3.3 numbered 1 to 153 except for *endo*H-p, *orf*RP-p, *whiG*-p *bldA*-p, *tylF*-p, *whiE*-p, *blaF*-p, *oct*D-p4, *oct*D-p5,6, *vph*-pA1, *vph*-p2 and *kmr*-p (promoters number 4, 20, 40, 43, 61, 67, 92, 122, 136, 137, 143 and 144). These were not included as their sequences did not extend as far as position -47 or because the data were not available at the time cluster analysis was performed. Group 2 consisted of the promoters listed in Table 3.3 numbered from 93 to 153 except for *oct*D-p4, *oct*D-p5,6, *vph*-pA1, *vph*-p2 and *kmr*-p.

The oligonucleotides selected were tested against either the Group 1 or 2 promoters. Each occasion they occurred within any promoter region was recorded. Also recorded was the position of the first base of the oligonucleotide in relation to the

Table 3.4 Information on oligomers used for promoter region oligomer cluster analysis

TYPE[a]	USING DATABASE SPRO#1			USING DATABASE SPRO#2		
	TEST[b]	MINIMUM VALUE NCF[c]	NUMBER TESTED[d]	TEST[b]	MINIMUM VALUE NCF[c]	NUMBER TESTED[d]
TRIMERS						
NNN	U			T	1.0	30
TETRAMERS						
NNNN	T	2.0	6	T	1.7	33
N-NN	U			T	1.2	16
NN-N	U			T	1.2	15
PENTAMERS						
NNNNN	T	2.8	13	T	3.5	32
N-NNN	U			T	1.5	36
NN-NN	U			T	1.6	33
NNN-N	U			T	1.5	45
N--NN	U			T	1.2	13
NN--N	U			T	1.2	9
N-N-N	U			T	1.2	11
HEXAMERS						
NNNNNN	U			T	7.0	125
N--NNN	T	1.5	6	T	1.5	37
NN--NN	T	1.5	18	T	1.5	39
NNN--N	T	1.7	10	T	1.5	39
N-N-NN	T	1.5	16	T	1.4	51
NN-N-N	T	1.5	15	T	1.4	42
N-NN-N	T	1.5	15	T	1.5	43
N---NN	T	1.3	5	T	1.1	21
NN---N	T	1.2	9	T	1.1	22
N--N-N	T	1.2	10	T	1.1	24
N-N--N	T	1.3	5	T	1.1	18
N----N	T	1.0	6	T	1.0	10

[a]Symbols; N, represents in turn each of A, G, C or T specifically; -, represents any base and is recorded as a space.

[b]Symbols; T, NCF value of oligomers was examined; U, NCF of oligomers was not examined.

[c]The minimum NCF value required for an oligomer to be tested by cluster analysis.

[d]The number of oligomers that had required NCF value and were tested by cluster analysis.

transcription start site of the promoter in which it was found. The number of times that an oligomer under consideration occurred at a particular distance (measured in numbers of bases) from the transcription start sites of the promoters was plotted as a bar graph. Putative promoter recognition sequences were aligned and the total number of times each base occurred in each position was recorded.

3.2.5 Protein coding region identification (FINDCODE)

A window (of the length of the oligomer to be used) was moved one base at a time down the sequence being analysed. At the first window position the frequency (F) of the oligomer in the window, using subset F1 data, was recorded. Likewise, in the second and third positions F2 and F3 data were used, always in that order. This pattern was repeated to the end of the sequence. The recorded frequencies of a set number of oligonucleotide windows (termed the averaged window) were then averaged. The total window used (termed the effective window) is therefore longer than the averaged window by the length of the oligonucleotide window minus 1. The averaged frequencies were then plotted against the sequence position. This process was repeated twice more, except that the process was started using F2 data in one case and F3 data in the third. The process was also repeated using R3, R2 and R1 data which were plotted in that order. The computer operated proceedure by which this was done has been termed FINDCODE.

If the sequence being examined had been used in the creation of the database utilised, that sequence was totally removed from the database before the analysis was conducted.

3.3 Results and Discussion

3.3.1 The distribution of oligomer frequencies

The distribution of oligomer frequencies for *E.coli* protein coding DNA (database Eco-C) was plotted. The oligomers used were 3 to 7 bases in length (Fig. 3.1 A to E).

Similar plots have been made from *B.subtilis* (database Bsub-C) and *Streptomyces* (database Strep-C) data. In these cases only the plots for pentamers are shown (Fig. 3.1 F and G).

Rogerson (1989) has previously reported that the frequencies of tetramers from *E.coli* protein coding DNA are normally distributed. Fig. 3.1 B corroborates this to a degree as it is not clear if any asymmetry exists. However, Figs. 3.1 C-E show that as the oligomer length is increased, the distribution of the oligomer frequencies becomes more skewed. Here the avoidance of certain oligomers appears to be compensated for by the increased usage of a limited number of common short sequences. Notice that the mode (or peak value) deviates from the mean value of 1. The same is true for *B.subtilis* and *Streptomyces*, both of which show an abnormal distribution with long oligomers (Fig. 3.1 F and G) that is more obvious than for the shorter ones (data not shown). This is particularly obvious for *Streptomyces*, as would be expected for a G+C rich organism.

The distribution of oligomer frequencies was analysed because an abnormal distribution precludes applying statistical methods to test the significance of the correlation coefficient (Clarke and Cooke, 1983).

3.3.2 Correlation between *E.coli* database subsets

It was argued in section 3.1.2 that certain sequences within the protein coding regions may be avoided for reasons that had nothing to do with the limitations imposed by translation. If this were true it would result in some degree of correlation between the frequency of an oligomer and the frequency of the same oligomer on the opposite strand. Likewise, correlation would occur between the frequencies of an oligomer in different

Figure 3.1 Distribution of bacterial oligomer frequencies

Plots of the total number of oligonucleotides of set frequency ranges against frequency. Frequency range groups are described in Methods (3.2.2). Data used in each case is from the protein coding sense strand (database subset C). A-E, *E.coli* trimers to heptamers (database Eco). F, *B.subtilis* pentamers (database Bsub). G, *Streptomyces* pentamers (database Strep). Dotted line, average value (=1).

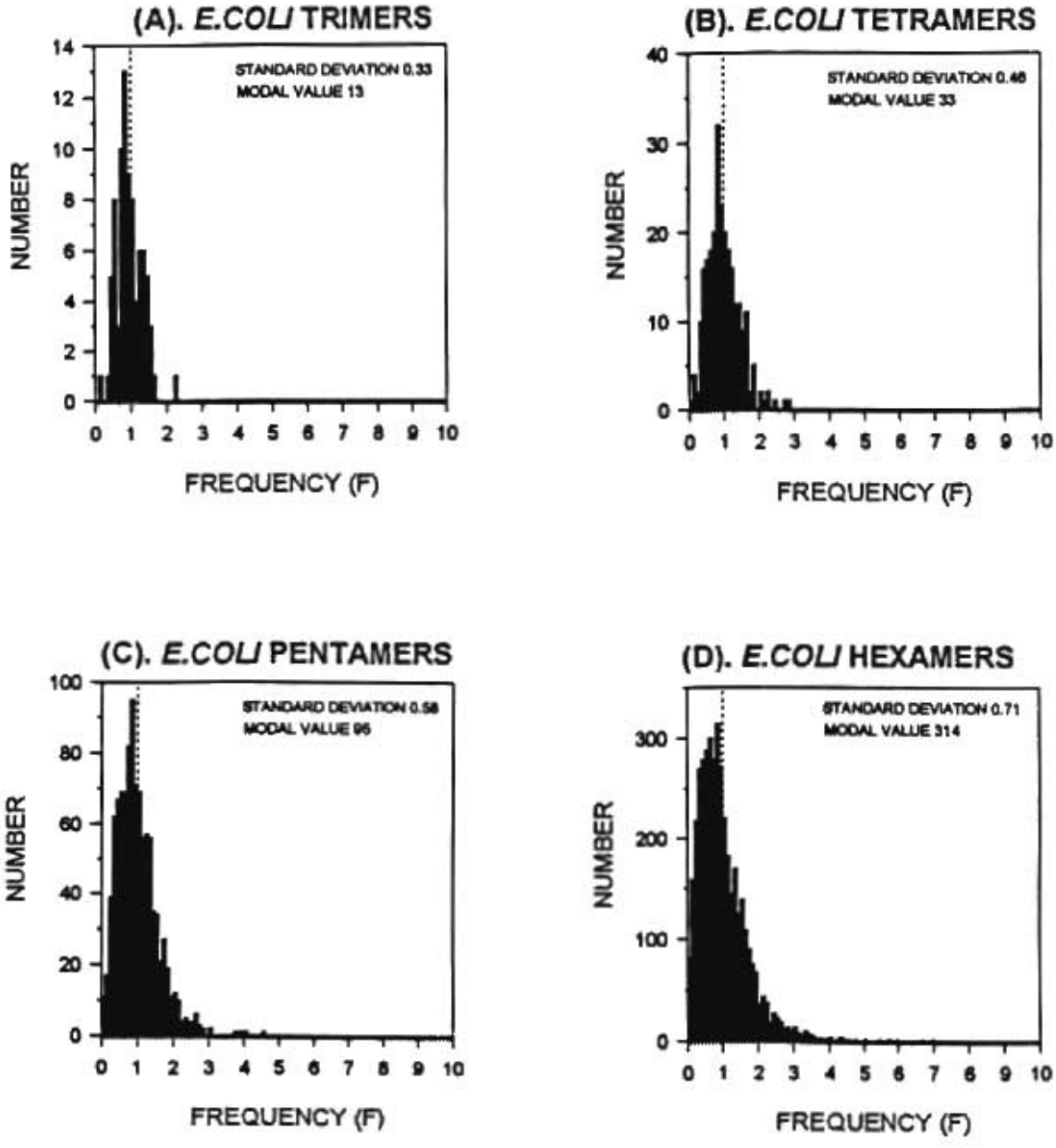
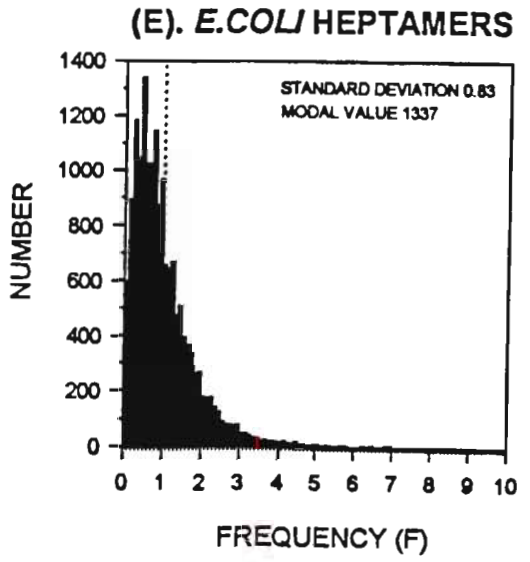
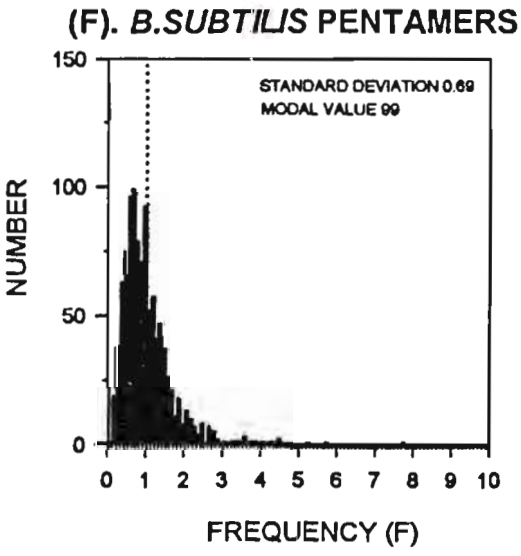
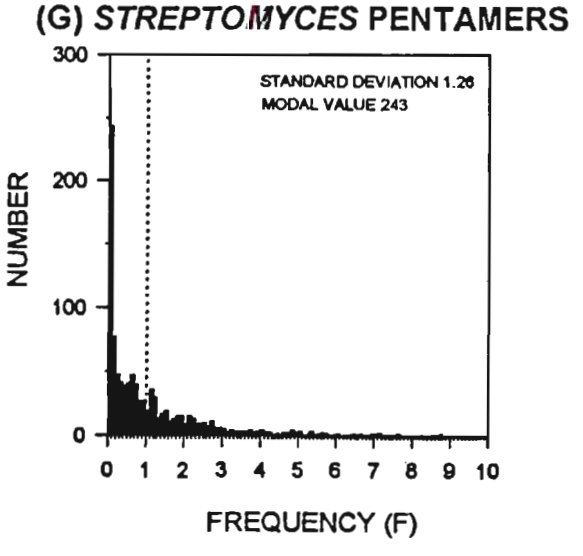


Figure 3.1 (continued)







frames. Moreover, the correlation should increase as the length of the oligomer increased.

Using *E.coli* and protein coding DNA sequence as a test system, the initial results were promising. The frequencies of oligomers from the sense strand (database Eco subset C) were plotted against the frequencies from the antisense strand (database Eco subset CR) to generate a scatter graph, and the correlation coefficient was calculated (section 3.2.3). The results for trimers to heptamers are shown in Fig. 3.2 A-E. Correlation between the two strands is obvious, and there appears to be a slight increase in the correlation as the oligomer length is increased. Notice that in 3.2 E only every fifth data point is plotted (due to the large number of data points) and the symmetry of the plots is therefore lost. The increasing correlation cannot be due to a G+C bias as the single nucleotide ratios for *E.coli* subset C are approximately 24/25/27/23 for A/C/G/T. The pattern generated in the scatter graphs using the longer oligomers is unexpected, with the points falling in a "horned" formation. There is no obvious explanation for this.

Similar tests between all combinations of the six possible reading frames (database Eco, subsets F1, F2, F3, R1, R2, and R3; Table 3.5) corroborates recently published work (Yumo and Urabe, 1994; section 3.1.2). The reasons for the correlation between the various frames can be easily understood. There are 6 possible reading frames on any stretch of DNA. As defined in Fig. 3.3, Frame 1 is the antisense frame to Frame 6, Frame 2 antisense to Frame 5 and Frame 3 antisense to Frame 4.

Testing the correlation of oligomer frequencies between the F (forward) type and R (reverse) type frame dependent database subsets is equivalent to testing correlation between each oligomer frequency in one of Frames 1-3 against the frequency of the same oligomer in one of Frames 4-6. Which of Frames 1-3 is to be tested is chosen prior to testing. However selection of the oligomer length determines which of the Frames 4-6 is represented by which of the different R type database subsets.

So (for example, see Fig. 3.4) when hexamers are considered, a correlation test between subsets F1 and R1 represents a test

Figure 3.2 Scatter analysis comparing oligomer frequencies from different database subsets.

The frequency (F) of each oligomer in a database subset was plotted against the frequency of the corresponding oligomer in the second subset. A to E, *E.coli* (database Eco) trimers to hexamers, protein coding sense strand (subset C) and protein coding antisense strand (subset CR). F, *B.subtilis* (database Bsub) hexamers, protein coding sense strand (subset C) and protein coding antisense strand (subset CR). G, *Streptomyces* (database Strep) hexamers, protein coding sense strand (subset C) and protein coding antisense strand (subset CR). H, *E.coli* (database Eco) pentamers, Frame 1 (subset F1) and Frame 5 (subset R1). I, *E.coli* (database Eco) pentamers, Frame 1 (subset F1) and Frame 6 (subset R2). J, *E.coli* (database Eco) pentamers, protein coding sense strand (subset C) and non-protein coding (subset NC).

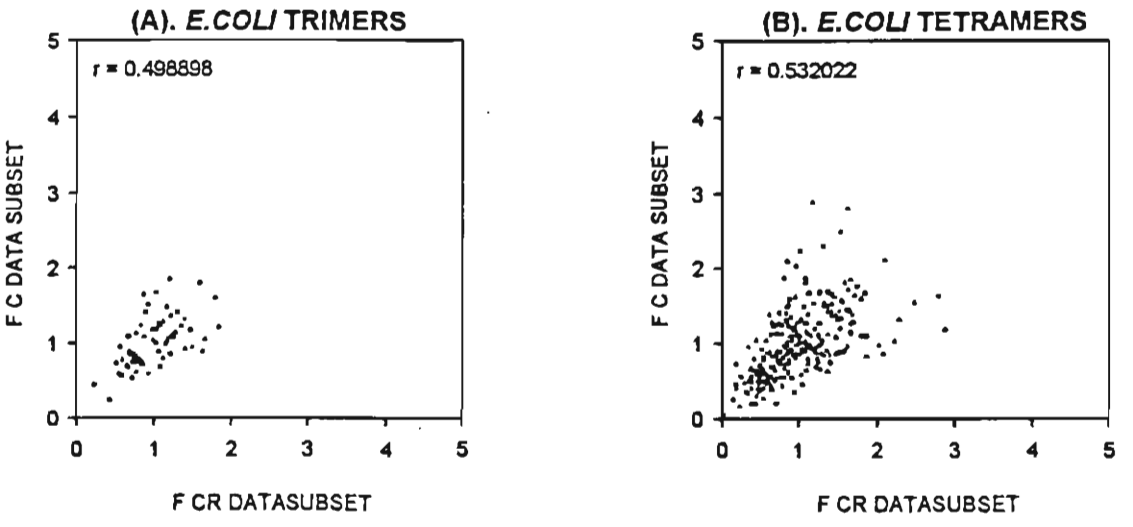
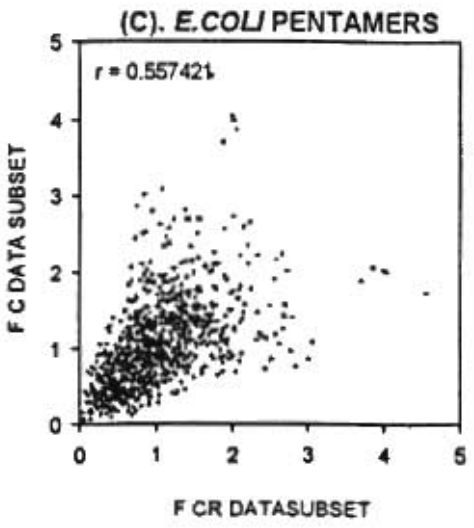
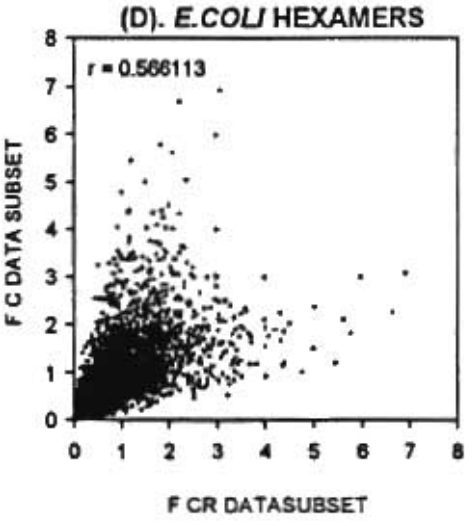
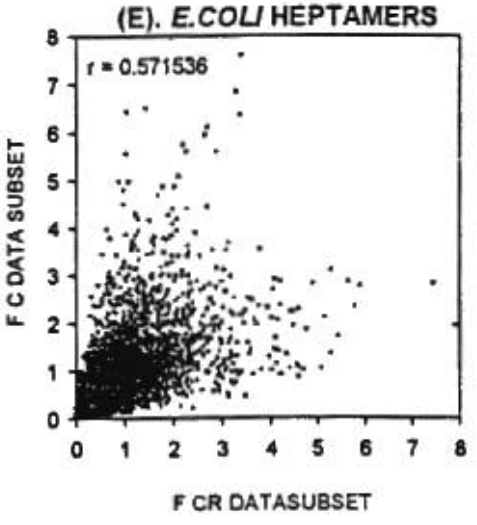
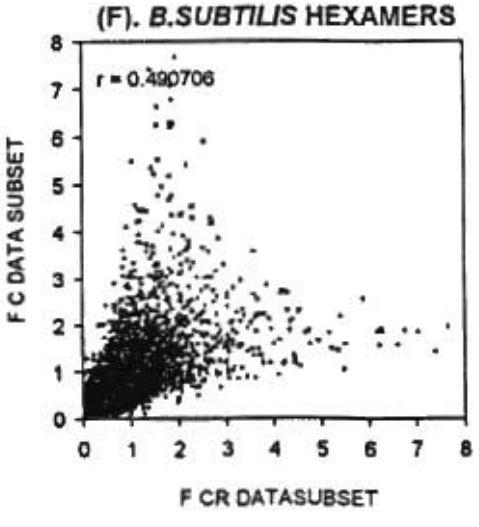


Figure 3.2 (continued)









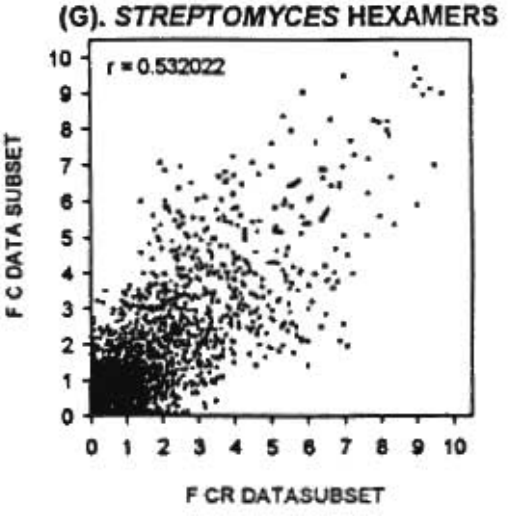
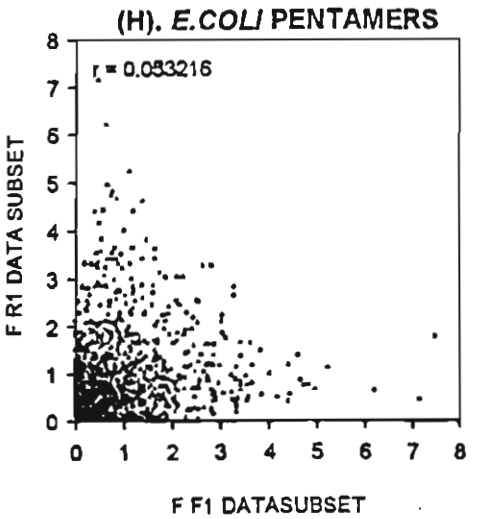
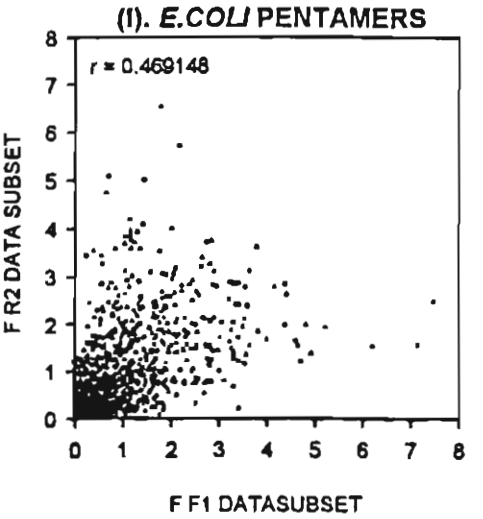


Figure 3.2 (continued)





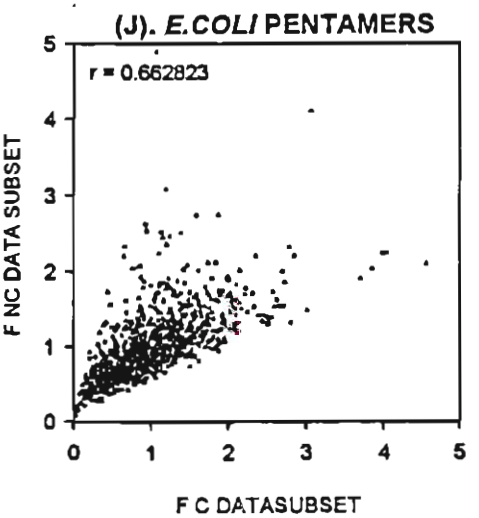


Table 3.5 Correlation between frequencies of oligomers in different database subsets

A. Database Eco (*E.coli*)

TRIMERS						
SUBSET	F1	F2	F3	R1	R2	R3
F1 F2 F3		-0.009255	-0.019331 -0.061691	0.521943	0.001604 0.045591	0.035559 0.56499 -0.161379
TETRAMERS						
SUBSET	F1	F2	F3	R1	R2	R3
F1 F2 F3		-0.058432	0.102082 0.015007	0.024407	0.051622 0.500144	0.595590 -0.014898 0.152891
PENTAMERS						
SUBSET	F1	F2	F3	R1	R2	R3
F1 F2 F3		0.500144	0.046609 0.102325	0.053216	0.569148 0.033269	0.015704 0.141988 0.492715
HEXAMERS						
SUBSET	F1	F2	F3	R1	R2	R3
F1 F2 F3		0.055620	0.060152 0.058864	0.422840	0.039810 0.132779	0.117411 0.450367 0.018704
HEPTAMERS						
SUBSET	F1	F2	F3	R1	R2	R3
F1 F2 F3		0.055863	0.087605 0.063878	0.040322	0.110224 0.403038	0.400041 0.035962 0.145438
OCTAMERS						
SUBSET	F1	F2	F3	R1	R2	R3
F1 F2 F3		0.036221	0.056298 0.078582	0.088606	0.346481 0.041004	0.035614 0.123484 0.358297

B. Database Bsub (*B.subtilis*)

HEXAMERS						
SUBSET	F1	F2	F3	R1	R2	R3
F1 F2 F3		0.203406	0.192767 0.192755	0.163946	0.187932 0.261920	0.264770 0.199969 0.182007

C. Database Strep (*Streptomyces*)

HEXAMERS						
SUBSET	F1	F2	F3	R1	R2	R3
F1 F2 F3		0.154433	0.165077 0.173370	0.609712	0.044761 0.476563	0.441198 0.630630 0.048360

Figure 3.3 Definition of frames

```
Frame 1          *--
Frame 2           *- -
Frame 3            * --
                AGA TCA TCA GCT ACG CA
                TCT AGT AGT CGA TGC GT
Frame 4            -- *
Frame 5             - -*
Frame 6               --*
```

* represents the initial base of the triplet in each frame
- represents the remaining two bases in the triplet

Figure 3.4 The relationship between frame and subset

HEXAMERS

A Database subset F1 and R1	**D** Database subset F2 and R2
F1 data set (Frame 1) N AGA TCA NNN NNN TGA TCT NNN N TCT AGT NNN NNN ACT AGA NNN R1 data set (Frame 6)	F2 data set (Frame 2) N NAG ATC ANN NNN NTG ATC TNN N NTC TAG TNN NNN NAC TAG ANN R2 data set (Frame 4)
B Database subset F1 and R2	**E** Database subset F2 and R3
F1 data set (Frame 1) N AGA TCA NNN NNN NTG ATC TNN N TCT AGT NNN NNN NAC TAG ANN R2 data set (Frame 4)	F2 data set (Frame 2) N NAG ATC ANN NNN NNT GAT CTN N NTC TAG TNN NNN NNA CTA GAN R3 data set (Frame 5)
C Database subset F1 and R3	**F** Database subset F3 and R3
F1 data set (Frame 1) N AGA TCA NNN NNN NNT GAT CTN N TCT AGT NNN NNN NNA CTA GAN R3 data set (Frame 5)	F3 data set (Frame 3) N NNA GAT CAN NNN NNT GAT CTN N NNT CTA GTN NNN NNA CTA GAN R3 data set (Frame 5)

HEPTAMERS

G Database subset F1 and R1
F1 data set (Frame 1) N AGA TCA TNN NNN ATG ATC TNN N TCT AGT ANN NNN TAC TAG ANN R1 data set (Frame 4)

* Position from which frame is determined.

using Frames 1 and 6. However using heptamers and the same subsets the test is between Frames 1 and 4 etc. This is shown (Fig. 3.4 A and G) for the hexamer AGATCA and the heptamer AGATCAT

For each of the six correlation tests between database subsets Eco F1-3 and Eco R1-3 there are two occasions in which a frame is tested against its antisense frame. This is shown for a hexamer in Fig. 3.4 (A and E). In this case, Frames 1 and 6, and Frames 2 and 5 are antisense to each other. The third possibility (Frames 3 and 4) is not shown because the correlation between Frames 2 and 5 for any oligomer is the same as the correlation between Frames 3 and 4 of the reverse compliment of that oligomer.

Table 3.5 A shows the expected high degree of correlation between Frames 1 and 6, Frames 2 and 5 and Frames 3 and 4 in *E.coli*. However, the correlation appears to decrease slightly with increasing oligonucleotide length. There also appears to be a low level of correlation between Frames 1 and 5, Frames 2 and 4 and Frames 3 and 6. This second, low correlation pattern was not observed by Yomo and Urabe (1994), and is likewise not obvious here using trimers. However, the correlation is maintained in oligomers of lengths greater than 3 bases (Table 3.5). The theory of Yomo and Urabe (1994) explaining the correlation between each frame and its antisense fails to explain this phenomenon. No strong correlation is observed between the remaining frames, except perhaps for F1 and F3, and F2 and F3, for tetramers and pentamers respectively. An example of the type of scatter graph generated when frame specific data is compared is shown in Fig. 3.2 H and I.

Large values of correlation coefficient do not necessarily require tests for significance especially when taken in conjunction with a scatter graph (see Fig. 3.2 A-E, for example). However very low values are not so readily interpreted. Unfortunately, because the oligomer frequencies are not normally distributed (Fig. 3.1) statistical significance tests cannot be applied.

3.3.3 Correlation between *B.subtilis* and *Streptomyces* database subsets

The correlation between database subsets using *B.subtilis* and *Streptomyces* data was tested. Only data for hexamers are shown (Table 3.5, B and C; Fig 3.2 F and G). In both cases the correlation between frames was completely different to that of *E.coli*.

In the case of *B.subtilis* correlation was recorded between all the frames (Table. 3.5 B). Such correlation might be expected in an organism whose genome is either A+T or G+C rich, however the nucleotide ratio found in database Bsub subset C is approximately 31/20/24/25 for A/C/G/T. Furthermore, the scatter graph of subset C plotted against subset CR (Fig. 3.2 F) shows a pattern, similar to that of *E.coli*, and not that expected to result from G+C bias.

In the case of *Streptomyces* data the scatter graph is exactly what would be expected in a G+C rich organism (the nucleotide ratio in database Strep subset C is approximately 15/37/35/13 for A/C/G/T). The correlation between individual frames (Table 3.5 C) is somewhat similar to that of *E.coli* in form, if not in scale, except that there appears to be a correlation between all the forward (F1, F2 and F3) reading frames.

These patterns (and those described in section 3.3.2) are difficult to reconcile with the theory of Yomo and Urabe (1994), and they do not appear to be directly related to the frequency of stop codons in different frames in any obvious way (data not shown). It is possible that they arose as described in the following tentative proposal.

At any point during the course of evolution an organism would be likely to have some form of codon bias, even if it was low and arose by chance. As evolution proceeded, proteins must have been assembled through rearrangements such as deletions, insertions and inversions. The chances that a functional protein would arise from such events would depend on the frequency of occurrence of stop codons and start codons. Because of the original codon bias, a bias in these would exist in the noncoding frames. Therefore different frames would have different chances of becoming incorporated into novel proteins.

Because some frames would be used in novel proteins more than others, the codon bias would then reflect those noncoding frames to different degrees. This dynamic process might never cease. The resulting bias would depend on many factors including the original bias, the G+C ratio (for example resulting from a biased DNA repair mechanism) and the relative rates of different types of rearrangement, all of which could differ between organisms. Such a process could have a profound effect on codon bias and would result in the kind of correlation patterns described here.

3.3.4 Can protein coding DNA sequence bias be used for protein-DNA binding site prediction?

The theory that protein coding DNA sequence would be biased as a result of the avoidance of inappropriate protein-DNA binding sites leads to the prediction that there will be a correlation between the oligomer frequencies of different frames and strands. This should be increasingly obvious as the lengths of the oligomers is increased. The pattern of correlation between frames observed here partially reflected the pattern previously established (Yomo and Urabe, 1994). In the case of the correlation between frames, no general and major increase was recorded as the oligomer length increased (indeed in some cases a consistent decrease was recorded; Table 3.5 A) However overall there was a slight increase (Fig. 3.2 A-E). Whether such an increase is as a result of the theory proposed is debatable. Even if it were so, such a minor effect is unlikely to be useful for the determination of protein binding sites. This is because the bias, if it exists at all, is totally overshadowed by the bias due to the protein coding considerations. Furthermore, the tests here have revealed patterns of correlation for which there is no obvious explanation. Such features include the odd "horned" pattern found in scatter graphs using database Eco and Bsub subsets C and CR. Nor can the correlation between frames in different species be accounted for. As these patterns cannot be explained their effects cannot be eliminated. Therefore, because of these factors, and because another approach had proved more fruitful (Section 3.3.6), this line of work was discontinued.

3.3.5 Test for NC subset contamination

The non-coding DNA sequences of an organism will not display sequence bias resulting from the requirements of the translation machinery, and are likely to be over represented in sequences that define protein binding sites. However it is unclear as to what degree the protein coding sequence in the Genbank database has been correctly recognised. Consequently, it is unclear how much of the non-coding (NC) database subsets are contaminated with unrecognised protein coding sequence.

To investigate this, the correlation between the protein coding and non-coding subsets of the *E.coli* database (Eco-C and Eco-NC) was tested using pentamers (Fig. 3.2 J). It can be seen that the correlation is extremely high, indicating that Eco-NC is heavily contaminated with protein coding data, rendering it useless for further analysis. The Bsub and Strep databases suffer a similar drawback (data not shown). In order to create non-coding databases that were free from contamination the databases Spro#1 and Spro#2 were generated using *Streptomyces* sequence data.

3.3.6 *Streptomyces* promoter sequence bias cluster analysis

Promoter region sequence bias cluster analysis was conducted on a number of *Streptomyces* promoters. Initially the selection of the oligomers, of up to six bases, to be used for the test was made using the Spro#1 database and NCF data. However the database was too small to support the use of hexamers without any gaps in the sequence. To overcome this, the oligomers used were such that not every position within their length was assigned as either A, G, C, or T. The databases used, types and numbers of oligomers examined, and the minimum NCF value required for them to be considered for further testing are listed in Table 3.4.

Cluster analysis was conducted using the Group 1 promoters unless otherwise specified. Selected results of the cluster analysis are shown in Fig. 3.5. Many of the oligomers gave clear, or highly suggestive results. These sequences have been termed potential promoter core sequence(s) (PPCS), and grouped into Classes A to H. Each Class consisted of either a single

Figure 3.5. Cluster analysis of oligonucleotides

Number of occurrences of the oligonucleotide is plotted against the position relative to the transcription start site. All plots utilize Group 1 promoters, except for F1 which utilizes Group 2 promoters. Dotted lines, limits in which positions it is possible to record oligomers occurring.

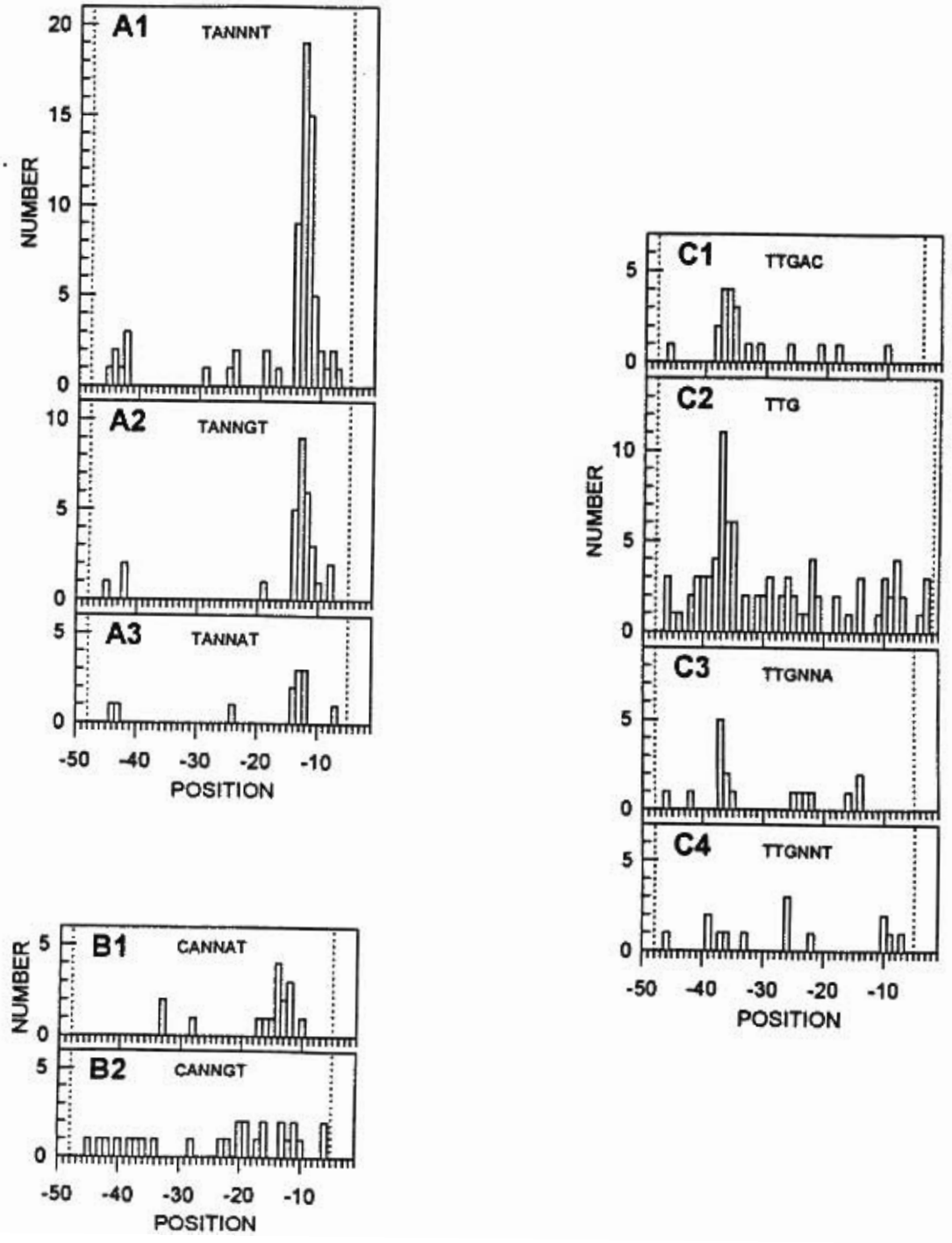


Figure 3.5 (continued)

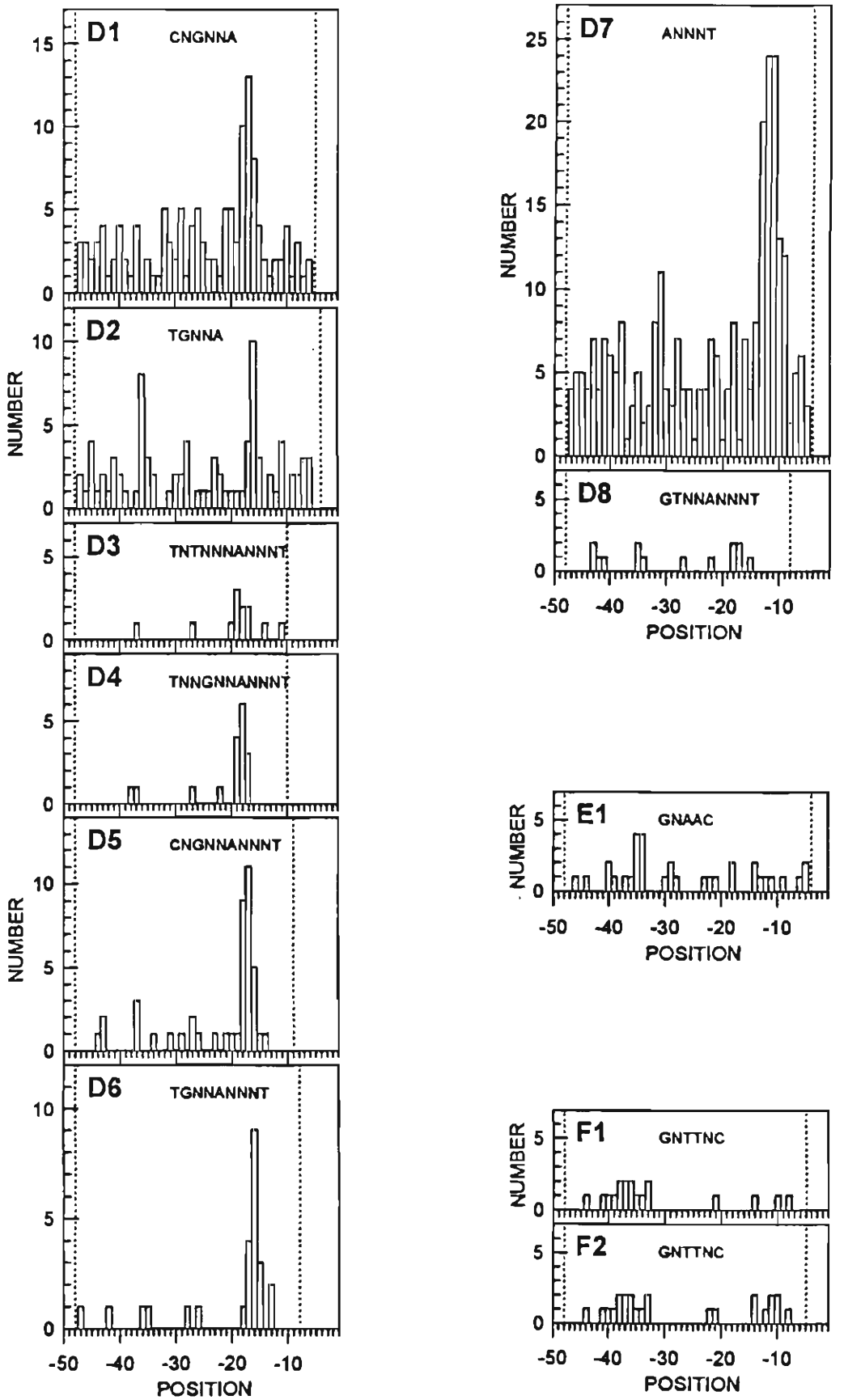
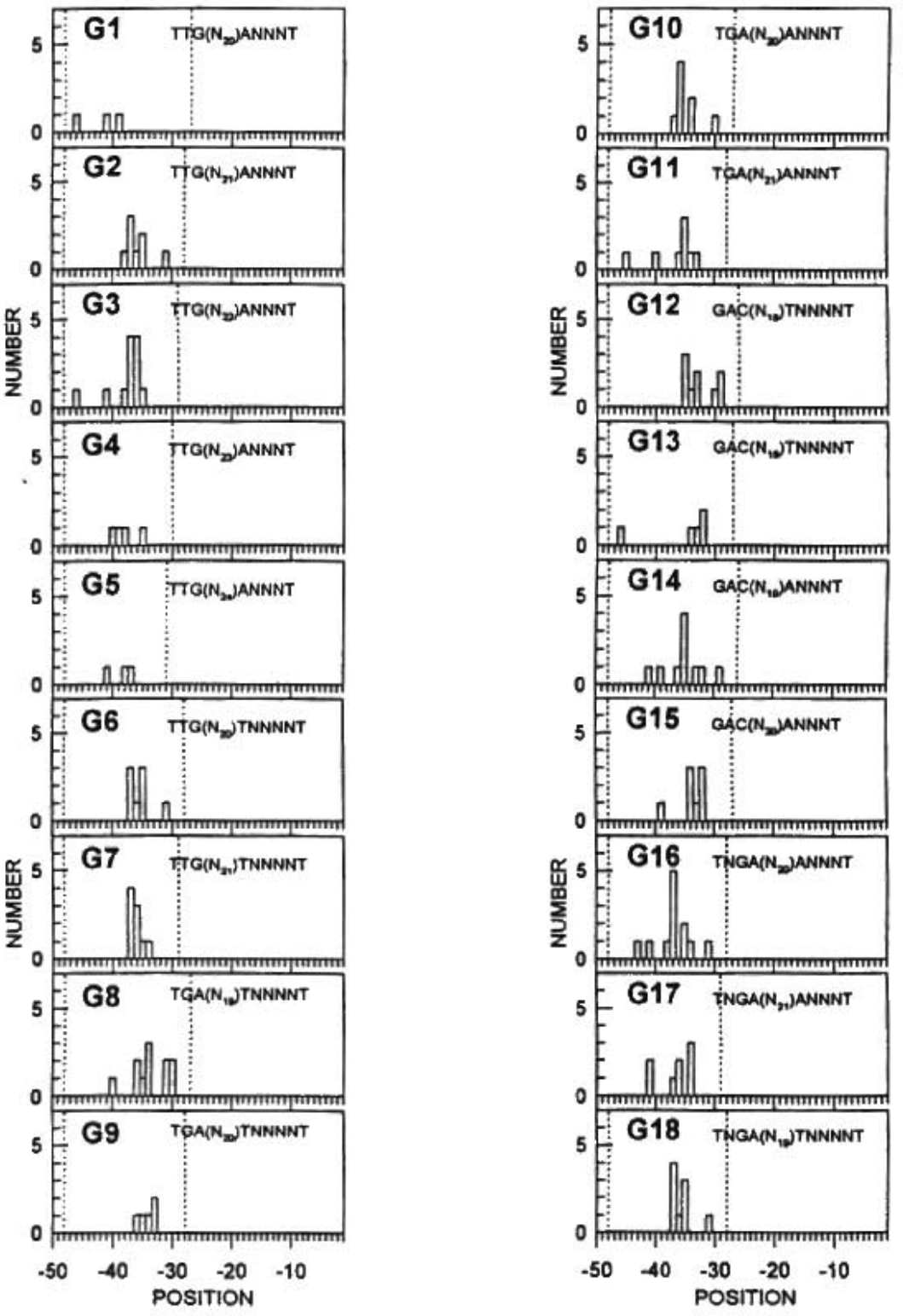


Figure 3.5 (continued)



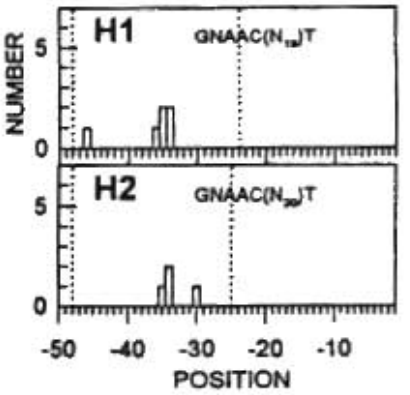


Table 3.6 Consensus analysis of PPCS classes

A. PPCS Class A

"POSITION"[a]		-19	-18	-17	-16	-15	-14	-13	-12	-11	-10	-9	-8	-7
E.COLI CONSENSUS[b]			T		T	G		T	A	T	A	A	T	
STREPTOMYCES CONSENSUS[c]		C/G	G/C/T	C/G	G/C/T	G	C/G/A	**T**	**A**	G/C	G/C/A	G/T	**T**	G/C
BASE	G	16	18	19	18	30	16	0	0	24	21	25	0	21
BASE	C	23	14	22	16	10	18	0	0	19	17	9	0	20
BASE	A	7	7	4	3	6	11	0	53	6	11	8	0	3
BASE	T	7	14	8	16	7	8	53	0	4	4	11	53	9

B. PPCS Class C

"POSITION"[a]		-41	-40	-39	-38	-37	-36	-35	-34	-33	-32	-31	-30	-29
E.COLI CONSENSUS[b]					T	C	T	T	G	A	C	A	T	
STREPTOMYCES CONSENSUS[c]		A			G	C	**T**	**T**	**G**	**A**	**C**		C	
BASE	G	3	3	6	7	5	0	0	15	0	0	6	5	6
BASE	C	2	4	2	5	8	0	0	0	0	15	3	10	5
BASE	A	8	4	2	0	1	0	0	0	15	0	4	0	1
BASE	T	2	4	5	3	1	15	15	0	0	0	2	0	3

[a] "Position" number is based on the positioning of the majority of the PPCS with respect to the transcription start site.

[b] Hawley and McClure, 1983.

[c] Bold sequence symbols define the PPCS class.

PPCS or a number of PPCS that were selected for testing by the same logical process and were similar in sequence.

The most striking result was for the sequence TANNNT (PPCS Class A; Fig. 3.5 A1). Of the 67 times this sequence occurred, 50 times (74.6%) it was found in the region -14 to -9. The clustering of this sequence could be due to the presence of *E.coli* σ^A-type promoters. However the results from cluster analysis using the sequences TANNGT and TANNAT (Fig. 3.5 A2 and A3) indicate that the preferred -10 region varies from the *E.coli* consensus.

A consensus calculation was performed using all the PPCS Class A sequences that fall between positions -14 and -9 (Table 3.6 A). There are 53 such promoters (Table 3.3), including *orf*RP-p, *whi*G-p and *tyl*F-p (numbers 20, 40 and 61), which were not used in the cluster analysis. It is difficult to determine a consensus sequence for an organism, such as *Streptomyces*, which has a strong G+C bias (calculation of monomer frequencies, or F values, showed that for database Spro#1 subset T the nucleotide frequencies are approximately 19/30/32/19 for A/C/G/T). This will tend to obscure the base preference due to the function of the DNA sequence under consideration. However, it would be expected that for positions within the consensus that have no function, G and C residues would be equally represented. A and T residues should likewise be equally represented. Wherever this was not the case it has been assumed that this is due to a base preference required for the protein binding activity of the DNA. The resulting consensus, TA(G/C)(G/C/A)(G/T)T, differs considerably from that of *E.coli* (Table 3.6 A). Most notable is the fact that T is the least common base at position "-11" while A is least common at position "-9", in direct contrast to the situation in *E.coli* (inverted commas are used to indicate that the "position" is as defined for Table 3.6).

The PPCS Class B is less obviously clustered than Class A (Fig. 3.5 B1). In this case the sequence CANNAT occurs in the -15 to -12 region 10 times out of a total of 16 (62.5%). It is the absence of this sequence in other regions that is significant, because it is likely that promoter-like regions will be avoided at positions where they have no function.

It could be argued that PPCS Class B merely represents a subset of promoters that are transcribed by the same RNA polymerase that recognizes Class A. However, it is reasonable to expect that were that the case, a sequence that retained features of Class B, but was closer to the consensus sequence of Class A, would be found clustered in the -10 region. CANNGT is such a sequence, yet in this case (and others, data not shown), no such clustering is observed (Fig. 3.5 B2). Thus the sequence CANNAT may represent the core components of a novel -10 region.

The PPCS Classes A and B appear to be associated with a third Class, C, in the -35 region (Fig. 3.5 C1). This sequence, TTGAC, is found over five times more often than predicted (database Spro#1 subset T, NCF value 5.43). 14 out of 20 (70%) sequences of this class start in the -38 to -33 region. Furthermore, the sequence is exactly that of the highly conserved part of the *E.coli* major σ factor -35 recognition site consensus sequence, except one base shorter (Hawley and McClure, 1983; Table 3.6 B). This strongly indicates that the cluster represents the -35 recognition site of *Streptomyces* *E.coli*-like promoters. It should also be noted that when the PPCS was found outside the cluster this was due to the phenomenon of overlapping multiple promoters (as was often the case with other PPCS classes, data not shown). Thus, the occurrence of the sequence TTGAC at positions -10 and -18 is as a result of the same sequence being found at different positions in the *amy*-P3 and *amy*-p2 promoters as well as in the *amy*-p1 promoter at position -36 (Table 3.3, promoters number 134, 142 and 7). The same is true of the *brp*A-p1 and *brp*A-p2 promoters at positions -33 and -26 (Table 3.3, promoters number 5 and 106).

Several attempts were made to identify more of these sites using limited subsets of the sequence TTGAC, but this was unsuccessful (data not shown). Although clusters were obvious it was impossible to distinguish in which instances the test sequence was likely to be functionally significant. This was due to the frequent occurrence of the same sequence in regions other than the -35 region. An example of this is shown for the sequence TTG (Fig. 3.5 C2).

A consensus sequence for the 15 PPCS Class C sequences found between positions -38 and -33 (including that of promoter number 4, *endo*H-p, which was not included in the cluster analysis) was generated (Table 3.6 B). Although there are too few of these sequences to be confident that small deviations are meaningful, it should be noted that there appear to be highly conserved A, G, C, and C residues at positions -41, -38, -37 and -30 respectively. This consensus is markedly different to that of *E.coli* in several features, most importantly the A residue at position -31, which is highly conserved in *E.coli*, does not appear to be exceptionally common in *Streptomyces*. However when the sequence TTGNNA was tested by cluster analysis (Fig. 3.5 C3) a cluster was observed in the -35 region, which argues that the A residue has some functional significance. The oligomers TTGNNT (Fig. 3.5 C4), TTGNNC and TTGNNG (data not shown) were also tested and only TTGNNT did not generate a cluster.

Another class of hexamer (CNGNNA) was seen to be clustered between bases -18 and -15 (Fig. 3.5 D1). The sequence was often associated with, and overlapped PPCS Classes A and B. This suggests a functional role for the sequence CNGNNA because it has only one base in common with those two other classes. In addition, the consensus sequence for the PPCS Class A (Table 3.6 A) shows that C and G are predominant in positions "-17" and "-15" respectively. This observation prompted a search for other oligomers clustered upstream of the -10 region, selected on the basis of their similarity to the extended PPCS Class A consensus sequence. Several were found and an example is shown for the sequence TGNNA (Fig. 3.5 D2). It is interesting to note that in this case there is a second cluster in the -35 region which represents a subset of PPCS Class C.

Thus the cluster analyses and the consensus sequence data together suggest that, as in *E.coli* (Hawley and McClure, 1983), an extended consensus sequence occurs for the -10 region in *Streptomyces*. Further cluster analyses were performed using sequences that were based on the PPCS Class A consensus in the "-18" to "-8" region. Again, some degree of clustering was recorded for a number of oligomers but only four were free from a high background. These have been termed PPCS Class D, and

include the sequences TNTNNNANNNT, TNNGNNANNNT, CNGNNANNNT and TGNNANNNT (Fig. 3.5, D3-D6).

It could be argued that the clustering of PPCS Class D simply reflects the fact that the sequence ANNNT itself occurs as a cluster (Fig. 3.5 D7). Two lines of evidence suggest that this is not the case. Firstly, it would then follow that the clustering of the sequences TGNNA and CNGNNA arises due to a preponderance of A residues in the -12 region. This would mean that most sequences that contain an A would be clustered in this region, which is not the case (data not shown). Secondly, if the ANNNT motif of the Class D sequence is retained intact, while the remaining nucleotides are assigned new positions, only a barely discernable peak is observable after cluster analysis. As an example of this the results for the sequences TGNNANNNT and GTNNANNNT can be compared (Fig. 3.5, D6 and D8).

Of the 153 promoters tested, 65 (Table 3.3, promoters number 88 to 153) did not fall into any of the classes that were defined using the Spro#1 database to select over represented sequences. There are several possible reasons for this. Firstly, it is possible that there are other classes of promoter which were not identified as they are rare and so do not appear as an obvious peak in cluster analysis. Such a problem would be exacerbated by the presence of large numbers of promoters of the major class(es) in the original sequence data entered. Alternatively, it is possible that the oligomers that define more common promoter recognition core sequences are not over represented in the database and were not tested because of this. This could occur if the over representation due to the oligomers appearance in the protein binding regions is counterbalanced by (not unexpected) under representation in other parts of the promoter. It is also possible that other PPCS were not identified because they are G+C rich and because the frequency of occurrence has been corrected for nucleotide bias. Any G+C rich oligomer would have to be numerically more highly represented than an A+T rich sequence in order to attain a similar NCF value. Finally it is possible that alternative PPCS classes were not identified because only hexamers were extensively tested.

Alternatively, many of the promoters that have not been classified on the basis of cluster analysis are probably members of the same promoter class(es) as those that have been tentatively defined above. In these cases, the -10 and -35 regions might vary too much from the consensus to be identified by the method used here.

In order to address these problems a number of approaches were taken. Firstly the biochemical data concerning certain promoters were considered in regard to the cluster analysis results. There are a limited number of *Streptomyces* promoters which are known to be expressed via different transcription factors. These include the promoters for the *dagA* and *gal* operons, the XP55 promoter and the σ^{whiG} dependent promoters P_{TH4} and P_{TH270} (section 1.2.3.1). As only *dagA*-p2, *gal*-p2 and the σ^{whiG} dependent promoters (numbers 91, 126, 110 and 111) do not contain sequences that can be classified as any of the PPCS Classes A to D, it is possible that these are representatives of minor classes.

The consensus sequence of the σ^{whiG} dependent promoters is thought to be similar to that of the motility related promoters of other bacteria, or the sequence TAAA(N_{15})GCCGATA(A/T) (Tan and Chater, 1993). In the case of the *dagA*-p2 -35 region recognition site there is evidence that the consensus may be CCGGAACTT (Lonetto *et al.*, 1994). Cluster analysis of all possible subsets of these sequences (triplets to octamers, with and without appropriate blank spaces) was carried out. No clusters were observed using any of the σ^{whiG} target-type sequences. This is not surprising as there has been no widespread effort to isolate and map this type of promoter. Furthermore, the majority of the transcription mapping of the promoters used in this study was performed using logarithmically growing mycelia from liquid medium, in which σ^{whiG} may not be expressed. These two factors would ensure that σ^{whiG} dependent promoters are poorly represented among the promoters examined here.

In the case of the *dagA*-p2 -35 region-type sequences, only one, GNAAC, gave any indication of clustering, albeit very weakly (Fig. 3.5 E1). This sequence has been termed PPCS Class E and its significance is questionable. It is interesting to note,

however, that in three cases, *hrd*D-p1, *whi*B-p1 and *dag*A-p2 (promoters number 89, 90 and 91) there is a correctly spaced TC dimer in the -10 region, as the putative recognition sequence for the σ^{28} promoter class suggests there should be (Lonetto *et al.*, 1994). This observation was used to define a further class of PPCS (see below).

To address the possibility that the presence of many of the major class(es) of promoters in the Spro#1 database was hindering the detection of minor classes, the Spro#2 database was created. This consisted only of promoters that did not contain appropriately positioned PPCS of Classes A to E (Table 3.3). As with the Spro#1 database, highly over represented sequences in the Spro#2 database were identified (Table 3.4). These were tested against the Group 2 promoters, which do not include among their number any promoters containing appropriately positioned PPCS Classes A to E, using triplets to hexamers.

Only one sequence, GNTTNC, gave results that might be interpreted, with caution, as a cluster (Fig. 3.5 F1). It is interesting to note, however, that when this sequence (Class F) was tested using the Group 1 promoters, the cluster pattern was almost identical (Fig. 3.5 F2), indicating that the sequence is very rarely found among the promoters that contain appropriately positioned PPCS Classes A to E.

Further tests were conducted to attempt to isolate promoter-like sequences that resembled the previously identified PPCS classes poorly, but which might still be functional. It was argued that while the -10 and -35 regions alone could not be recognized, together there could be enough sequence information for their identification. As the PPCS Class A and B sequences appear to be associated with those of Class C, subsets of these classes, with a variety of spaces between them were tested.

To first establish the limits of the correct spacing, the sequences TTG and ANNNT were used, with the spacing between them set at 20 to 24 bases. Only in the cases of a 21 and 22 base separation were obvious clusters recorded, indicating that this was the correct spacing. It should be noted that this represents a spacing of 18 and 19 bases between the TTGAC and

(T/C)ANNNT, and is similar to the spacing found for the *E.coli* major class of promoter. However, the 17 base spacing found in *E.coli* was not observed (Fig. 3.5 G1 to G5).

With the spacing between the -10 and -35 regions set at either 18 or 19 bases, all possible combinations of subsets of TTGAC and (T/C)ANNNT were tested. Thirteen sequences, TTG$(N_{20/21})$TNNNNT, TGA$(N_{19/20})$TNNNNT, TGA$(N_{20/21})$ANNNT, GAC$(N_{18/19})$TNNNNT, GAC$(N_{19/20})$ANNNT, TNGA$(N_{20/21})$ANNNT and TNGA(N_{19})TNNNNT showed some clustering and were together termed Class G (Fig. 3.5 G6 to G18).

A similar approach was adopted to test for a -10/-35 association within the PPCS Class E sequences. As noted above, inspection had revealed that in some cases there appeared to be an appropriately positioned TC in the -10 region of promoters that fall within this class. For this reason cluster analysis was conducted on the sequences GNAAC(N_{19})T and GNAAC(N_{20})T. While similar sequences with lesser or greater spacing between the -10 and -35 regions were tested and did not show clustering (data not shown), when the spacing was either 19 or 20 bases, clusters could be observed (Fig. 3.5 H1 and H2). These two sequences have been termed PPCS Class H. The clustering that is observed strongly indicates that these sequences represent a distinct class of promoters, and gives support to the contention that the indistinct clustering of PPCS Class E sequences is not coincidental.

The promoters that fall into the various classes are shown in Table 3.3. Up to 59% of the promoters can be considered to be of the Classes A, B, C, D and G, and it seems that these represent *E.coli*-like promoters. Within this group are found the XP55-p, *dag*A-p4, *dag*A-p1, *dag*A-p3 and *gal*-p1 promoters (numbers 1, 21, 33, 79 and 45). Also in this group are the *B.subtilis veg* and *ctc* gene promoters (numbers 154 and 155). It is interesting to note that although the promoters are so similar, there are at least three distinct transcribing activities that recognize different members of this group. Perhaps these consist of (at least in part) the different *hrd* homologues.

Focusing primarily on *Streptomyces* developmental genes, other *E.coli*-like promoters of note include *whiG*-p, *whiE*-p, *whiB*-p2, *bldA*-p and *hrdB*-p. In the case of the *bldA* promoter (promoter number 43) the potential -10 region is rather distant from the transcription start site. With *whiB*-p2 (promoter number 9) the spacing between the -35 and -10 regions is very unusual (21bp), which in *E.coli* is thought to be a feature of promoters that are regulated by DNA supercoiling (Wang and Syvanen, 1992). This is intriguing, as *whiB* is probably a fundamental and early regulator of the *whi* gene network (section 1.3.4.6) and the *whiB*-p2 promoter is strongly activated at the onset of sporulation (Soliveri *et al.*, 1992). The possibility that chromosomal supercoiling plays a central role in *Streptomyces* sporulation has not been considered. Support for this speculation is found in the observation that *whiB* expression could not be detected with plasmid borne indicator gene fusions (Soliveri *et al.*, 1992). Notice that *actII2,3*-p (promoter number 8) also has an unusual 21 bp spacing between the potential -10 and -35 sequences.

The unusual spacing between the -35 and -10 regions of *whiB*-p2 only occurs if the PPCS Class A motif is considered as the -10 region. However, *whiB*-p2 also contains a Class G PPCS and this may indicate that a different -10 region is used. It is also possible that both potential -10 regions may be functional.

Promoters in the Classes E and H might be of the σ^{28} class. This is particularly relevant because these include the promoters of a number of fundamentally important genes including *hrdB* and *hrdD* (promoters number 85 and 89; see also section 1.2.3.1). In the case of *hrdB*-p, Class B and H sequences overlap, suggesting that *hrdB* could be transcribed from either an *E.coli*-like or σ^{28} type promoter, or indeed from both. Also of note is the fact that *whiB*-p1 (promoter number 90) could be a σ^{28} type promoter. *SapA*-p (number 87; see also section 1.3.4.4) too could fall in this class, although it also has features indicating it could be of the *E.coli*-like type. It is interesting to note that it has been argued that *sapA* does not need sequences upstream of the -8 region in order for transcription to occur (see section 1.3.4.4).

The *bla*F-p promoter (number 92) contains PPCS Class E and H sequences, although these are not within the cluster but 3 bases removed from it. However extensive mutation analysis of this promoter has been performed (Forsman and Granstrom, 1992) and comparison of these results with those obtained here strongly suggest that *bla*F-p is a σ^{28} type promoter. Both point mutation and deletion analysis are compatible with the results reported here.

The *gal*-p2 promoter (number 126; see also section 1.2.3.1) is an interesting case as it contains both the PPCS Class E and H sequences, but in the wrong position, at -45. Inspection of this promoter reveals that it matches the putative σ^{28} consensus recognition sequence (Lonetto *et al.*, 1994) in ten of the eleven positions, with the correct spacing between the -10 and -35 regions. It seems unlikely that such a sequence could occur by chance (especially as it is relatively A+T rich and the database is so small) and that it should have no function.

The *bld*A promoter (number 43; see also section 1.3.4.2) has similar features. In this case there is a potential Class A PPCS, but it is further from the transcription start site than usual (position -17). Furthermore, the potential *E.coli* like -35 region identified by Leskiw *et al.* (1993) is an unusual 17 bp from that potential -10 region (notice that this could imply regulation by DNA supercoiling, but in the opposite fashion to that speculated about with regard to *whi*B). In addition to this, like *gal*-p2, *bld*A-p contains a PPCS Class E sequence very similar to the putative consensus sequence for the σ^{28} (position -43, with 8 out of eleven matches).

There are no well studied promoters that fall in the PPCS Classes B and F. It is also difficult to determine the significance, if any of these classes. It is noteworthy, however, that the potential *bld*A gene target ORF1590 promoter (number 95; see also section 1.3.4.2) is of the Class F.

Noteworthy promoters that did not fall in any class include *afs*A-p, *afs*B-p, *afs*R-p and *gal*-p2 (numbers 131, 138, 152 and 126).

3.3.7 Possible promoters in the Bluescript multiple cloning site

It was noted previously (Chapter 2) that the Bluescript multiple cloning site can act to promote transcription in *S.lividans* TK24. The results of the cluster analysis were used to identify *Streptomyces* promoter like sequences found on this DNA fragment. The results are shown in Fig 3.6.

3.3.8 FINDCODE analysis of *Streptomyces* DNA

Numerous different *Streptomyces* genes and gene clusters were tested using FINDCODE (section 3.2.5) and in each case the correct open reading frame was identified (data not shown). Loci tested included *dag*A (Buttner *et al.*, 1987), *whi*B (Davis and Chater, 1992), *whi*G (Chater *et al.*, 1989), *afs*B (Horinouchi *et al.*, 1986), *gln*A (Wray and Fisher, 1988), *glg*B (Bruton, 1993), *bar* (Thompson *et al.*, 1987), *whi*E (Davis and Chater, 1990), *tsr* (Bibb *et al.*, 1985), *sph* (Voegtli and Huetter, 1987), *phr* (Kobayashi *et al.*, 1989), *act*II (ORF1-4; Fernandez-Moreno *et al.*, 1991), *arg*C (Ludovice *et al.*, 1992), *chi*A (Miyashita and Fujii, 1993), *sti*1 (Strickler *et al.*, 1992) and *dna*K (Bucca *et al.*, 1994)

The example shown here is for the *whi*E gene cluster (sections 1.3.4.3 and 1.3.4.6). This sequence was first examined using the protein coding sequence determination computer package CODONPREFERENCE (Genetics Computer Group, Wisconsin, USA). The results are shown in Fig. 3.7. Figure 3.7 A shows the analysis of the gene cluster sense strand, while Fig. 3.7 B shows the antisense strand. Several features are pertinent. It should be noted that because analysis of both strands requires the reversal of the DNA sequence Figs. 3.7 A and B are not aligned with respect to the protein coding regions.

The protein coding regions can easily be determined in Fig. 3.7 A. However, in Fig. 3.7 B it can be seen that the third position bias plot is exactly the same as that of Fig. 3.7 A and, moreover, this tends to align with the very similar codon bias plot. Furthermore, the distribution of rare codons on the antisense strand to a degree mirrors that of the sense strand. While this is not a crucial problem in the example given, these

Figure 3.6 Potential *Streptomyces* promoters in the Bluescript restriction cassette.

Potential promoter -10 and -35 regions are underlined. PPCS classes are shown within dashed lines, and the defining sequence is bold.

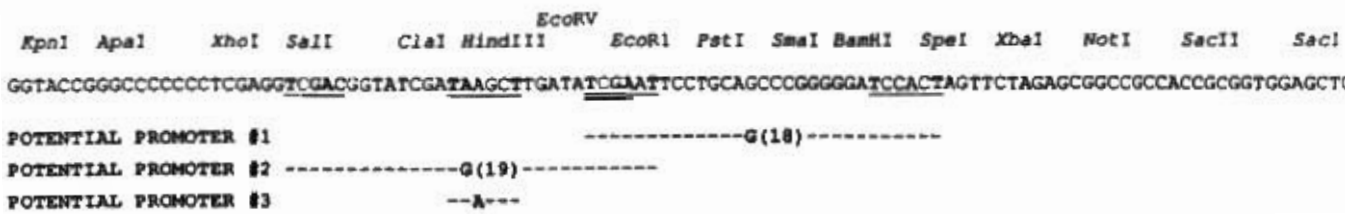
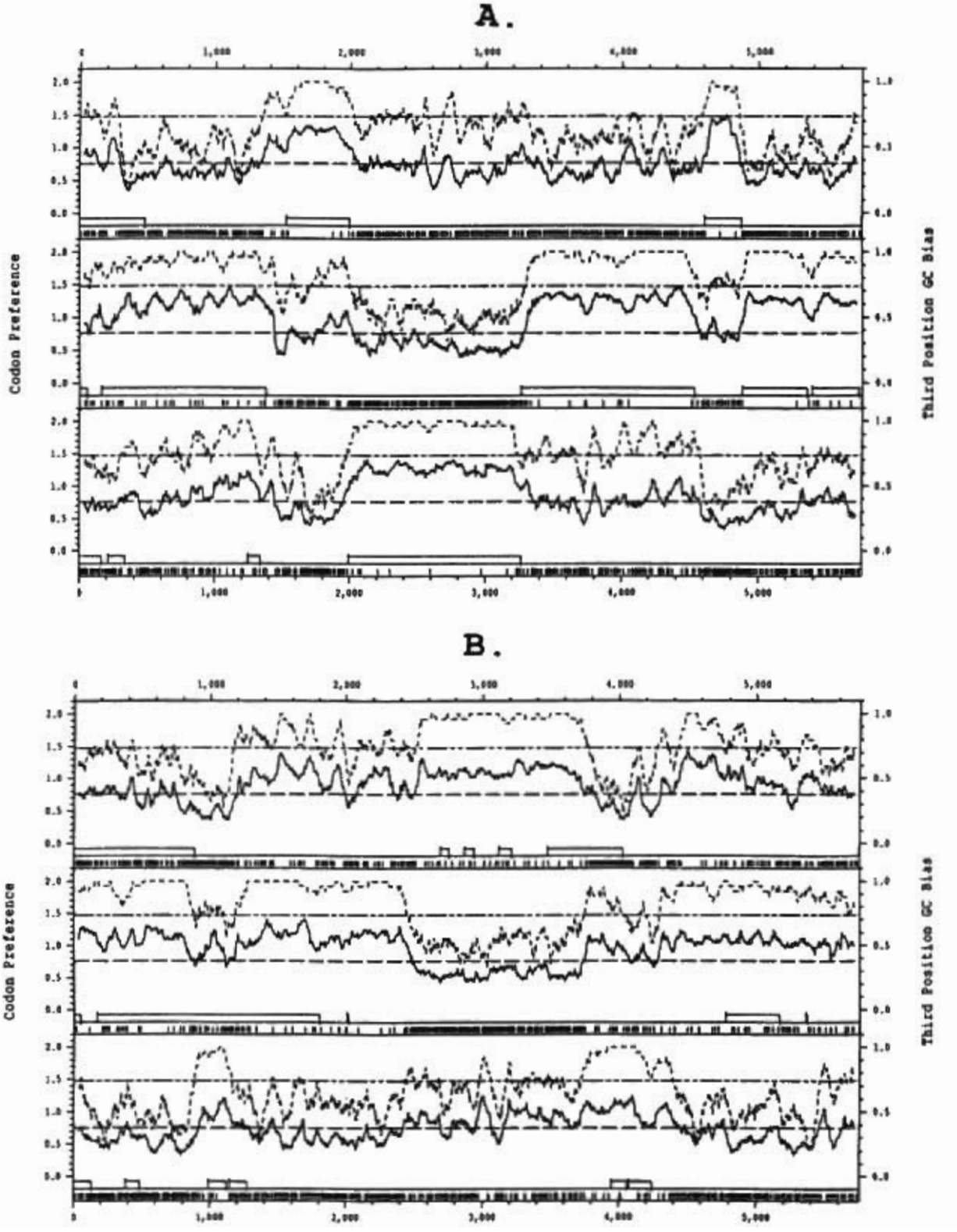


Figure 3.7 FRAME and codon bias analysis of *whiE*

The Genetics Computer Group (575 Science Drive, Madison, Wisconsin, USA) program CODONPREF was used to analyse the *whiE* gene cluster. Third position bias (dotted line), codon bias (solid line), potential ORF search (boxes, only ATG considered as a start codon) and rare codon search (vertical dash). Window size was set at 25 amino acids (75 bases). Rare codon threshold was 0.1.

A. The *whiE* gene cluster sense strand.

B. The *whiE* gene cluster antisense strand.



features make the determination of the limits of protein coding sequences very difficult on short *Streptomyces* DNA sequences.

To test the FINDCODE system, the contribution made by the 5749 base pair *whiE* DNA sequence was removed from the Strep database. FINDCODE analysis was then conducted on the *whiE* gene cluster and the results are shown in Fig. 3.8. Using hexamer data with an averaged window of 150, and starting with each of F1, F3 and F2 data (Fig. 3.8, A, B and C) the protein coding regions are immediately obvious. This is also true when shorter averaged windows are used (Fig. 3.8, D to F) although plots using averaged windows of less than 40 are more difficult to interpret. However, although the protein coding regions become less obvious, this is not true of the protein non-coding regions, or protein coding regions which are tested with data in the wrong frame. These regions have numerous data points at or below an F value of 1. This was never the case for any protein coding region when it was tested with data in the correct frame. The patterns were essentially the same in form using heptamers and octamers, but protein coding regions were even more obvious. (Fig. 3.8 G). The results using oligomers of shorter than six bases were difficult to interpret.

With each of the three tests, only one set of protein coding regions are identified. Which set is identified by each test is dependent on the position (with respect to the first base in the sequence) of the start codon of the protein coding regions. All protein coding sequences that initiate at a position (in number of bases) that is a multiple of three will be identified together. Similarly, those that initiate at a position that is a multiple of three plus one, comprise the second set. The protein coding regions in the third set start at positions that are a multiple of three plus two.

Using the database subsets R1, R2 and R3 (Fig. 3.8 A, B and C) it can be seen that for each test peaks occur for two sets of protein coding sequences. The reasons for this are as follows. Considering a coding region sense strand that is being tested with the data subsets R3, R2, and R1, three possible repeating patterns can be used. Firstly, R3 data is plotted in what is correctly the F1 position, R2 data plotted in the F2 position and R1 data plotted in the F3 position. From the correlation

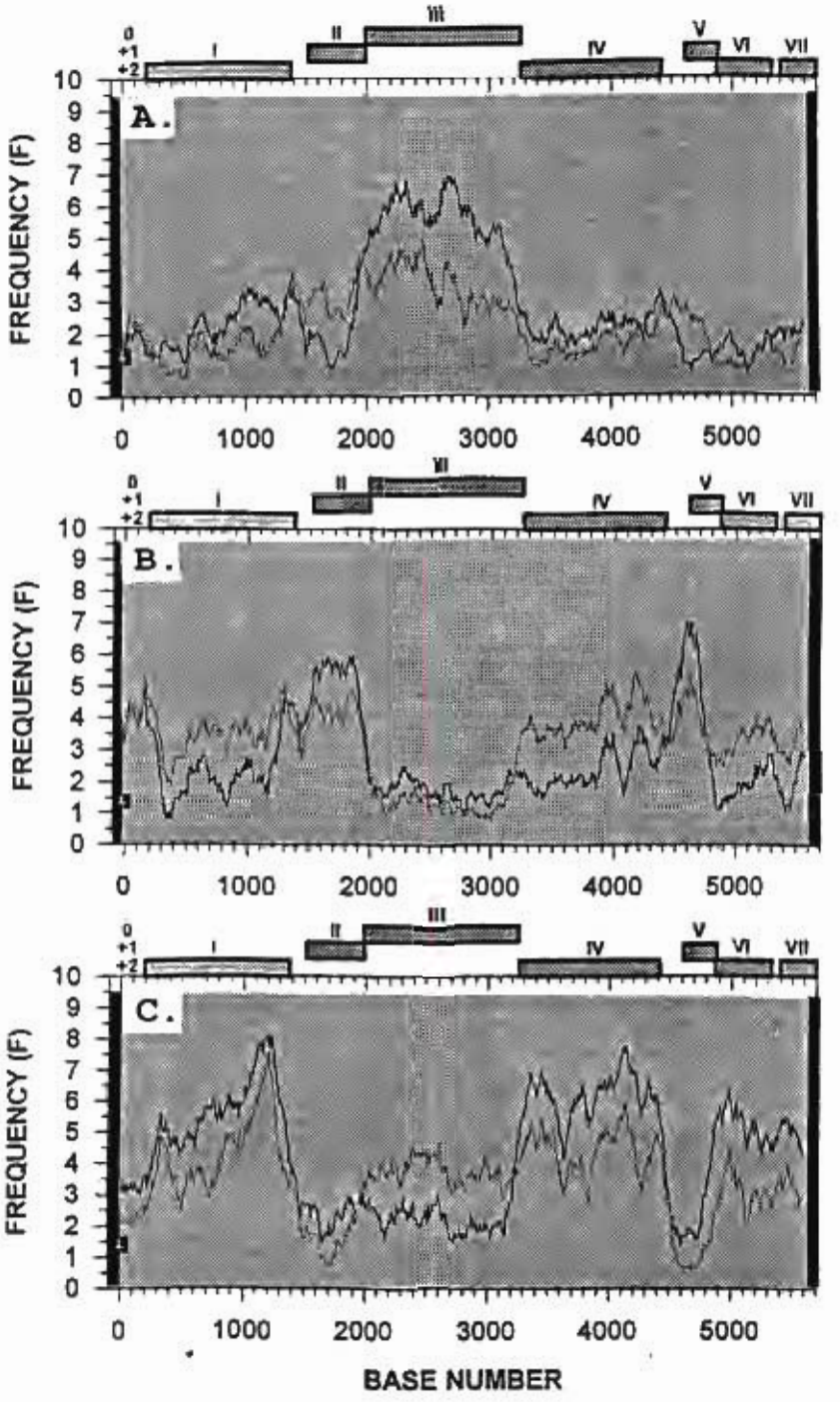
Figure 3.8 FINDCODE analysis of *whiE*

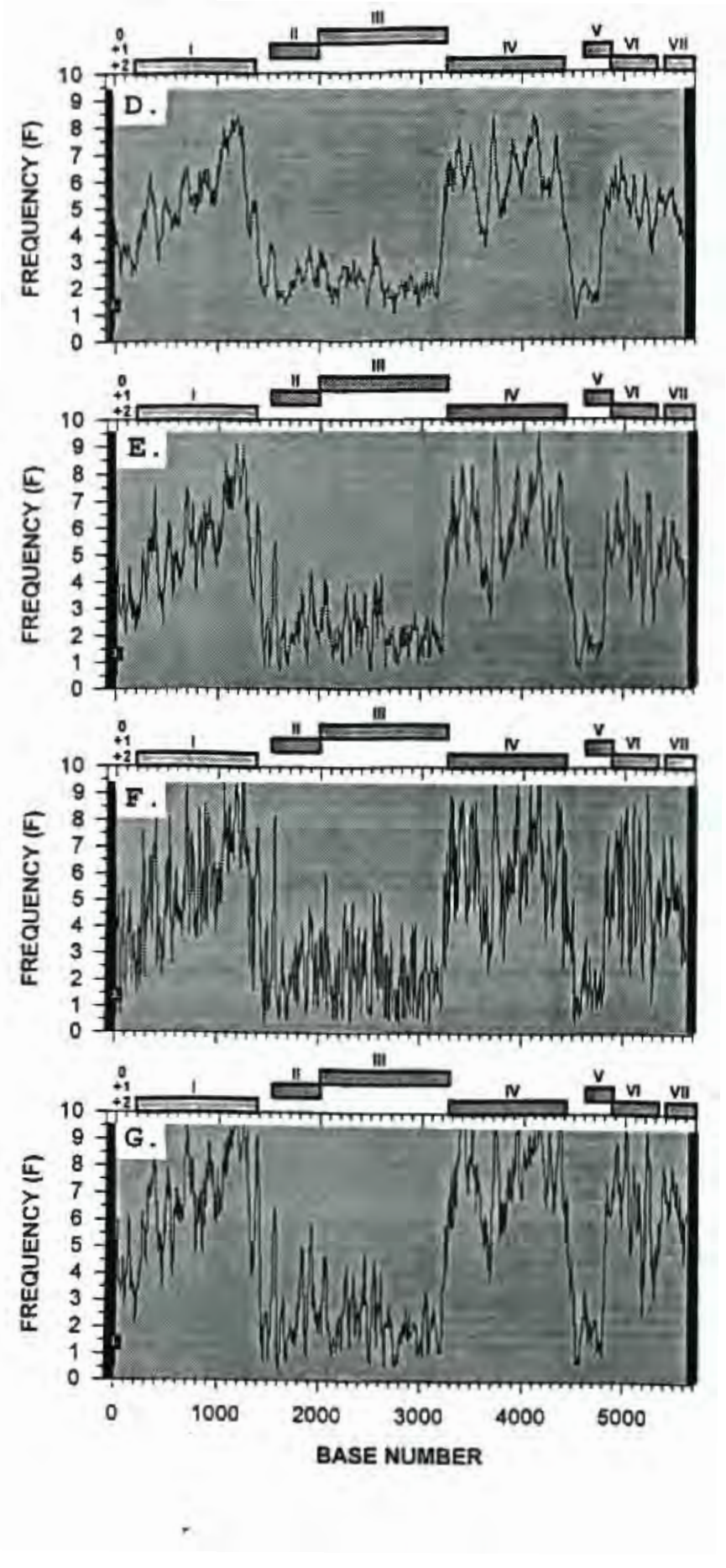
Averaged window frequency plotted against sequence position (in number of bases). **A.** Hexamers, averaged window 150, starts F1 or R1; **B.** Hexamers, averaged window 150, start F3 or R2; **C.** Hexamers, averaged window 150, start F2 or R3; **D.** Hexamers, averaged window 75, start F2; **E.** Hexamers, averaged window 40, start F2; **F.** Hexamers, averaged window 20, start F2; **G.** Octamers, averaged window 40, start F2.

Shaded bars, *whiE* protein coding regions.

Black line, F1, F2 and F3 data; grey line, R1, R2 and R3 data.

Computer program settings, as described in Appendix G, are FINDCODE choice #2, #1 and #3 for graphs A, B, and C respectively.




0
+1
+2
I
II
III
IV
V
VI
VII
D.
E.
F.
G.
FREQUENCY (F)
0 1 2 3 4 5 6 7 8 9 10
0 1000 2000 3000 4000 5000
BASE NUMBER

patterns discussed previously (section 3.3.3, see also Table 3.5 C) it can be seen that a moderate correlation exists between each of these pairs of subsets, implying that the bias is similar and therefore a peak is expected in the FINDCODE plot. In the second instance the pairing is between F2 and R3, F3 and R2, and F1 and R1. Inspection of Table 3.5 C shows that in the case of these pairs the correlation is very high, so again a peak is expected. The same argument follows for the third possibility of pairing between F3 and R3, F1 and R2, and F2 and R1. However, no correlation exists between any of these pairs so no peak is expected or obtained. This observation means that using FINDCODE with *Streptomyces* DNA sequence, it is possible to determine which is the true coding frame, with little doubt as to which strand it is found on.

FINDCODE analysis has a significant advantage over other systems. The small windows needed for FINDCODE ensure that only a short (one effective window length) stretch of bases at the end of the subject sequence is not tested.

Of all the *Streptomyces* genes tested only one, *dagA* (Buttner *et al.*, 1987), showed an averaged frequency of below 3 when tested using FINDCODE using hexamer data in the correct frame (Fig. 3.9 A). This gene was also difficult to identify using codon bias plots (Fig. 3.9 B). It should be noted that the *dagA* start codon is GTG so the ORF identified by CODONPREFERENCE (which utilises only ATG start codons) is shorter than the actual protein coding region.

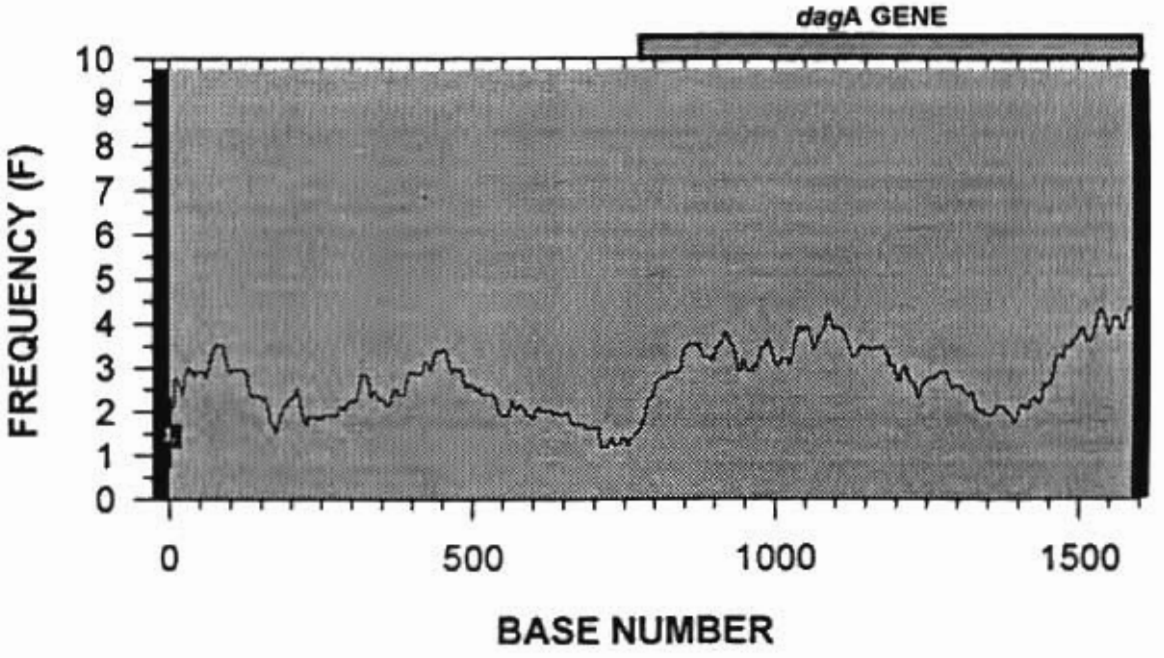
The FINDCODE system has also been tested using both *E.coli* and *B.subtilis* sequences using the Eco and Bsub databases. In these cases protein coding sequences could be determined, but they were not obvious (data not shown). Presumably the system works for streptomycetes because they are so G+C rich. It is probable that it would work for all organisms that have either an A+T or G+C rich genome, although this has not been investigated. To fully use the system with other organisms it would first be necessary to determine the correlation between different frames.

Figure 3.9 FINDCODE analysis of *dag*A gene

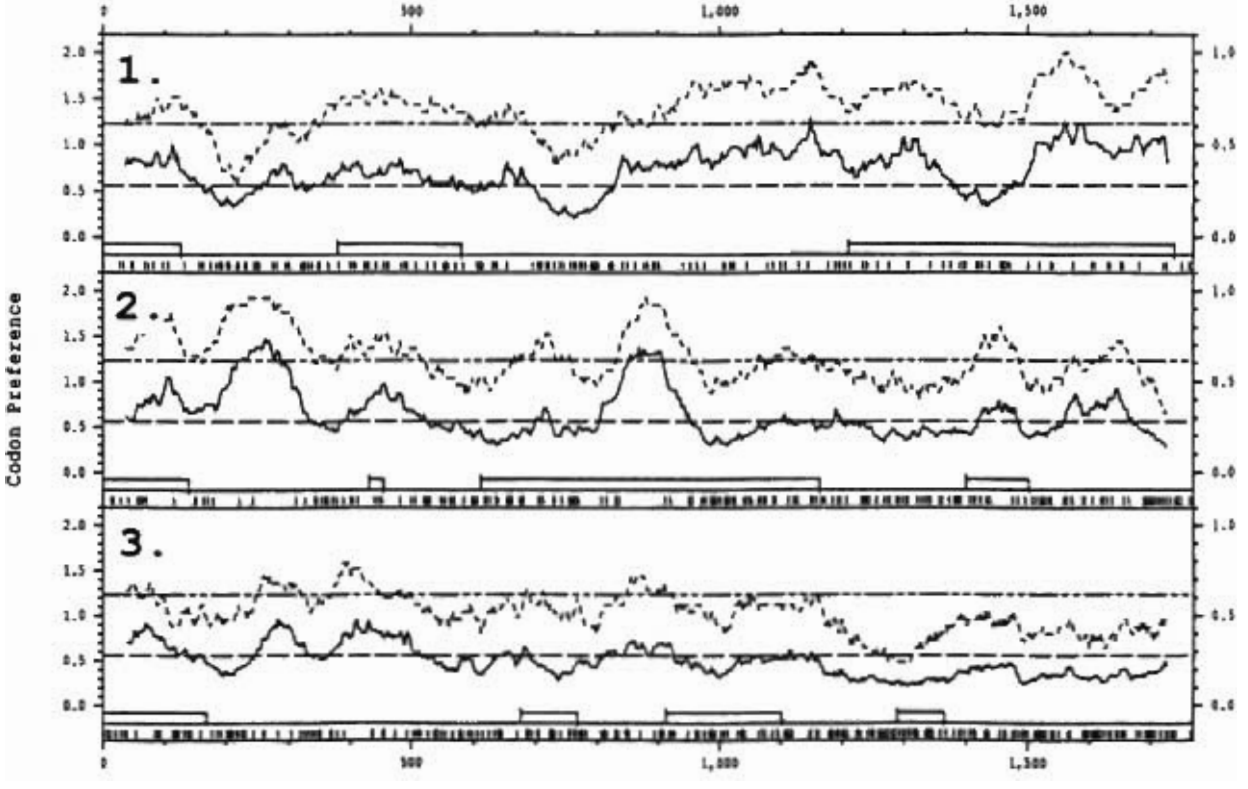
A. Findcode of the *dag*A gene using octamers with an averaged window of 150. The computer settings were such that the *dag*A gene was expected to be identified.

B. CODONPREFRENCE analysis of the *dag*A gene. Settings were as for Figure 3.7.

A.



B.



3.4 Conclusion

The study of the genetic regulation of *Streptomyces* is in its infancy, yet it is already clear that there are a multiplicity of promoter types in this organism. Such a feature is fundamental to the gene regulation and will certainly have bearing, not only on the general metabolism of the bacteria, but also on their development and antibiotic production. To date there have been no formal studies on what constitutes a *Streptomyces* promoter. While it has been possible to recognise those that bear some similarity to the *E.coli* major class of promoters, there has been no way to estimate the probability that such assignments are correct.

The initial approach (tested here using *E.coli*) of using protein coding sequence to define potential protein-DNA binding sites failed. Whether the idea behind the approach is valid remains questionable. However, the unexplained correlation patterns indicate that unknown factors play a role in protein coding DNA sequence bias and this precludes its use in the manner proposed. The patterns are interesting in themselves and merit further investigation.

The cluster analysis has proved successful in defining some of the *Streptomyces* promoter sequences. However, the results must be viewed with caution, as a definitive classification is not intended. Rather the classification system is a guide to the possibilities, and is meant to assist in the further study of *Streptomyces* promoters. Such future work will need to take the form of mutation studies and *in vitro* transcription experiments. This study should assist in selection of the promoters with which to work and the bases to target.

The approach of using cluster analysis to identify promoters has great potential, particularly as the list of sequenced and mapped *Streptomyces* promoters is growing rapidly. This will also have bearing on the genetics of other actinomycetes, and the system may also be applied to other organisms. The approach is clearly subjective to a degree, particularly with regard to defining what constitutes a cluster in cases where it is not immediately obvious. However, further statistical analysis must

ultimately rely on pre-set conditions, the selection of which is also subjective.

The FINDCODE analysis should also prove to be a useful addition to the methods by which the protein coding regions in *Streptomyces* DNA can be determined.

CHAPTER FOUR

CLONING AND ANALYSIS OF PROMOTERS

Abstract

The promoter probes previously constructed were used to try and isolate late expressing *Streptomyces* promoters. Several attempts failed. On the basis of the presumption that this was because the colonies were utilising all of the tyrosine substrate as a nutrient source before sporulation commenced, the conditions which enabled phosphate limited growth were established. Promoters that expressed upon phosphate limitation were isolated, as was a promoter that appeared to be glucose repressed. Further tests established that three of the phosphate dependent promoters were down regulated in a *S.coelicolor bldA* mutant. These promoters, the glucose repressed promoter and a late expressing promoter that had previously been isolated were sequenced. The previously developed computer programs were used to identify potential promoter sequences and ORFs. During the course of the work it was noted that sporulation by a *bldA* mutant strain of *S.coelicolor* could be induced on glucose minimal medium under phosphate limiting conditions.

4.1 Introduction

It is the usual course, in molecular biological research, that certain genes are selected for study because of the characteristics of the protein that they encode. This is a sensible and valuable approach, particularly if much is already known about the system that is under examination. Where this is not the case, genes are selected for study largely on the basis of the effects of mutations within those genes.

However, the biological function of any gene is not determined only by what that gene encodes, but also by when it is expressed. The timing, or induction of expression can, therefore, also be used as a method of identifying genes that are involved in particular genetic pathways. This approach is particularly useful in cases were identifiable mutants are difficult to isolate, and has the potential to select for genes that would not be isolated by other methods.

Given the number and scale of the attempts to isolate *Streptomyces* developmental mutants, it is surprising how few have been identified (section 1.3.4). Thus, the isolation of late expressed, developmentally regulated promoters is a promising method by which the genes involved in differentiation could be identified. Because of the simplicity of this method it is possible to search for other types of promoters at the same time.

The following work was aimed at identifying developmentally regulated promoters, and glucose repressed promoters, using the *mel* operon as a reporter system.

4.2 Methods

General methods are described in Appendix C. Bacterial strains and plasmids are listed in Appendix B. Materials used are described in Appendix D.

4.2.1 Cloning of promoter bearing fragments

Plasmid pWB150 or pWB151 DNA (isolated from *E.coli* LKIII and *E.coli* GM41 respectively; C.3.2) was digested with *Xba*I and the

recessed ends partially filled in with dGTP, dATP and dTTP (C.4.6). The plasmid used in each case is specified in Results and Discussion (section 4.3). Chromosomal DNA from *S.coelicolor* M130 (C.3.6) was digested with *Sau*3A and the recessed ends were partially filled in with dCTP, dATP and dTTP (C.4.6). The DNA samples were mixed and ligated (C.4.7).

S.lividans TK24 was transformed with the ligation mix (C.2.4). *S.lividans* was used as the host due to its high transformation rate (Hopwood *et al.*, 1985a). Selection was on either MMTCuP medium or MMTCuP1 medium. Where specified, colonies were subcultured on MMTCuP2 and MMTCuGlc(-) medium after transfer with Whatmans 540 paper (C.2.4). MMTCuP, MMTCuP1, MMTCuP2 and MMTCuGlc(-) media are specified in Table 4.1.

4.2.2 Nutrient limited culture tests

Spores of both *S.lividans* TK24 (pWB150) and *S.coelicolor* J1501 (pWB150) were spotted as before (2.2.2) on media containing different levels of either uracil, histidine or K_2HPO_4. For uracil and histidine limitation tests the growth medium consisted of 0.05% L-asparagine, 0.02% $MgSO_4$, 0.001% $FeSO_4$, 1% glucose, 0.01% $CuCl_2.2H_2O$, 0.029 M Na_2HPO_4, 0.021 M NaH_2PO_4, 0.075% L-tyrosine and 2% agar (Oxoid #1). In the first case histidine was at a concentration of 400 μgml^{-1} with uracil at a concentration of 60 μgml^{-1} to 0 μgml^{-1}. In the second case uracil was at a concentration of 60 μgml^{-1} with histidine at concentrations of 400 μgml^{-1} to 0 μgml^{-1}. For phosphate limitation tests the growth medium consisted of 0.05% L-asparagine, 0.02% $MgSO_4$, 0.001% $FeSO_4$, 1% glucose, 0.01% $CuCl_2.2H_2O$, 0.573% TES Buffer (pH 7.0), 0.075% L-tyrosine and 2% agar (Oxoid #1) with concentrations of K_2HPO_4 ranging from 0.025% to 0%.

4.2.3 Promoter expression tests in developmental mutants

Plasmid DNA (specified in Results and Discussion, section 4.3.5) was isolated (C.3.5) and used to transform developmental mutants with selection on YEME medium (C.2.4). The developmental mutant strains utilised were *S.coelicolor* J1501

Table 4.1 Media

	MEDIA NAME				
COMPONENTS	MMTCuP	MMTCuP1	MMTCuP2	MMTCuGlc(-)	MMTCuP4
Casamino acids	0.6%	-	-	-	-
L-asparagine	0.05%	0.3%	0.3%	0.3%	0.3%
$MgSO_4$	0.02%	0.02%	0.02%	0.02%	0.02%
$FeSO_4$	0.001%	0.001%	0.001%	0.001%	0.001%
$CuCl_2.2H_2O$	0.01%	0.01%	0.01%	0.01%	0.01%
K_2HPO_4	0.05%	0.0031%	-	-	0.01%
Agar	2%	2%	2%	2%	2%
Glucose	1%	1%	1%	-	1%
Tiger Milk	0.75%	-	-	-	-
L-tyrosine	0.0375 %	0.0075 %	0.00075 %	0.0075 %	0.0075 %
TES Buffer(pH7)	-	0.573%	-	-	0.573%
Na_2HPO_4	0.029 M	-	0.029 M	0.029 M	-
NaH_2PO_4	0.021 M	-	0.021 M	0.021 M	-
pH	7.0	7.0	7.0	7.0	7.0

(wild type), *S.coelicolor* J1700 (*bldA*), *S.coelicolor* J669 (*bldB*), *S.coelicolor* J660 (*bldC*), *S.coelicolor* 1169 (*bldD*) and *S.coelicolor* J1820 (*whiG*). In each case transformants were isolated and grown in YEME medium before mycelial samples were harvested and resuspended in water (C.1.2). Melanin expression was tested by spotting mycelial samples on MMTCuP4 (Table 4.1) medium as before (section 2.2.2). For all the media used nutrient additives (uracil 7.5 μgml^{-1}, histidine 50 μgml^{-1}, cystine 37 μgml^{-1}, methionine 37 μgml^{-1} and phenylalanine 37 μgml^{-1}) were included, as recommended by Hopwood *et al.* (1985), unless otherwise specified.

4.2.4 Southern blotting

Chromosomal DNA from *S.coelicolor* M130 was isolated (C.3.6). Plasmid pADS911, pLP2, pLP3, pLP13 and pGS100 DNA was isolated from *E.coli* GM41 (C.3.2). DNA samples were subjected to restriction endonuclease digestion (C.4.1; enzymes specified in text, section 4.3.8). Approximately 0.0008 pM of each DNA sample was subjected to agarose gel electrophoresis (C.4.2) and Southern blotting (C.4.9). Probing was with plasmids labelled with ^{32}P by nick translation (C.4.8, specified in section 4.3.6). Membranes were subjected to autoradiography.

4.2.5 Shortening and sequencing of the pGLC100 insert

The procedures utilised below are described in Appendix C: plasmid DNA isolation (C.3.1 and C.3.2), partial and complete restriction endonuclease digestion (C.4.1), ligation (C.4.7), fill in of recessed 3' DNA ends and removal of overhanging 3'DNA ends (C.4.5), partial fill in of recessed DNA ends (C.4.6), agarose gel electrophoresis (C.4.2), transformation of *E.coli* (C.2.2), creation of nested deletions by partial digestion (C.4.12) and sequencing (C.4.10).

The plasmid DNA used in each cloning step was obtained by large scale isolation from the organism in which that particular plasmid was originally constructed, or in which it was originally maintained (Appendix B). Antibiotic selective agents are specified in brackets at concentrations specified in

Appendix B. The identity and structure of each plasmid constructed was confirmed by small scale isolation and restriction mapping.

In order to sequence the pGLC100 promoter bearing fragment it was first inserted into the multiple cloning site of pMTL22. Plasmid pMTL22 DNA was digested with *Hin*dIII and the recessed ends were partially filled in using dGTP and dATP. The recessed ends of *Xba*I digested pGLC100 DNA were partially filled in using dCTP and dTTP. The two samples were ligated. After transformation of *E.coli* LKIII (ampicillin selection) the plasmid pGS100 was identified by isolation and restriction mapping. Shortening of the cloned insert involved three approaches and was as follows. In each case the construct was identified by isolation and restriction mapping.

In the first instance nested deletions of plasmid pGS100 were created using the partial digest method. Partial digestion was with *Hha*I and was followed by complete digestion with either *Xba*I or *Pst*I.

In the second case shortening was achieved by digestion of pGS100 with *Kpn*I followed by recircularisation ligation and transformation of *E.coli* LKIII.

The final plasmid was constructed by partial digestion of pGS100 with *Kpn*I, gel purification of the linear plasmid, digestion of the DNA with *Sal*I, conversion of the overhanging ends to blunt ends, recircularisation by ligation and transformation of *E.coli* LKIII (ampicillin selection).

Sequence data was obtained by sequencing plasmid pGLC100 using the MEL or ORI primers, or by sequencing the shortened constructs using the 1212 primer or 1233 primer (specified in Results and Discussion; see also Appendix D for primer sequence).

4.2.6 Sequencing of the pLP2, pLP3 and pLP13 inserts.

Plasmids pLP2, pLP3 and pLP13 were sequenced (C.4.10) using the MEL and ORI primers (specified in Results and Discussion; see also Appendix D for primer sequence).

4.2.7 Shortening and sequencing of the pPS9 insert

The procedures utilised below are described in Appendix C: plasmid DNA isolation (C.3.1 and C.3.2), partial and complete restriction endonuclease digestion (C.4.1), CIAP treatment (C.4.11), ligation (C.4.7), fill in of recessed 5' DNA ends and removal of overhanging 3'DNA ends (C.4.5), partial fill in of recessed DNA ends (C.4.6), agarose gel electrophoresis (C.4.2), transformation of *E.coli* (C.2.2), creation of nested deletions by partial digestion or using Bal31 (C.4.12 and C.4.13) and sequencing (C.4.10).

The plasmid DNA used in each cloning step was obtained by large scale isolation from the organism in which that particular plasmid was originally constructed, or in which it was originally maintained. In all cases the ligation mix was used to transform *E.coli* LKIII, with ampicillin selection (100 $\mu g ml^{-1}$). The identity and structure of each plasmid constructed was confirmed by small scale isolation and restriction mapping.

In order to shorten the *S.coelicolor* M130 chromosomal DNA insert in pPS9 the DNA fragment was first subcloned. Plasmids pPS9 and pUC19 were digested with *Pst*I. The pUC19 DNA was then treated with CIAP. The DNA samples were mixed, ligated and after transformation of *E.coli*, plasmid pADS91 was identified. Following this, pADS91 DNA was digested with *Sph*I and recircularised by ligation. After transformation of *E.coli* plasmid pADS911 was identified.

Several approaches were adopted in the creation of nested deletions suitable for sequencing. In the first procedure plasmid pADS91 DNA and Bal31 exonuclease was used, with the initial restriction endonuclease digestion utilising *Bgl*II while the second digestion was with *Sph*I (C.4.13). Nested deletions were also created from plasmid pADS911 by partial restriction endonuclease digestion with *Hha*I followed by digestion with *Xba*I (C.4.12). Specific deletions were created as follows: *Sal*I digestion of pADS911 followed by self ligation; *Xho*I and *Xba*I digestion of pADS911 followed by conversion of overhanging DNA ends to blunt ends and recircularisation by ligation; *Hind*III and *Apa*I digestion of

pADS911 followed by conversion of overhanging DNA ends to blunt ends and recircularisation by ligation.

4.3 Results and Discussion

4.3.1 Initial attempts to isolate developmentally regulated promoters

*Sau*3A digested *S.coelicolor* M130 chromosomal DNA fragments were ligated into the *Xba*I site upstream of the *mel* reporter operon of either pWB150 or pWB151. *S.lividans* TK24 colonies carrying these constructs were grown on MMTCuP medium (4.2.1, Table 4.1). In a number of independent experiments, several thousand colonies were examined. Those that produced black pigment at the onset of sporulation were identified and spores from these colonies were streaked on MMTCuP medium to test if the regulated melanin production was reproducible.

The initial results obtained were as expected in that approximately 10% of the colonies eventually turned black and some of the colonies showed delayed pigmentation. However, very few colonies displayed concomitant melanin production and sporulation, and in these cases the pigment was produced at very low levels. Furthermore, upon retesting those clones, melanin production was erratic, and was usually not visible.

4.3.2 Nutrient limited growth

Although it had previously been shown that it was possible to isolate late expressed promoters using the *mel* reporter system, this had only been done on complex medium (section 2.3.4). The attempts, as described above (section 4.3.1), to isolate such promoters utilised a defined medium. The fact that no late expressing promoters could be isolated suggested that the medium might somehow be unsuitable for the purpose. It should be noted that the factors that induce sporulation on either complex or defined medium are unknown. Possibly the limitation of one or a number of nutrients, or other factors such as the build up of toxins plays a role. The factors are likely to be different for different media.

It is possible to explain why late expressed promoters were not identified using defined medium if certain assumptions are made. If sporulation on the defined medium is brought about by a limitation of carbon and/or nitrogen then it is probable that the *mel* operon can not be used as a reporter for late expressed genes. This is because the tyrosine that is used as the indicator substrate is an amino acid and so presumably can act as both a carbon and nitrogen source for the growing colony. This means that sporulation may not occur until most or all of the indicator substrate had been utilised for growth. Once this had occurred differentiation could proceed, but any subsequent *mel* operon expression would not be detected. It should be noted that Bascaran *et al.* (1991) have shown that when *Streptomyces clavuligerus* was grown in liquid medium that contained 1% glycerol as a carbon source and approximately twice as much asparagine and half as much phosphate as MMTCuP medium the asparagine was growth limiting.

In order to eliminate the possibility that the surrounding medium was free of tyrosine at the time of sporulation it was therefore necessary to isolate developmentally regulated promoters under conditions where growth was not limited by carbon or nitrogen availability. Growth conditions in which this was the case therefore had to be established.

Two possibilities were tested. In the first case the *his*$^-$ *ura*$^-$ auxotrophic mutant strain *S.coelicolor* J1501 was grown on a series of different media with decreasing amounts of either histidine or uracil present. Such systems have been used to induce sporulation in *B.subtilis* (Lopez *et al.*, 1979). In an alternative approach *S.lividans* TK24 was grown in the presence of different levels of phosphate (section 4.2.2). The results are shown in Table 4.2.

In the case of the auxotrophic mutant, it was clear that reduction in the concentration of the essential supplements resulted in poor growth and an absence of sporulation. However, the pertinent observation was that increased levels of supplements, as far as these were tested, did not cause any visible delay in the timing of sporulation. These results are difficult to interpret fully and it cannot be stated that starvation caused by the absence of either supplement induced

Table 4.2 The effect of nutrients on growth and sporulation[a]

BACTERIAL STRAIN (+pWB150)	BASAL MEDIUM[b]	HISTIDINE CONC. μgml^{-1}	URACIL CONC. μgml^{-1}	PHOSPHATE CONC. (K_2HPO_4) %	GROWTH	SPORULATION
S.coelicolor J1501	#1	400	60	-	Growth	Sporulation
S.coelicolor J1501	#1	200	60	-	Growth	Sporulation
S.coelicolor J1501	#1	100	60	-	Growth	Weak
S.coelicolor J1501	#1	50	60	-	Growth	None
S.coelicolor J1501	#1	25	60	-	Weak	None
S.coelicolor J1501	#1	12.5	60	-	None	None
S.coelicolor J1501	#1	6.25	60	-	None	None
S.coelicolor J1501	#1	3.13	60	-	None	None
S.coelicolor J1501	#1	1.56	60	-	None	None
S.coelicolor J1501	#1	0.00	60	-	None	None
S.coelicolor J1501	#1	400	60	-	Growth	Sporulation
S.coelicolor J1501	#1	400	30	-	Growth	Sporulation
S.coelicolor J1501	#1	400	15	-	Growth	Sporulation
S.coelicolor J1501	#1	400	7.5	-	Growth	Sporulation
S.coelicolor J1501	#1	400	3.75	-	Growth	Weak
S.coelicolor J1501	#1	400	1.88	-	Growth	None
S.coelicolor J1501	#1	400	0.94	-	Growth	None
S.coelicolor J1501	#1	400	0.47	-	Weak	None
S.coelicolor J1501	#1	400	0.23	-	Weak	None
S.coelicolor J1501	#1	400	0.00	-	Weak	None
S.lividans TK24	#2	-	-	0.050	Growth	None[c]
S.lividans TK24	#2	-	-	0.025	Growth	None[c]
S.lividans TK24	#2	-	-	0.0125	Growth	Weak[c]
S.lividans TK24	#2	-	-	0.0063	Growth	Sporulation
S.lividans TK24	#2	-	-	0.0031	Growth	White spores, pigment[d]
S.lividans TK24	#2	-	-	0.0015	Growth	White spores, pigment[d]
S.lividans TK24	#2	-	-	0.0007	Weak	None
S.lividans TK24	#2	-	-	0.0000	Weak	None

[a] Inspected after 5 days of growth.

[b] Media base #1: L-asparagine 0.05%, $MgSO_4$ 0.02%, $FeSO_4$ 0.001%, agar 2%, after autoclaving glucose 1%, $CuCl_2.2H_2O$ 0.01%, Na_2HPO_4 0.029 M, NaH_2PO_4 0.021 M, Tyrosine (added as neutral suspension) 0.075 % (pH 7.5). Media base #2: L-asparagine 0.05%, $MgSO_4$ 0.02%, $FeSO_4$ 0.001%, agar 2%, after autoclaving glucose 1%, $CuCl_2.2H_2O$ 0.01%, TES Buffer (pH 7.0) 0.573%, Tyrosine (added as neutral suspension) 0.075 %.

[c] Sporulation occurred after extended incubation.

[d] Colonies had a white fluffy appearance. Purple pigment appeared in the medium.

sporulation. It is interesting to note that under the conditions used here, poor growth and sporulation occurred at concentrations of supplement that are routinely used for the growth and maintenance of *Streptomyces* auxotrophic mutants (Hopwood *et al.*, 1985). It is possible that this is because large numbers of spores were patched on a limited area of the plate.

In contrast, while very low levels of utilisable phosphate caused poor growth and sporulation, at increased concentrations sporulation occurred (Table 4.2). Remarkably it was found that at phosphate concentrations of between 0.0015% and 0.0031% the sporulating colonies had a white fluffy appearance typical of *whi* mutants (section 1.3.4.3). Furthermore, the pigmented antibiotic actinorhodin was produced, and this is not typical of *S.lividans*, in which actinorhodin expression is much weaker than in *S.coelicolor*. It is important to note that at phosphate levels of 0.025% sporulation was delayed. This may indicate that at concentrations lower than this sporulation was induced by phosphate starvation. Caution should be applied in the interpretation of these results, however. The responses of bacteria to nutrient levels must be complex and subtle, so elucidation of these phenomena requires complex and subtle experiments. Furthermore, it is probable that different parts of the colony experience different environmental conditions and nutrient levels. The test utilised here was crude, considering the subject under investigation. Nutrient levels in the medium were not actually measured and it can therefore only be presumed that sporulation is induced by phosphate starvation. This should be borne in mind when later results (sections 4.3.3, 4.3.5 and 4.3.6) are considered.

4.3.3 Cloning of promoters

*Sau*3A digested *S.coelicolor* M130 chromosomal DNA fragments were ligated into the *Xba*I site upstream of the *mel* reporter operon of pWB151. *S.lividans* TK24 colonies carrying these constructs were grown on 3 types of medium. Colonies from regenerated protoplasts were subcultured on MMTCuP1, MMTCuP2 and MMTCuGlc(-) medium (section 4.2.1; Table 4.1). Melanin

producing colonies that showed different degrees of melanin production on the different media were isolated.

It should be noted that the concentration of tyrosine used in the above media is ten fold less than that used in the test for phosphate limitation. This was because, in preliminary tests, the strong and early production of melanin by promoter carrying clones caused difficulty in distinguishing colonies that displayed late melanin production, unless low numbers of colonies were grown on each plate. The reduction in the amount of tyrosine in the medium resulted in lower levels of melanin production in general, and enabled larger numbers of clones to be screened. This adjustment meant that it was possible that the carbon and/or nitrogen sources might again be growth limiting. However, it can be argued that if late expressed, developmentally regulated promoter bearing clones could be distinguished on MMTCuP1 medium then this implies that sufficient tyrosine was present for the purpose. It should also be noted that these media contained increased concentrations of asparagine in order to try and compensate to some degree for the reduction in carbon and nitrogen source level due to the decrease in tyrosine concentration.

The sole difference between MMTCuP2 and MMTCuGlc(-) medium is the absence of glucose in the latter. It is probable, therefore, that colonies that produced melanin on MMTCuGlc(-) medium but not MMTCuP2 carried copies of glucose repressed promoters cloned in pWB151. In contrast, the main difference between MMTCuP1 and MMTCuP2 lies in the fact that the former contained a limited concentration of phosphate, at a level which might induce sporulation and antibiotic production. Thus, colonies which produced melanin on MMTCuP1 medium but not on MMTCuP2 medium were likely to carry developmentally regulated promoters cloned in pWB151. The phosphate concentration used in MMTCuP1 (0.0031% or 0.2 mM) was selected because this allowed both a reasonable amount of growth and the very early onset of sporulation. A clear difference in the timing of sporulation on phosphate limiting medium and non-phosphate limiting medium was essential for the isolation of developmentally regulated promoters. Over 5500 colonies were screened, of which approximately 92% contained plasmid constructs in which DNA

inserts were found upstream of the *mel* operon. Approximately 20% of these clones produced melanin.

Colonies that showed differences in expression on the different media were used to inoculate YEME medium and subsequently spore suspensions (in water) were produced (C.1.2). Spore samples were spotted onto MMTCuP1, MMTCuP2 and MMTCuGlc(-) medium and the melanin production pattern was monitored. While in the majority of cases differences in expression were small and/or not reproducible, 27 isolates showed consistent melanin production on MMTCuP1 but not MMTCuP2 medium. A single isolate showed consistent strong melanin production on MMtCuGlc(-) medium but not on MMTCuP2 medium. These 27 plasmids were isolated and used to transform *S.coelicolor* J1501 with selection on R2YE medium (C.3.1 and C.2.4). Transformants were isolated and mycelia grown in YEME medium were resuspended in water and spotted onto the different media (C.1.2). Melanin production by the transformants was monitored. Examples are shown in Fig. 4.1. The results were essentially the same as those for *S.lividans* TK24. In cases where the blue pigment actinorhodin was produced, this could easily be distinguished from the black melanin pigment, as has been noted by other authors (Paget *et al.*, 1994). Low levels of melanin production by *S.coelicolor* J1501 (pWB151) were observed in the form of a faint discolouration around the colonies (this was not the case with *S.lividans* TK24 (pWB151); see also Chapter 2, 2.3.10), however such a weak effect did not preclude the use of pWB151 as a promoter probe in *S.coelicolor*.

4.3.4 Testing of a glucose repressed promoter

The single plasmid bearing *S.lividans* TK24 strain which showed glucose repression of melanin production was further tested in liquid medium. Mycelia were grown in YEME medium to promote dispersed growth, before being isolated by centrifugation, washed and resuspended in either MMTCuP2 or MMTCuGlc(-) liquid medium. Melanin expression was monitored. Strong melanin expression was recorded only in the absence of glucose (Fig. 4.2). It was noted that there was a low level of melanin

Figure 4.1 Different melanin production on different media

Plates: **A**, mycelia grown on MMTCuP1 medium; **B**, mycelia grown on MMTCuP2 medium; **C**, mycelia grown on MMTCuGLC(-) medium.
Strains: **1**, *S.coelicolor* J1501 (pWB151); **2**, *S.coelicolor* J1501 (pLP2); **3**, *S.coelicolor* J1501 (pLP3); **4**, *S.coelicolor* J1501 (pLP13); **5**, *S.coelicolor* J1501 (pGLC100).

PLATE A PLATE B

VIEW OF LOWER SIDE OF PETRI DISH

STRAINS
1
2
3
4

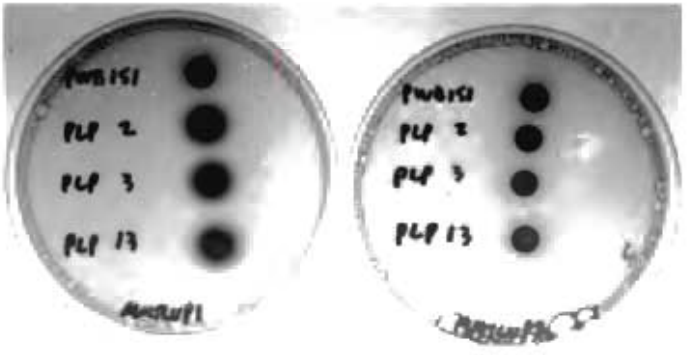


PLATE A PLATE B

VIEW OF UPPER SIDE OF PETRI DISH

STRAINS
1
2
3
4

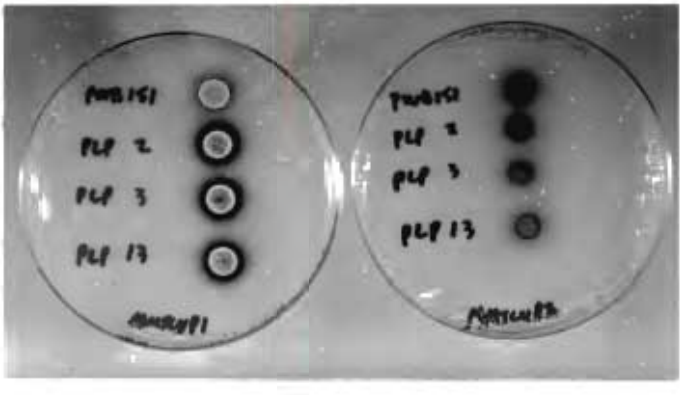


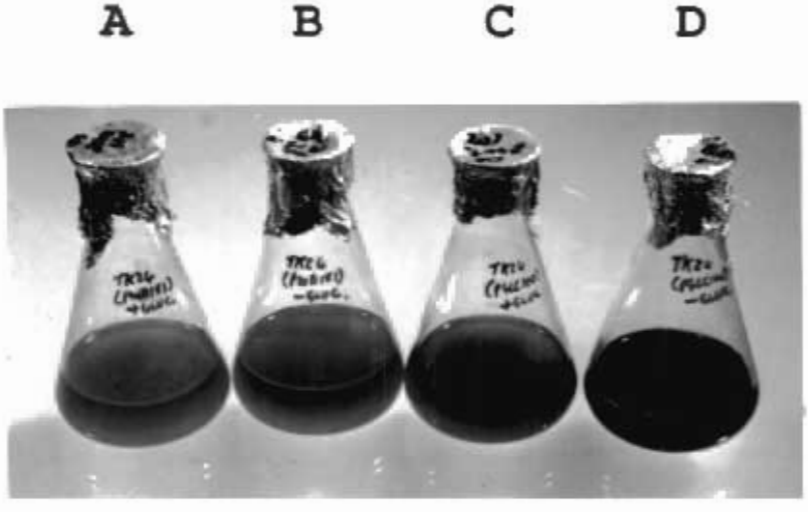
PLATE C PLATE A

VIEW OF LOWER SIDE OF PETRI DISH

STRAINS
1
5

Figure 4.2 Glucose repression of pGLC100 melanin production in liquid medium

A, *S.lividans* TK24 (pWB151) in MMTCuP2 medium; **B**, *S.lividans* TK24 (pWB151) in MMTCuGlc(-) medium; **C**, *S.lividans* TK24 (pGLC100) in MMTCuP2 medium; **D**, *S.lividans* TK24 (pGLC100) in MMTCuGlc(-) medium.



production by *S.lividans* TK24 (pWB151) in both MMTCuP2 and MMTCuGlc(-) medium, which resulted in a grey-green colouration.

The plasmid carrying the glucose repressed promoter was termed pGLC100.

4.3.5 Testing of phosphate dependent promoters

The above procedure, in which promoters that are expressed upon phosphate starvation were isolated, does not ensure that only developmentally regulated promoters are found. The processes undergone by a colony experiencing starvation must be highly complex. Indeed, the line between events associated with starvation or stress and those events associated with differentiation probably exists mainly in the minds of researchers. However, artificial as this distinction is, promoters which are not active in mutants that are blocked at certain stages of development can certainly be termed "developmentally regulated promoters".

The promoters previously isolated were tested to determine which were inactive in developmental (*bld* and *whi*) mutants. Also tested was the previously constructed plasmid pPS9 (which consisted of the promoter probe pWB102 carrying a *S.coelicolor* M130 chromosomal DNA fragment that caused late melanin expression; Chapter 2, 2.3.4). Plasmid DNA from each strain was isolated and used to transform developmental mutants with selection on R2YE medium (C.2.4). The *S.coelicolor* developmental mutant strains utilised were J1700 (*bldA*), J669 (*bldB*), J660 (*bldC*), 1169 (*bldD*) and J1820 (*whiG*) with *S.coelicolor* J1501 (wild type development) as the control. In each case transformants were isolated and grown in YEME medium before mycelial samples were resuspended in water (C.1.2). Melanin expression was retested by spotting mycelial samples on MMTCuP4 (Table 4.1) medium as before (section 2.2.2). The phosphate concentration in MMTCuP4 was set at 0.01% (0.5 mM) as this was the level which, in the case of *S.lividans* TK24, generated the maximum amount of growth while still affecting the timing of sporulation (section 4.3.2).

With each developmental mutant strain tested there was at least one plasmid construct which appeared to be able to cause

melanin production in the parental (wild type for sporulation) strain, *S.coelicolor* J1501, but not the mutant itself (data not shown). It should be noted, however, that the *bldB*, *bldC* and *bldD* mutant strains used here are genotypically different from the sporogenous strain in loci other than the *bld* genes.

Three plasmid isolates, and the plasmid pPS9, showed strong expression of melanin in *S.coelicolor* J1501, with weak or no expression in *S.coelicolor* J1700. The three plasmids were termed pLP2, pLP3 and pLP13. *S.coelicolor* J1501 and *S.coelicolor* J1700 strains carrying these plasmids were again tested by patching them on MMTCuP4 medium as before. In this case however, only histidine and uracil were added as nutrient supplements at concentrations of 500 $\mu g ml^{-1}$ and 75 $\mu g ml^{-1}$ (or 10 fold the previous and recommended level). This was to ensure that vigorous growth and sporulation occurred when possible, as it had previously been noted (section 4.3.2) that low levels of supplements affected development. The results are shown in Fig. 4.3. Again it was seen that none of the plasmids caused strong melanin production by *bldA* mutants. Remarkably, in a later repetition of this work, it was noted that upon extended incubation of over a week, the *bldA* mutants started to produce an orange pigment and sporulate weakly (Fig. 4.3). This is discussed further below (section 4.3.6).

It is not possible to state with certainty what factor limited growth and/or induced sporulation in this experiment as the previous phosphate limitation tests were conducted with *S.lividans* TK24 and not *S.coelicolor* J1501. Furthermore, it is difficult to ascertain the effect that the auxotrophic phenotype of the *S.coelicolor* strains used, and consequent addition of supplements to the media, had. However, it is plain from Fig 4.3 that at the time the cloned promoters were activated, there remained in the medium sufficient tyrosine for the *mel* operon to act as a reporter.

4.3.6 Sporulation of a *bldA* mutant on glucose medium

In the course of work not reported here the *bldA* mutant *S.coelicolor* J1700 carrying various plasmids (including pIJ702, pLP2, pLP3, pLP13 and pWB102, among others) was plated on

Figure 4.3 Melanin expression in developmental mutants

Mycelia were grown on MMTCuP4 medium (see section 4.3.5).

Strains: **1**, *S.coelicolor* J1501 (wild type sporulation); **2**, *S.coelicolor* J1700 (*bldA*);.

Plasmids: **A**, pWB151; **B**, pLP2; **C**, pLP3; **D**, pLP13; **E**, pWB102; **F**, pPS9.

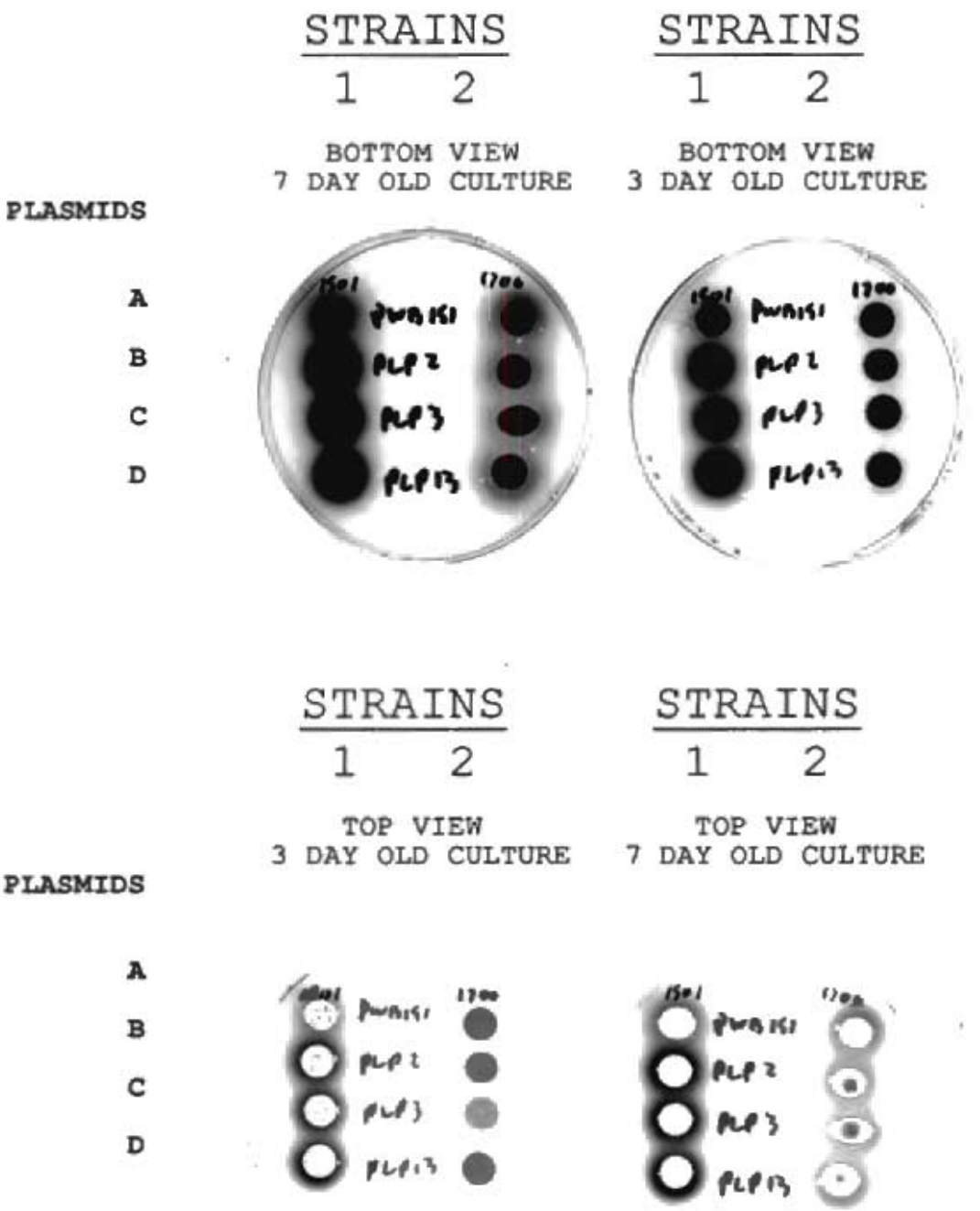
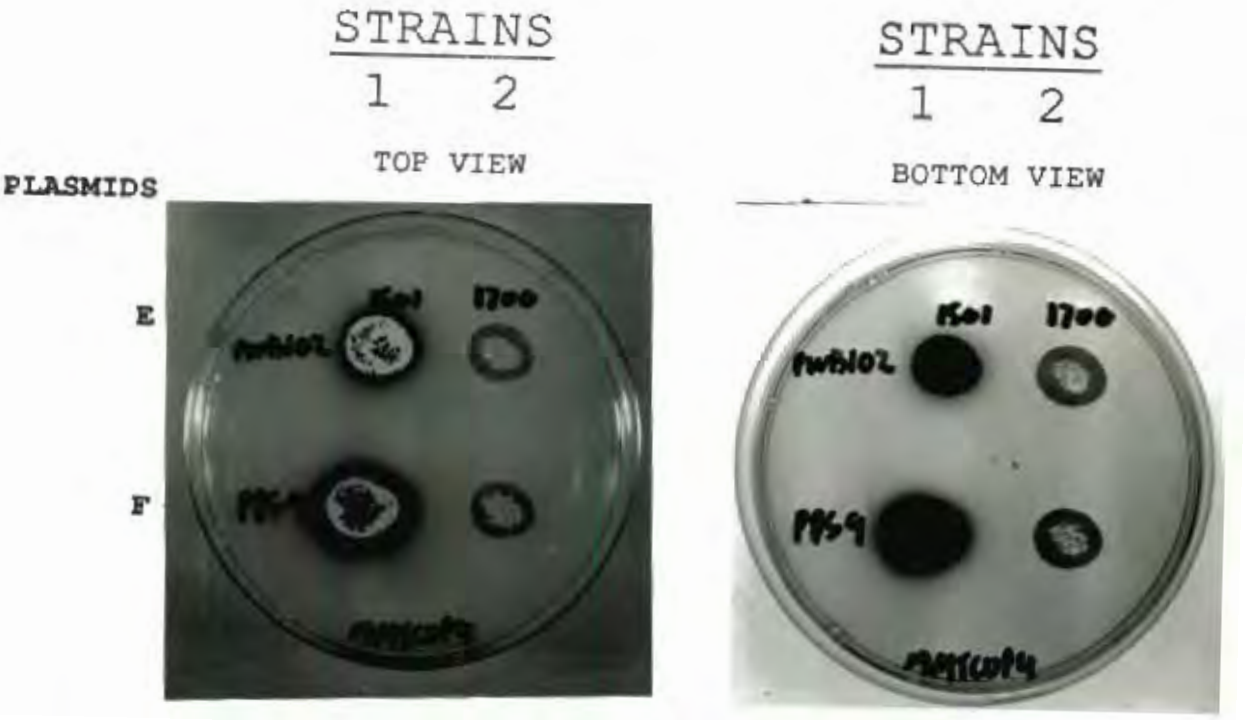


Figure 4.3 Continued



medium containing different concentrations of phosphate. Mycelia were spotted on the medium as before (section 4.2.3). In this case, however, the media were identical to MMTCuP4 medium except that the concentrations of K_2HPO_4 used were set at 0%, 0.002%, 0.02% and 0.2% and the histidine and uracil nutrient additives were supplied at concentrations of 400 $\mu g ml^{-1}$ and 60 $\mu g ml^{-1}$ (or 8 fold the recommended level). The results are shown in Fig. 4.4. At low levels of phosphate (0% and 0.002% or 0 mM and 0.2 mM) production of an orange pigment, presumably undecylprodigiosin, was observed. This could be expected as expression of the *red* genes in *bldA* mutants grown on 0.04 mM phosphate has been previously reported (section 1.3.4.2; Guthrie and Chater, 1990). However, as shown in Fig. 4.4, after extended incubation of over one week, weak sporulation of the *bldA* mutant colonies occurred at 0.02% phosphate. In the example shown this was limited to the edge of the colony, but in repeats of the test in some cases the sporulation covered the entire colony. Such sporulation was not observed by Guthrie and Chater (1990) and is, as far as the author is aware, the first report of sporulation of a *bldA* strain on a variation of glucose minimal medium.

As this line of investigation was not pursued further it is unknown if the sporulation of *S.coelicolor* J1700 on phosphate limited medium is as a result of the specific combination of the different additives in the medium, or if this characteristic is displayed by other *bldA* mutant strains.

It should be noted that there appeared to be a small degree of growth even when no phosphate was added to the medium. This was also true with previous tests (see section 4.3.2). This implies that the other components of the medium were not totally free of phosphate or that the mycelium contained phosphate reserves.

4.3.7 The *mel* promoter and expression in a *bld* A mutant

It was previously shown that melanin production by tyrosinase was possible at late stages of colony development (Chapter 2). However, it was not shown that *bldA* mutants could support melanin production. Thus speculation that the pleiotropic

Figure 4.4 *S.coelicolor* J1700 sporulation

Patches of *S.coelicolor* J1700 carrying the plasmids pIJ702 (left colony) and pPS9 (right colony) were grown on MMTCuP4 medium except that the concentrations of K_2HPO_4 used were: **A**, 0%; **B**, 0.002%; **C**, 0.02%; **D**, 0.2%.

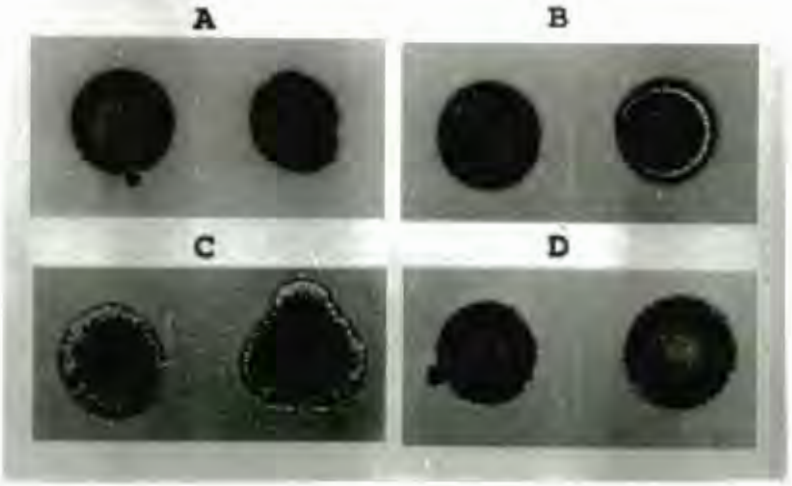
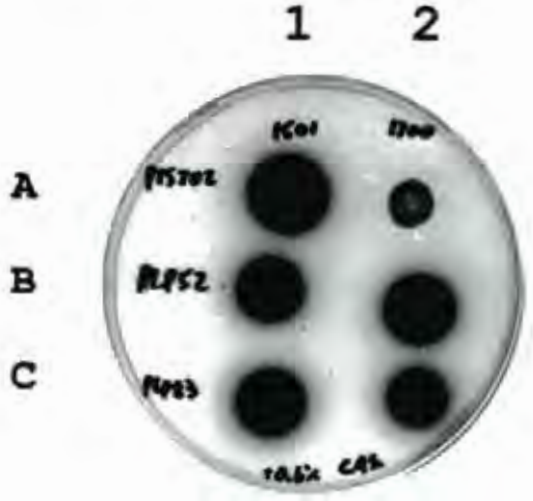


Figure 4.5 Melanin production by pIJ702 in a *bldA* mutant

Patches were grown on MMTCuP4 medium supplemented with 0.6% casamino acids. The photograph shows the Petri dish viewed from the underside.

Columns: **1**, *S.coelicolor* J1501; **2**, *S.coelicolor* J1700. Each carries the plasmid designated by row number.

Rows: **A**, pIJ702; **B**, pLP52; **C**, pLP83.



effects of the *bld* lesion could result in cellular changes that prevent the action of tyrosinase remained plausible.

In order to eliminate this possibility *S.coelicolor* J1501 (pIJ702) and *S.coelicolor* J1700 (pIJ702) mycelia were spotted onto MMTCuP4 medium that had been supplemented with histidine and uracil as before and also with 0.6% casamino acids. It was expected that melanin production by both strains would be observed as casamino acids had previously been shown to be required for induction of *mel* expression from pIJ702 (section 2.3.3). The results are shown in Fig. 4.5. Remarkably no melanin production by the *bldA* mutant was observed. Repetition of the experiment utilising freshly transformed bacteria gave the same result. Parallel experiments, however (reported above, see section 3.4.5), in which cloned fragments of *S.coelicolor* chromosomal DNA were inserted upstream of the *mel* reporter operon of pWB151 gave different results. In this case numerous plasmids displayed the ability to promote melanin production in *S.coelicolor* J1700. Typical examples are shown in Fig 4.5. Thus it is clear that melanin production by *bldA* mutants is possible, and the failure of *S.coelicolor* J1700 (pIJ702) to produce the pigment may be due to regulatory features specific to the original *mel* promoter. As this subject was pursued no further it cannot be presumed that this feature will be found in other strains of *bldA* mutants or even in other culture lines of the same mutant maintained in other laboratories. It is interesting to note, however, that Ochi (1987c) has observed that a in a *rel* mutant of *S.antibioticus* melanin production was abolished. This implies that tyrosinase production might somehow be dependent upon the stringent response. Given the possible links between starvation, the stringent response and sporulation (see chapter 1), it is not improbable that the *mel* operon may appear to be regulated in a fashion usually described as developmentally regulated. Such considerations must be borne in mind when the *mel* operon is used as a reporter gene in developmental mutants.

4.3.8 Southern blots with promoter bearing plasmids

To ensure that the cloned promoters were derived from *S.coelicolor* M130, Southern blotting analysis was conducted (section 4.2.5). The results are shown in Fig. 4.6. In the case of pPS9 and pGS100 (for construction of pGS100 and pADS911 see below, section 4.3.9) bands of the same, correct, size and pattern are generated by both the plasmid and chromosomal DNA samples (Fig. 4.6 A and B). This indicated that the source of the cloned DNA was *S.coelicolor* M130.

In the case of plasmids pLP2, pLP3, and pLP13 the size of the chromosomal DNA inserts, as gauged by restriction mapping, was 100 bp or less. As such small fragments are unsuitable for use in Southern blots due to weak binding to the filter (Maniatis *et al.*, 1982) *Sau*3A digests were not tested. Thus the evidence that the DNA inserts were derived from *S.coelicolor* M130 is not as strong as for the other plasmids. However, using other restriction enzymes, each plasmid generated only a single band (Fig. 4.6 C, D and E). In two cases, however, no obvious bands were observed (Fig. 4.6 D2 and E1) possibly because the DNA fragments were too small or (more probably) too large for efficient detection. The bands which do appear are clearly not as a result of hybridisation between the chromosomal DNA and the promoter probe vector pWB151 DNA because the band sizes are all different. These observations, in conjunction with the fact that the DNA of the insert is G+C rich in the manner typical of *Streptomyces* DNA, (see below, section 4.3.9) suggest that *S.coelicolor* M130 was the source of the cloned promoters.

4.3.9 Sequencing and analysis of promoters

The five tyrosinase expressing promoters on plasmids pGLC100 (section 4.3.4), pPS9 (section 2.3.4), pLP2, pLP3 and pLP13 (section 4.3.5) were sequenced. Sequencing protocols are described in Methods (sections 4.2.3, 4.2.4 and 4.2.5). Initially the promoter bearing fragment of plasmid pGLC100 was subcloned to create pGS100. Plasmid construction is shown in Fig. 4.7. The insert was subjected to restriction mapping (C.3.2 and C.4.2). The results are shown in Fig. 4.8. The promoter bearing fragment of pPS9 was also subcloned to create

Figure 4.6 Southern blots

Probes used (also shown in brackets): **A**, pADS911; **B**, pGS100; **C**, pLP2; **D**, pLP3; **E**, pLP13.

Lanes: **A1**, *SphI/PstI* digested *S.coelicolor* M130 chromosomal DNA; **A2**, *SphI/PstI/XhoI* digested *S.coelicolor* M130 chromosomal DNA; **A3**, *SphI/PstI* digested pPS9 DNA; **A4**, *SphI/PstI/XhoI* digested pPS9 DNA; **B1**, *Sau3A* digested *S.coelicolor* M130 chromosomal DNA; **B2**, *Sau3A* pGS100 DNA; **C1**, *XhoI* digested *S.coelicolor* M130 chromosomal DNA; **C2**, *PstI* digested *S.coelicolor* M130 chromosomal DNA; **D1**, *XhoI* digested *S.coelicolor* M130 chromosomal; **D2**, *PstI* digested *S.coelicolor* M130 chromosomal DNA; **E1**, *XhoI* digested *S.coelicolor* M130 chromosomal DNA; **E2**, *PstI* digested *S.coelicolor* M130 chromosomal DNA.

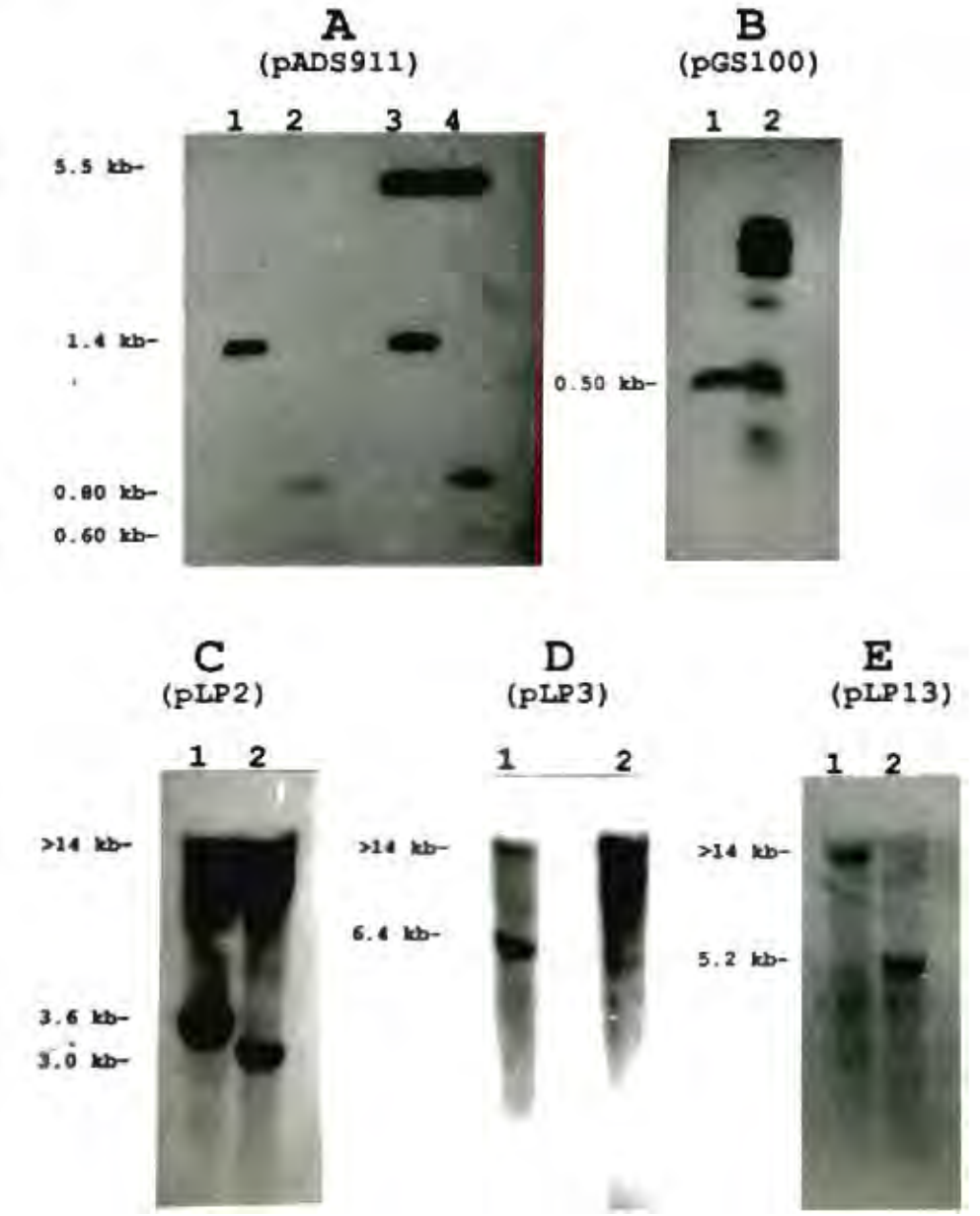


Figure 4.7 Construction of plasmids

A, plasmids digested with *Pst*I, pUC19 treated with CIAP. B, pADS911 digested with *Sph*I. C, *Sau*3A digested chromosomal DNA and *Xba*I digested pWB151 both partially filled in. D, *Hin*dIII digested pMTL22 and *Xba*I digested pGLC100 both partially filled in. Details are given in sections 4.2.1, 4.2.4 and 4.2.6.

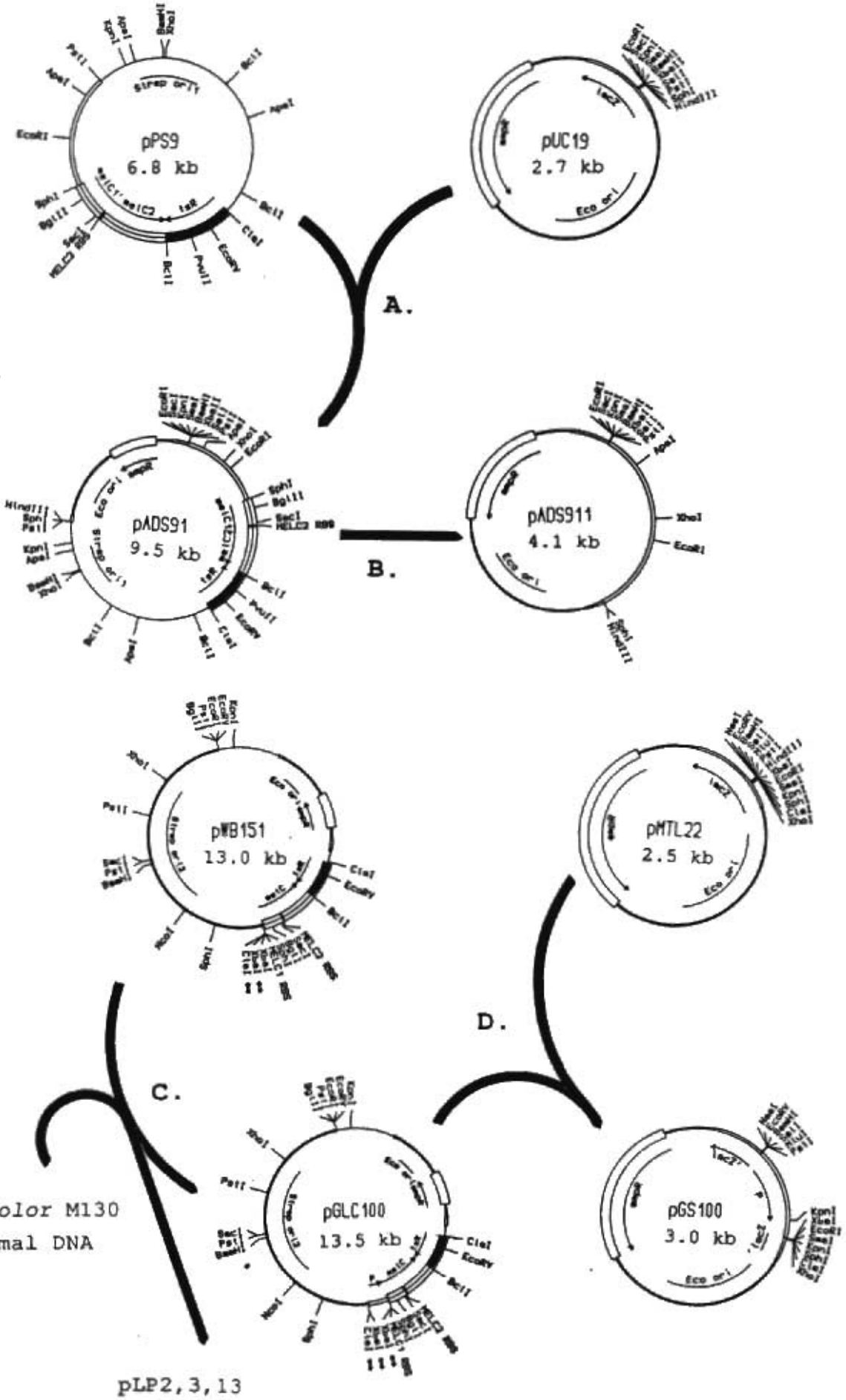
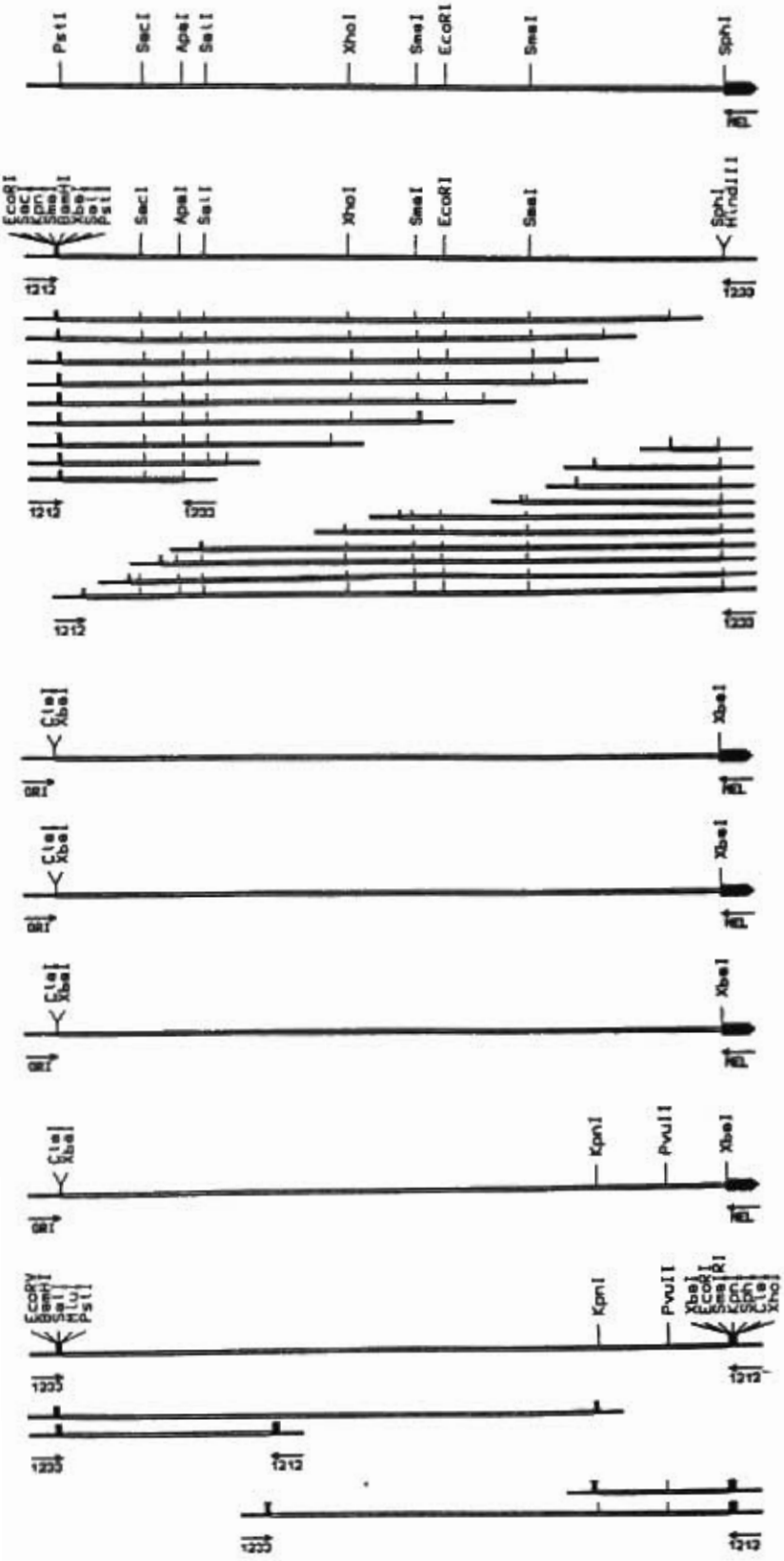


Figure 4.8 Nested deletions and primers used for sequencing



PLASMID NAME	INSERT SIZE	PRIMERS USED FOR SEQUENCING
pPS9	1.36 kb	-
pADS911	1.36 kb	1233, 1212
DELETIONS		1233
		1212
pLP2	0.09 kb	MEL, ORI
pLP3	0.11 kb	MEL, ORI
pLP13	0.14 kb	MEL, ORI
pGLC100	0.48 kb	MEL, ORI
pGS100	0.48 kb	1233, 1212
DELETIONS		1212
		1233

Figure 4.9 Sequence of *S.coelicolor* chromosomal DNA fragments

From plasmids: A, pGLC100; B, pLP2; C, pLP3; D, pLP13; E, pPS9.
Symbols: **A**, **B**, **C**, **D**, **E**, **F**, **G** and **H**, potential promoter core sequences (PPCS) as determined in Chapter 3 and defined in Table 3.3 (double underlined) numbers in brackets represent distance between potential -35 and -10 regions; **RBS**, potential ribosome binding site (underlined); **START**, potential translation start codon (underlined).

A. pGLC100.

```
   1 GATCACGCCGCTCGCTCGCCTCGGTCCTGGCTTCGGGAACCGGACCCGCGCTCCCCGACG
                                                  H(20)
  61 TCGGTTCTTCCGACGACTCGCGCTGCTTCGAACTGTCGGGGGACTCGCCCGCCACCGATG
 121 CCTCCTCCATGCGCCGCACGTCCTGTGCGCGCTGCCGAACCCGTACCGAAACCATTTACC
                                                DE      H(19)
 181 AGTGTCCTGTGTACGGGTGGAACCCACGCGGAAGACGAGAACGACATACCTACTGGTTCC
                          E                                  F
 241 ACTACGAAGCGGTCACGCACCCTCGACAGACCAATGTGAGAGGGGTCACCCTGTCTGTCA
                                              G(19)
 301 TCCACGCGGGGAGGCATGGATGGGCAGGAGCCGCAGAACACTTCCGGAGGAGCTTCTGCT
               RBS      START
 361 GCTGGCGTTGGACCCGACCACGGGTACCACCGCACAGCCGCAGTCGCTCGACCTCGGTCT
 421 GGCCGGAGCACAGCTGGTGGAGCTGGCGCTGGCCGGACGGATAGCCCCAGACGGGGATC
```

B. pLP2.

```
   1 GATCGCCTCATCAAGGCCATCGACCGGTTCGAACTGACCCGAGAGGTCGAGTTCACCTCC
  61 TTCGCCGTGCCGTACATCGTCGGGGAGATC
```

C. pLP3.

```
    1 GATCGTCACTTCGGTCGTTCCCCAGCCAGAACCAGCCATGGCGGAATGCGCCACGTCGCC

   61 CTCCCCTTAGCACGCCATCGAGCTCTCGAGTGGTGGCCGCATCTCAGCACGGATC
```

D. pLP13.

```
    1 GATCTTCCGTACATCCACCACGGGGTGAGCTTCGAGGTGCCGGCTCGGGGGATGACGGCG

   61 TTCGTCGGTCCGTCCGGTGCGGGCAAGACCACCGTGTTCTCCCTCGTCGAGCGGTTCTAC

  121 GACCCGGACTCCGGGGAGATC
```

E. pPS9.

```
    1 CTGCAGCGTCCCACCGCAGCGTCCGCGCGGATCGGACGCGGTCGTGCGCACCTTCCTCGG

   61 GGAGCTCGCGGTGGAGCGTGCCTGTTTTCAGGGCGAACGCGGCGGCGCACCGCCGGGCGG
                             F

  121 GCCCCCCGGAAGCGGCGGCGTGCGCCGGACGACCGTGCCGGGCGCTGGGCACAACGTCGT
                                                  G(19)

  181 GCTCGACCGCCCGACGCGTTGCGCGCGGTCGCCGCGCGGGACCGACTCAGCCCGTGCGGG

  241 TCGCCACGGCCAGGAACCGATCGTCCTCGTCGACGTACGCCGTCATCGCCCATCCGGAAC
                   E                                          E

  301 CGGCCAGCACCGGCCGAGGTTGGCCTCGGCACGAAGGTCGTCCGGTGTGAGCCGCCGCCC
          H(19/20)

  361 CTGCCGTGCCGCGAGCGCCGTCCGGCCGATCGGGTGGAACAGCGCCAGCACACCGCCGGG
                                        E

  421 GCGCACTACGCGGGCCAGCTCCCGCAGGTTTCCGGCCGGATCAGGCAGGTGCGCGACGAG
                             F

  481 ACCCGCCGCGAAGACCGCGTCGAGCGTCCGGGAACGCAGGGGCAACGCGGCCACGTCGGT
                   G(19)              E     H(20)   E

  541 CAGCAGCAGGCGTCCGTCACGGTCCCTCCCGGCCCGTACGGCGGCCTCGAGCATGGCAGG
```

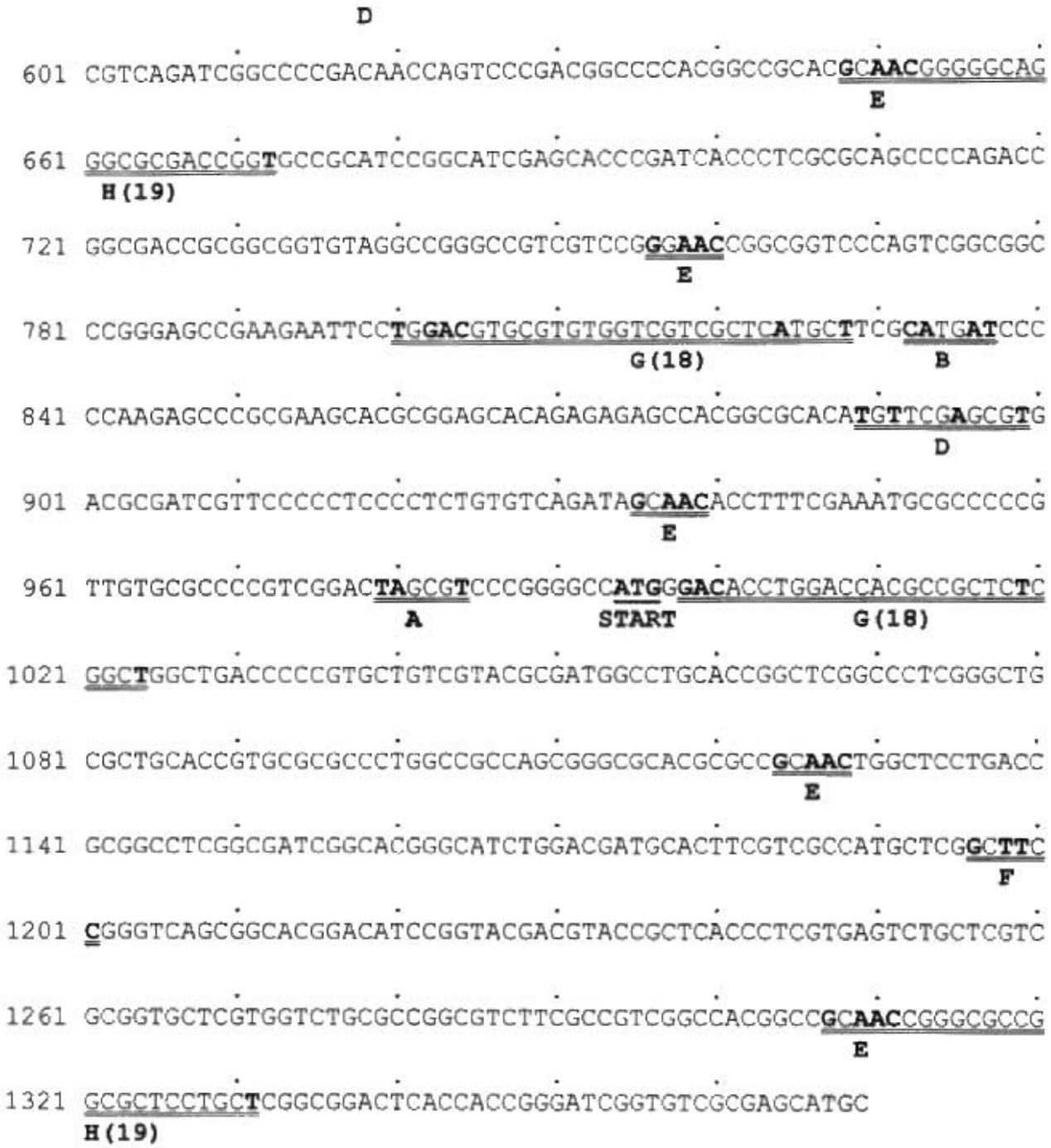

D
601 CGTCAGATCGGCCCCGACAACCAGTCCCGACGGCCCCACGGCCGCACGCAACGGGGGCAG
E
661 GGCGCGACCGGTGCCGCATCCGGCATCGAGCACCCGATCACCCTCGCGCAGCCCCAGACC
H(19)
721 GGCGACCGCGGCGGTGTAGGCCGGGCCGTCGTCCGGGAACCGGCGGTCCCAGTCGGCGGC
E
781 CCGGGAGCCGAAGAATTCCTGGACGTGCGTGTGGTCGTCGCTCATGCTTCGCATGATCCC
G(18)
B
841 CCAAGAGCCCGCGAAGCACGCGGAGCACAGAGAGAGCCACGGCGCACATGTTCGAGCGTG
D
901 ACGCGATCGTTCCCCCTCCCCTCTGTGTCAGATAGCAACACCTTTCGAAATGCGCCCCCG
E
961 TTGTGCGCCCCGTCGGACTAGCGTCCCGGGGCCATGGGACACCTGGACCACGCCGCTCTC
A
START
G(18)
1021 GGCTGGCTGACCCCCGTGCTGTCGTACGCGATGGCCTGCACCGGCTCGGCCCTCGGGCTG
1081 CGCTGCACCGTGCGCGCCCTGGCCGCCAGCGGGCGCACGCGCCGCAACTGGCTCCTGACC
E
1141 GCGGCCTCGGCGATCGGCACGGGCATCTGGACGATGCACTTCGTCGCCATGCTCGGCTTC
F
1201 CGGGTCAGCGGCACGGACATCCGGTACGACGTACCGCTCACCCTCGTGAGTCTGCTCGTC
1261 GCGGTGCTCGTGGTCTGCGCCGGCGTCTTCGCCGTCGGCCACGGCCGCAACCGGGCGCCG
E
1321 GCGCTCCTGCTCGGCGGACTCACCACCGGGATCGGTGTCGCGAGCATGC
H(19)

Figure 4.10 FINDCODE and CODONPREFERENCE analysis

Promoter DNA sequence from: A, pGLC100; **B**, pLP2; **C**, pLP3; **D**, pLP13; **E**, pPS9.

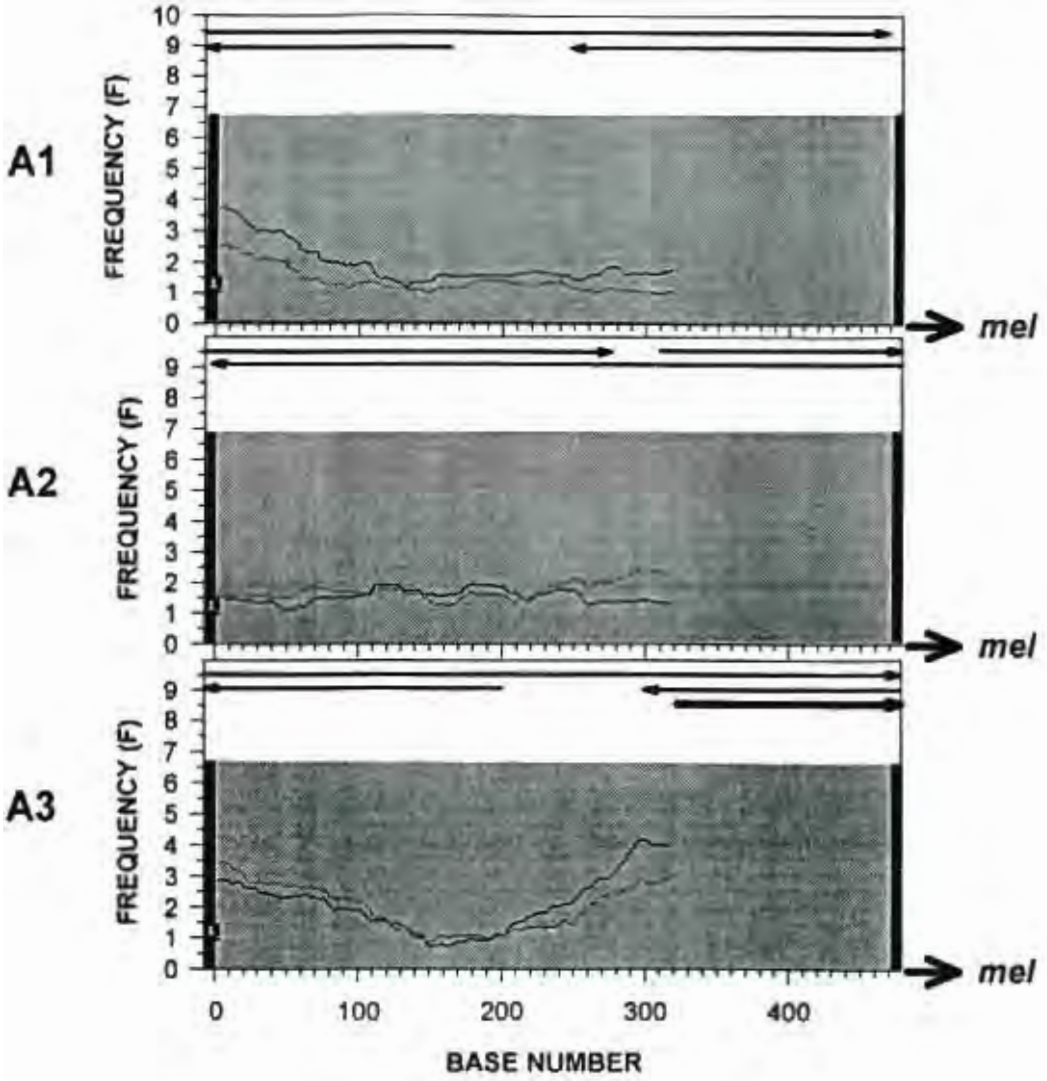
Graph numbers: 1-3, FINDCODE analysis (computer program settings, as described in Appendix G, are FINDCODE choice #2, #1 and #3 for graphs 1, 2, and 3 respectively); **4**, CODONPREFERENCE analysis.

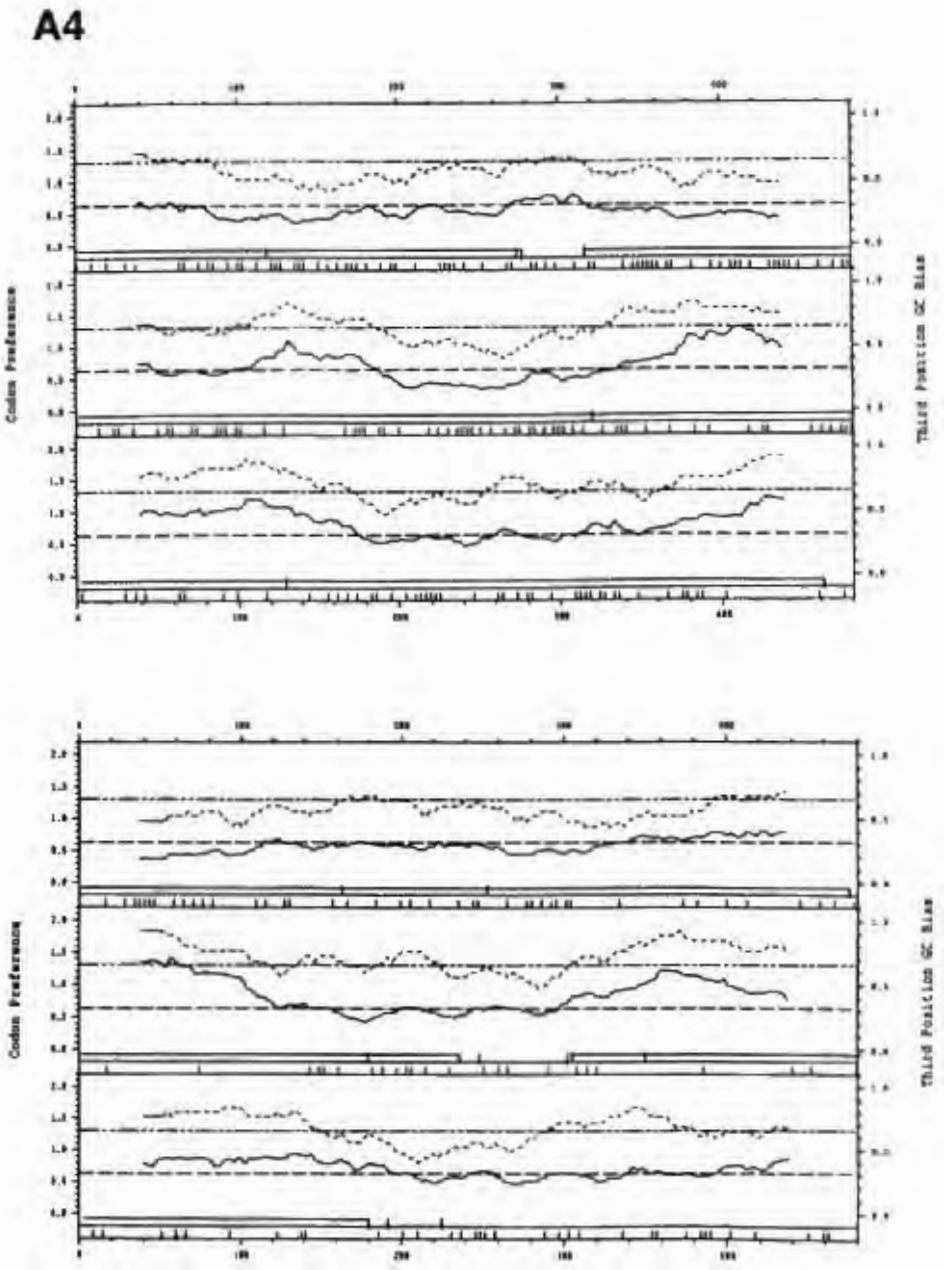
FINDCODE settings and display: Graphs A and E, hexamers with averaged window of 150, Graphs B, C and D, hexamer with averaged window of 20; Black line, F1, F2 and F3 data; grey line, R1, R2 and R3 data; arrows, open reading frames (with ATG and GTG considered as start codons); bold arrows, probable protein coding sequence.

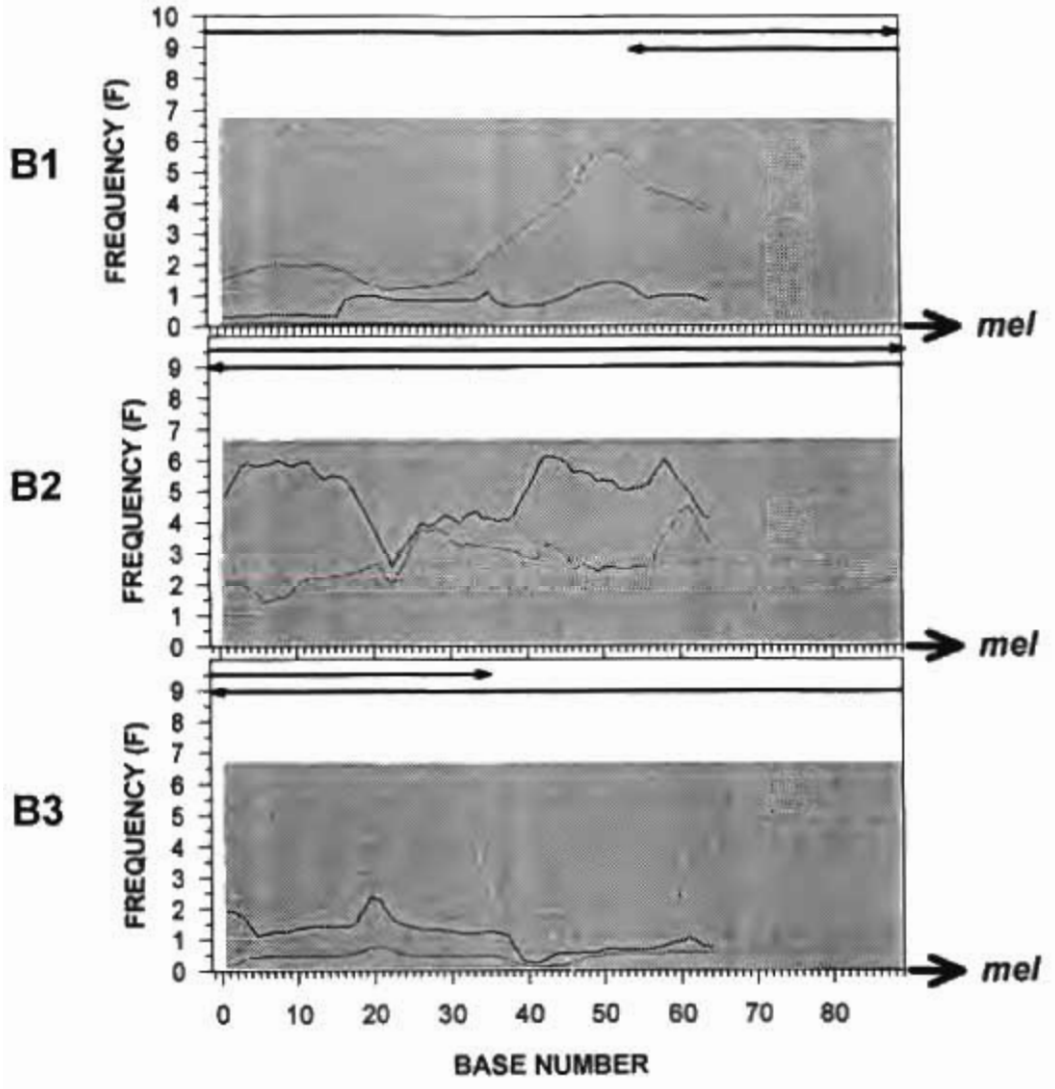
FINDCODE and protein coding region start/stop determination: (black line described only), graph numbers 1, 2, and 3 will record peaks for protein coding regions with start codons at base numbers that are mutiples of 3, mutiples of 3 +1 and mutiples of 3 +2 respectively.

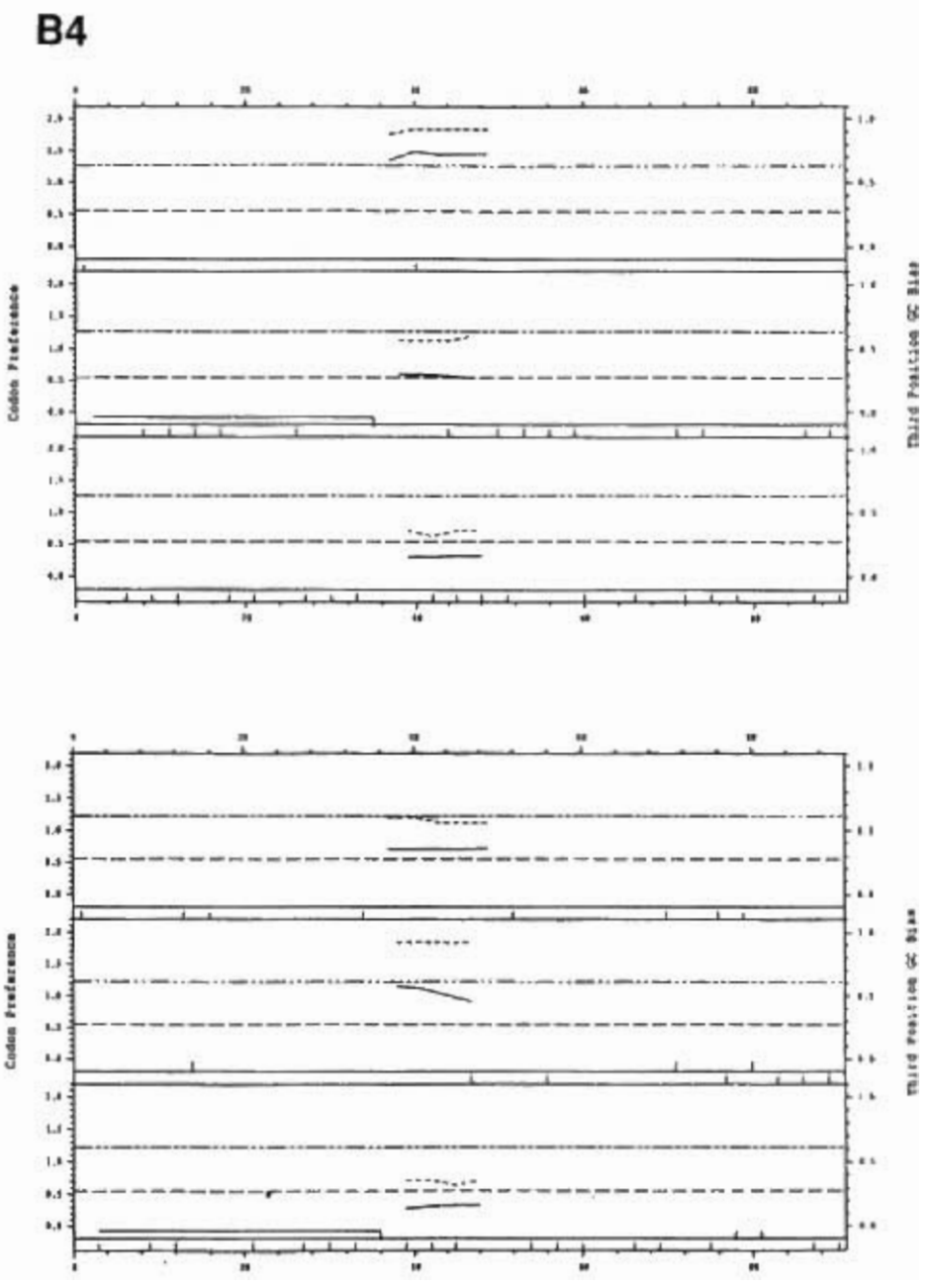
CODONPREFERENCE settings and display: dotted line, third position bias; solid line, codon bias; boxes, potential ORF search (only ATG considered as a start codon); vertical dash, rare codons; window sizes set at 25 amino acids (75 bases); rare codon threshold 0.1; upper graph, DNA strand as shown in Fig. 4.9, lower graph, DNA strand antisense to that shown in Fig. 4.9.

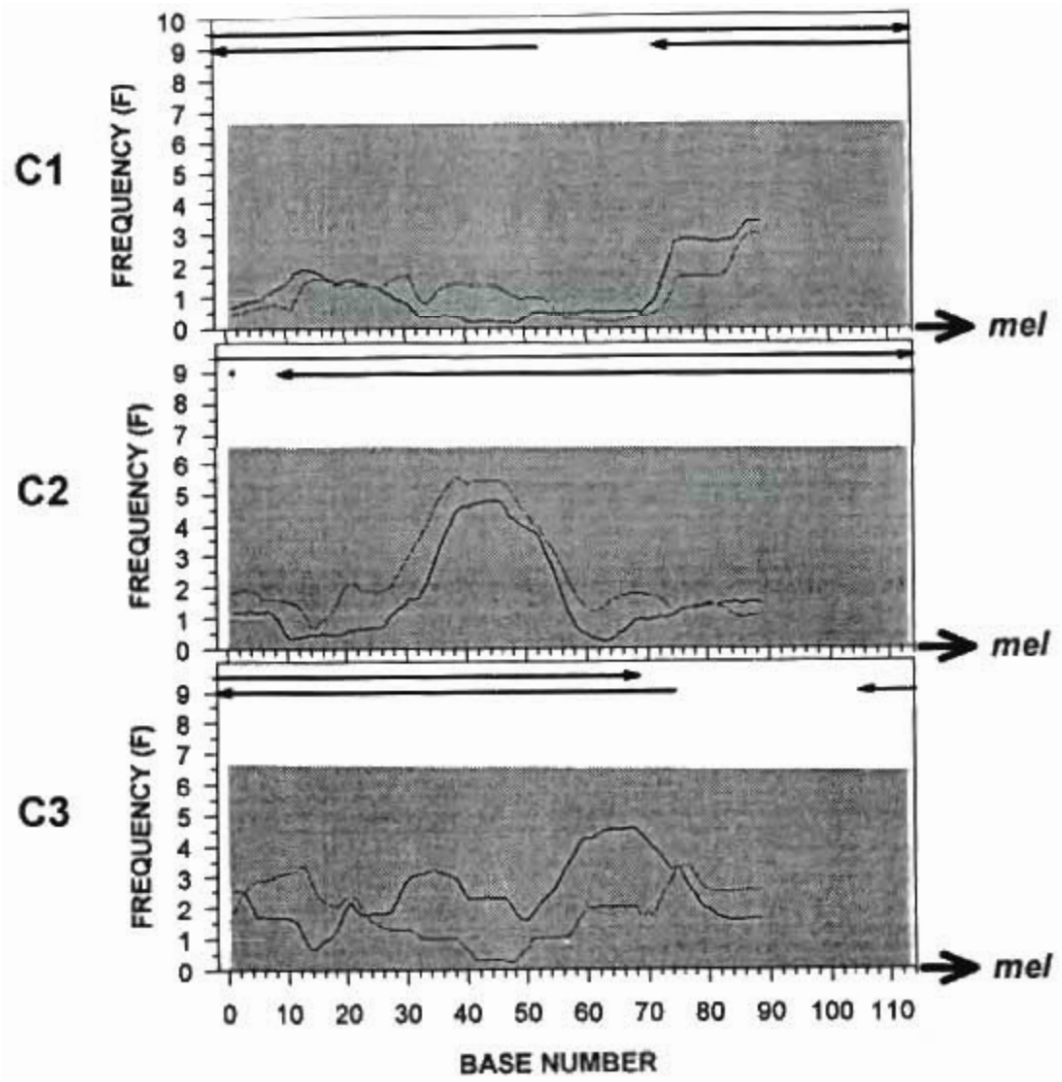
CODONPREFERENCE and protein coding region start/stop determination: Upper graph, in order from the top each of the three plots will record peaks for protein coding regions with start codons at base numbers that are multiples of 3 +1, multiples of 3 +2 and multiples of 3 respectively. Lower graph, the relationship between the protein coding region start/stop and the plot number is dependent on the sequence length and will vary accordingly.


A1
A2
A3
FREQUENCY (F)
mel
BASE NUMBER
0
100
200
300
400

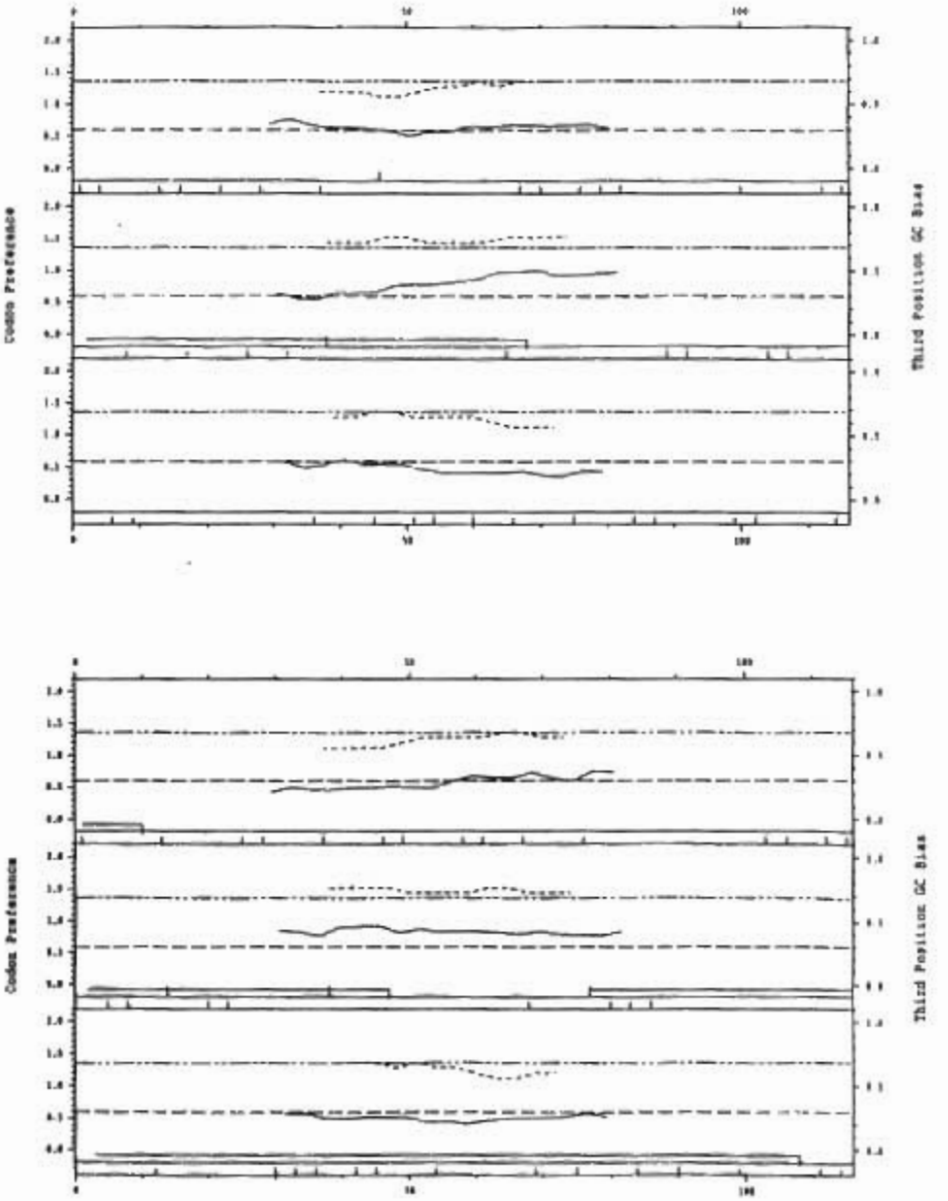

A4
Codon Preference
Third Position GC Bias

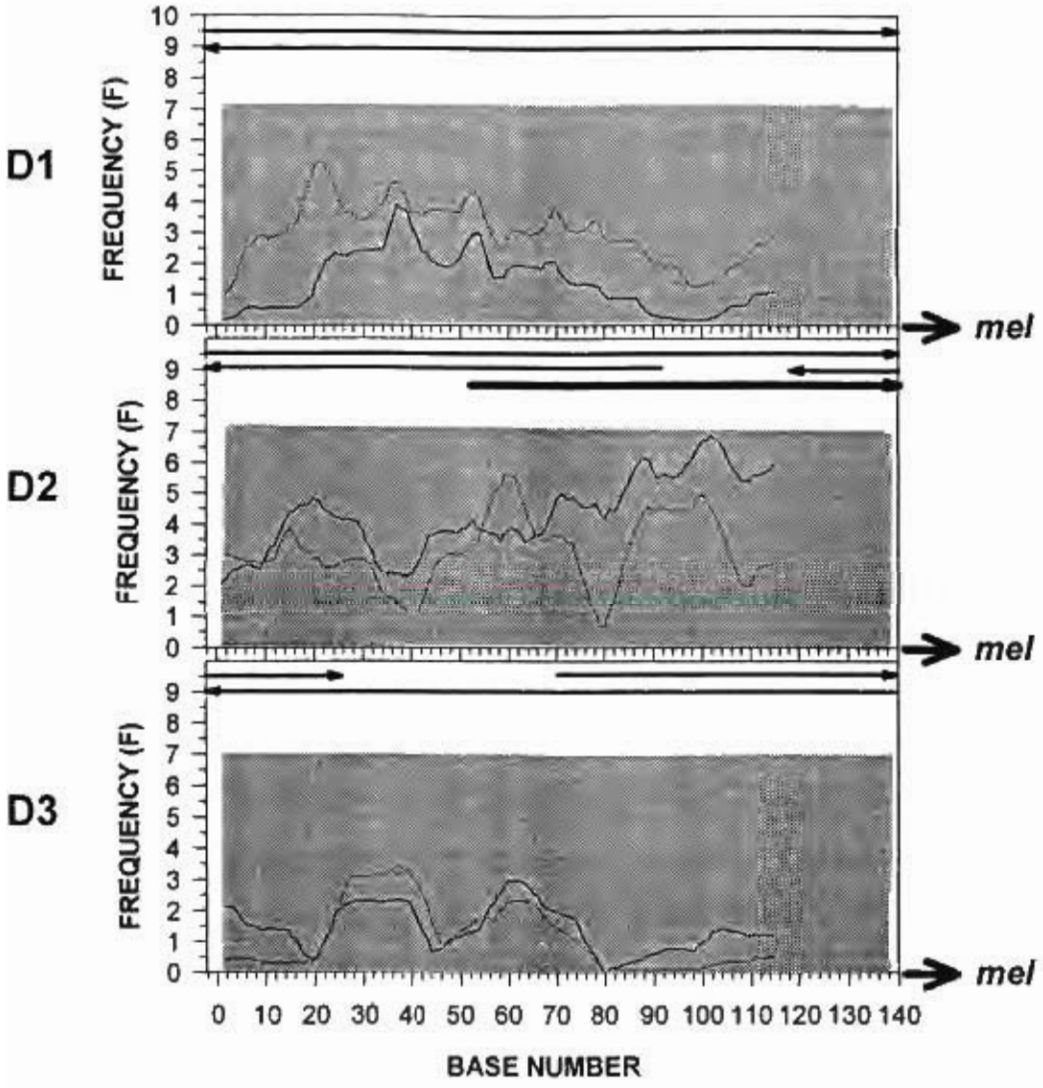

B1
B2
B3
FREQUENCY (F)
mel
BASE NUMBER
0 10 20 30 40 50 60 70 80

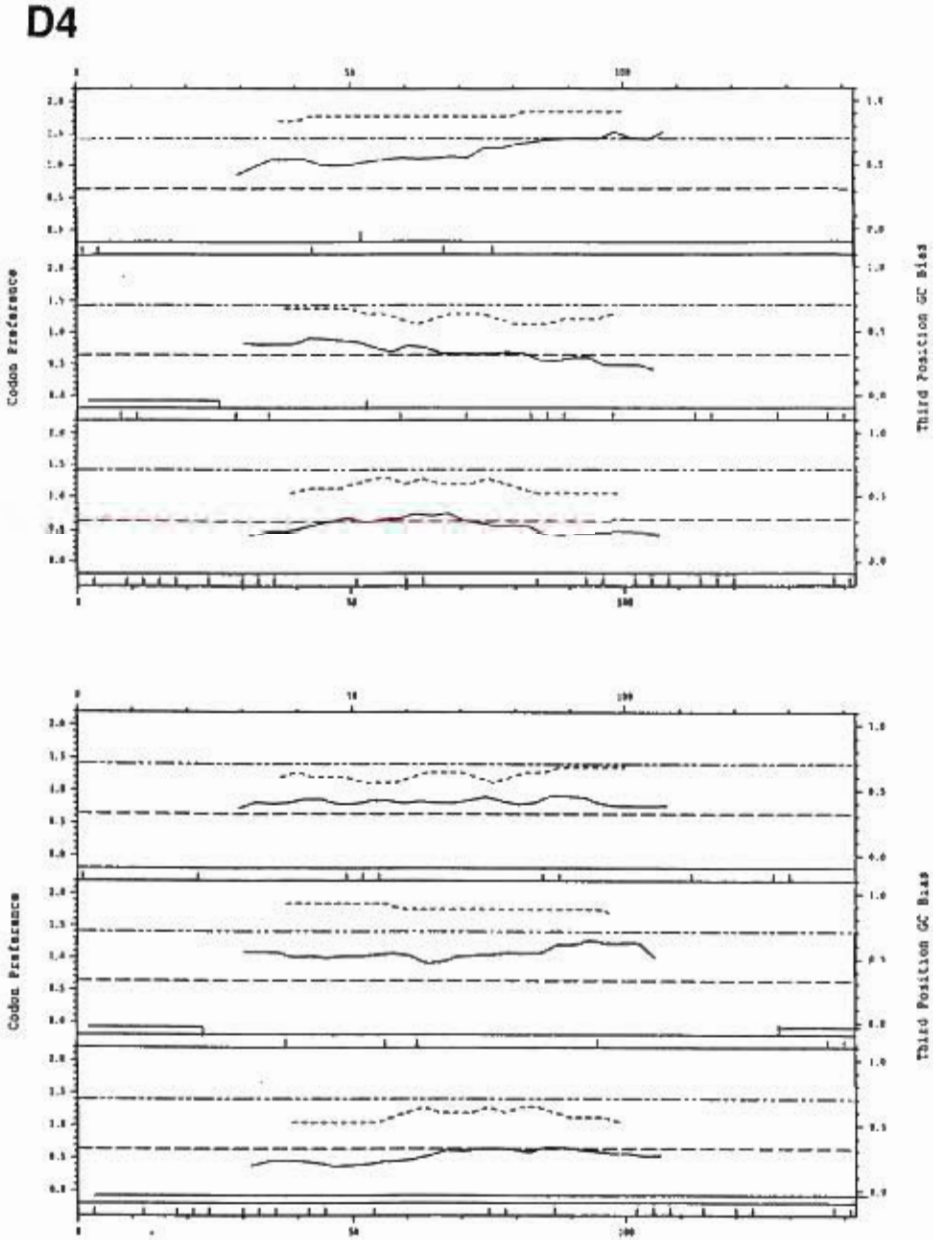

B4
Codon Preference
Third Position GC Bias

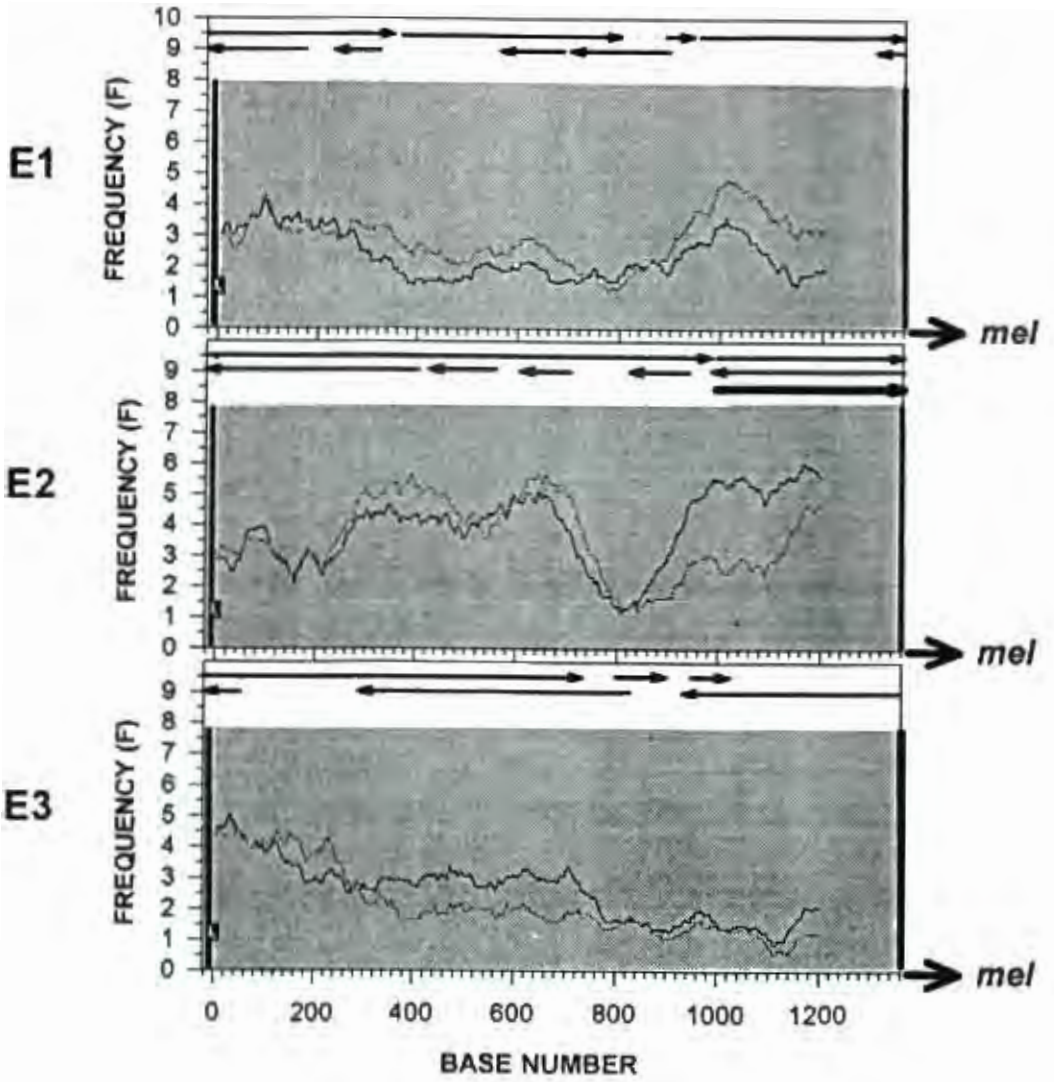

C1
C2
C3
FREQUENCY (F)
mel
BASE NUMBER
0 10 20 30 40 50 60 70 80 90 100 110

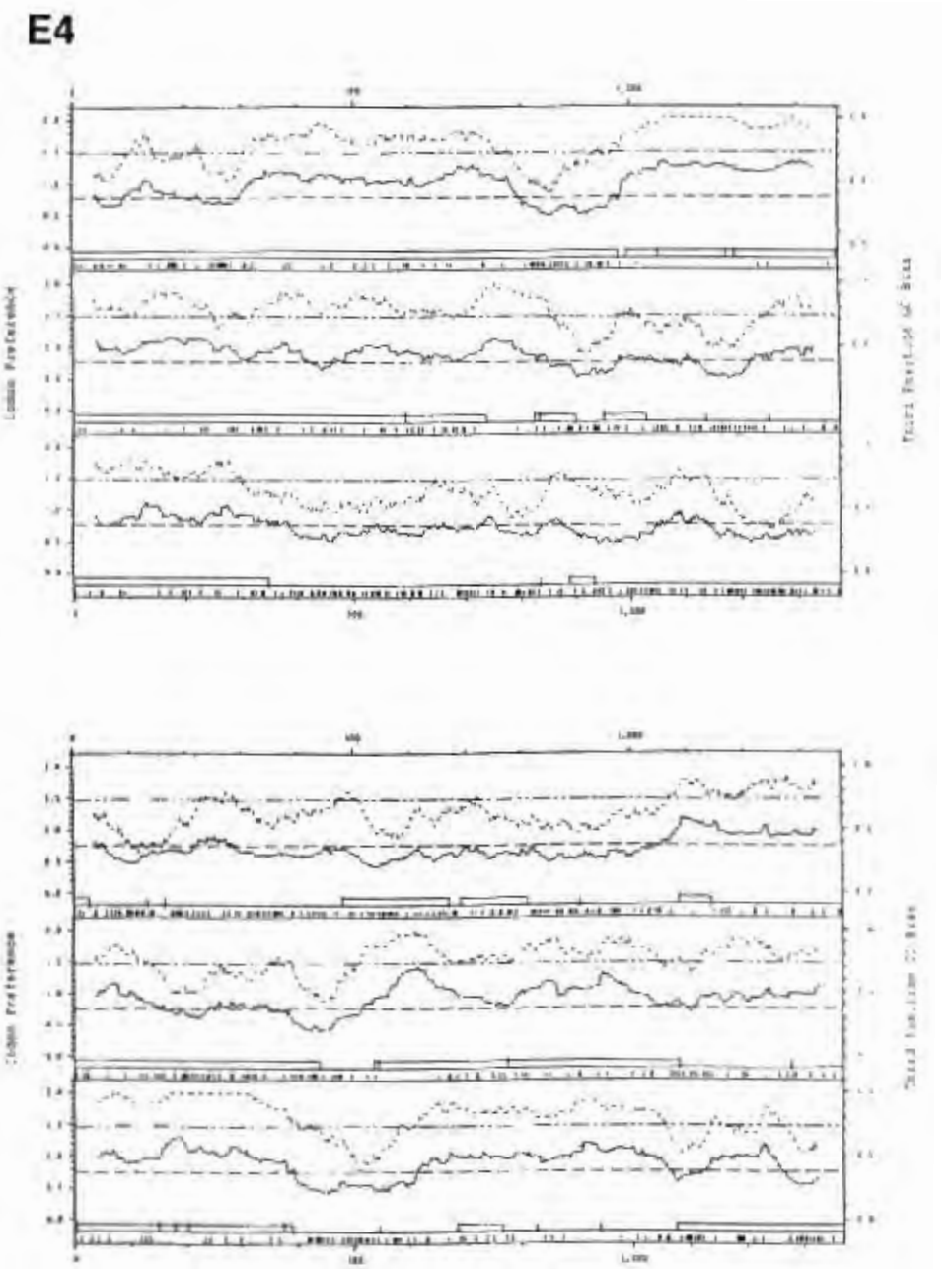
C4


Codon Preference
Third Position GC Bias


D1
D2
D3
FREQUENCY (F)
mel
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140
BASE NUMBER


D4
Codon Preference
Third Position GC Bias


E1
E2
E3
FREQUENCY (F)
mel
0
200
400
600
800
1000
1200
BASE NUMBER


E4

pADS91 and pADS911 (Fig 4.7). Nested deletions used for sequencing of the five different promoter bearing DNA fragments, and the primers used for sequencing are shown in Fig. 4.8. Sequence data is presented in Fig. 4.9.

Database searches using the FASTA TFASTA and BLAST programs (Genetics Computer Group, Wisconsin, USA; Pearson and Lipman, 1985; Altschul *et al.*, 1990) to search the GenBank, EMBL, PIR-Protein and Swiss-Prot databases were conducted. FASTA searches utilised nucleotide sequence data while TFASTA and BLAST searches utilised the backtranslated, amino acid sequences of each possible protein coding region identified (see below sections 4.3.10.1-3). Total daily updates of the databases were used. The last search was conducted on 17th of November 1995. Only in one case (that of pLP2, see section 4.3.10.2) was sequence with extensive homology to any of the promoter bearing fragments of pGLC100, pLP2, pLP3, pLP13 or pPS9 identified.

To identify potential promoter sequences, each sequence was searched for the characteristic PPCS motifs identified previously (section 3.3.6). These are shown in Fig. 4.9.

Possible protein coding regions were identified by testing each sequence using the CODONPREFERENCE (Genetics Computer Group, Wisconsin, USA) and FINDCODE (section 3.3.8) computer programs. The results are shown in Fig. 4.10.

4.3.9.1 The glucose repressed promoter bearing DNA

Inspection of Fig. 4.10 A reveals only one region that almost certainly encodes a protein. This is shown by the peak in the graph near base number 300 (Fig 4.10. A3), which indicates that the coding region starts at approximately this position and that the start codon is found at a position number that is a multiple of 3, plus 2. Such a suitable ATG start codon is found at position 320 and is upstream of a ribosome binding site-like sequence at the correct distance (Fig 4.9 A).

FINDCODE analysis of the DNA sequence between base numbers 150 and 300 indicates that this region is unlikely to encode a protein. However, the results of the FINDCODE analysis of the stretch of DNA between positions 1 and 150 are difficult to

interpret. The peak in this area seen in Fig. 4.10 A1 implies the possible existence of a protein coding region, and that the sense strand is the one used in the analysis. However, there are no stop codons suitably positioned for such a protein coding region. The alternative, that the protein coding region is found on the other strand, does not fit the observation that the peak generated by the F-type data is greater than that using R-type data (Fig. 4.10 A1, black and grey lines respectively), although a suitably positioned possible start codon and ribosome binding site exist. The possible existence of a protein coding region in the same area implied by the peak in Fig. 4.10 A3 is refuted by the lack of correctly positioned start (including GTG) or stop codons on either DNA strand.

Six promoter like sequences were identified. These consisted of two of Class H, and one each of Classes D, E, F and G.

4.3.9.2 The phosphate repressed promoter bearing DNA

Due to the limited length of the promoter bearing DNA fragments of pLP2, pLP3 and pLP13 it was difficult to determine if any region within them consisted of protein encoding sequence. FINDCODE analysis was conducted using only a 20 base averaged window and consequently the results are not completely obvious. However it is possible to discount the existence of protein coding sequence of certain frames in most cases. Specifically, in Fig. 4.10 B1, B3, C1, C2, C3, D1 and D3 the lines of the plot frequently achieve values of less than one. As shown previously (section 3.3.8) this is typical of sequence that either does not code for a protein or which is being tested in the wrong frame. It is therefore possible to state with some certainty that the promoter bearing fragment found on pLP3 is derived from a region which does not encode a protein, as each of the three FINDCODE tests failed to generate consistently high value plots (Fig. 4.10 C1-3).

In the case of the pLP2 and pLP13 DNA fragments, FINDCODE analysis indicated that it was possible that these were derived from protein encoding regions of the chromosome, as two of the plots generated (Fig. 4.10 B2 and D2) were consistently above a value of two. Such a possibility does not preclude the

existence of promoters within the region as it is known that many *Streptomyces* promoters are found in functional open reading frames (section 1.2.3.1).

Remarkably, searches for promoter-like sequences (PPCS) revealed none whatsoever.

Searches for amino acid or DNA sequence homology were conducted, although these were difficult to interpret due to the fact that the query sequences were so short. Only for the pLP2 chromosomal DNA insert was significant homology found (Fig 4.11) and, furthermore, the frame of translation was that predicted by FINDCODE analysis. The sequence was almost identical to that of a recently described σ factor (Kato *et al.*, 1995). This σ factor appears to govern the expression of the genes involved in carotenoid biosynthesis in *Streptomyces setonii*. In this organism the carotenoid biosynthesis genes are normally cryptic, and are only expressed if the σ factor is produced, which requires some unknown trigger, possibly a mutation. The region of homology lies in the 2.4 region of the protein, which is essential for promoter -35 sequence recognition, so it is difficult to believe that the putative promoter could generate RNA capable of being translated to produce a functional σ factor. It is also unexpected that a transcription start site should be found so deep within a protein coding region (assuming that the σ factor gene is complete and functional in *S.coelicolor*). Usually promoters found within protein encoding regions in streptomycetes are found at either the beginning or end of the gene (Strohl, 1992 and references therein). However cases in which promoters are found some considerable distance from either end of the protein encoding sequence in which they occur have been recorded. Examples include the *hrd*D-p1 promoter which is found within the *bar* gene, approximately 240 bp from the *bar* stop codon (Buttner *et al.*, 1990). Similarly a transcription start point is found within the *afs*R gene, approximately 2000 bp and 1000 bp from the start and end of that gene, respectively (Horinouchi *et al.*, 1990).

Figure 4.11 Homology between pLP2 insert and sigma factors

ORGANISM	SIGMA	AMINO ACID SEQUENCE AND NUMBER[a]			IDENTITY	POSITIVES	REFERENCE[b]	ACCESS #
S.coelicolor	?	?	LIKAIDRFELTREVEFTSFAVPYIVGEI	?	-	-	This work	-
			:::::::::*:+::::::::::::::::::					
S.setonii	crt	65	LIKAIDRFDLSREVEFTSFAVPYIVGEI	92	92	100	C	D17466
			:+:++:*:+:: *: :+++::: :+:::					
B.subtilis	F	70	LLKSVDKFDLTYDVRFSTYAVPMIIGEI	97	53	89	G	P07860
			:+:+:: :+:+ :+:+++::: :+:::					
B.subtilis	G	76	LMKSIDNFDLSHNVKFSTYAVPMIIGEI	103	53	85	D	P19940
			:: ::::::* : :*: *::+: +::::					
S.coelicolor	F	100	LINAIDRFDPERGVQFPTFAMPTVVGEI	127	60	78	E	L11648
			:: ::++:++ ::++: ++:+ : :					
S.coelicolor	G	86	LIDAIEKFDVDREIKFETYAITRIRG	111	42	76	F	P17211
			:+ :: :++ : +::+: :+:::					
B.subtilis	B	67	LLGAIKRYDPVVGKSFEAFAIPTIIGEI	94	42	64	B	P06574
			:+::*:*:: : *:+++: +:					
E.coli	70	412	LMKAVDKFEYRRGYKFSTYATWWI	435	41	75	A	P00579

[a]Symbols. Underlined, amino acid sequence of *S.coelicolor* M130 chromosomal DNA insert in pLP2 of DNA sequence numbers 7-90; :, identitcial; +, positive.

[b]References. **A**, Burton *et al.* (1981); **B**, Duncan *et al.* (1987); **C**, Kato *et al.* (1995); **D**, Masuda *et al.* (1988); **E**, Potukova *et al.* (1995); **F**, Tan and Chater (1993); **G**, Yudkin (1987).

4.3.9.3 The pPS9 promoter bearing DNA

FINDCODE and CODONPREFERENCE analysis of the late expressed promoter bearing DNA fragment of pPS9 highlighted some unusual features in this region. There were several sections of the sequence which could obviously represent protein coding DNA. Foremost among them was the stretch from base number (approximately) 1000 to 1369. As shown in Fig. 4.10 E2 and E1 the FINDCODE plot has high values in this area. Furthermore, a potential ATG start codon in the correct frame is found at position 994, but this is not accompanied by any sequence that resembles a ribosome binding site (Fig 4.9 E). This may indicate that the potential protein coding region is of that class that does not appear to need a ribosome binding site for translation (section 1.2.3.2).

Other regions within the DNA fragment generated high value FINDCODE plots, notably the stretch between base numbers 300 and 750 (Fig. 4.10 E2). The profile of the plot was such that it indicated that the protein coding sense strand was found on the opposite strand to that used for the test. Remarkably, this region also displayed numerous stop codons and a lack of suitably positioned open reading frames. This is difficult to explain, as such features were never found in the original tests of the FINDCODE system. Perhaps that particular stretch of DNA was originally derived from protein encoding sequence but has in the course of evolution become functionless. Thus enough of the original bias could remain to generate a protein coding sequence-like FINDCODE profile. Likewise, the lesser peaks found in the region between base numbers 1 and 200 (Fig. 4.10 E1 and E3), which also do not correlate with appropriate open reading frames might be explained.

The search of the sequence for promoter-like regions revealed 20 possible candidates. Of these there were one each of Classes A, B and D, three of Class F, four each of Classes G and H and six of Class E (Fig 4.9 E). The high number of Class E PPCS was not unexpected given the fact that this sequence is common in regions other than the -35 of promoters. Furthermore, there is scant evidence that the Class F PPCS do, in fact, encode promoters (section 3.3.6). The remaining eleven potential PPCS are all strong candidates as the promoters which drive *mel*

expression in pPS9. Of particular note is the PPCS Class A sequence found at position 979 (Fig 4.9 E). It has already been noted that the potential protein coding region with the start codon at position 994 has no ribosome binding site. In other such cases where this feature occurs for functional open reading frames the start codon is always within one base pair of the transcription start site (section 1.2.3.2; Strohl, 1992). The position of the Class A PPCS is such that it would enable exactly that relationship to occur. The possibility that this PPCS represents a functional promoter can also be argued strongly as this type of sequence is very rarely found when it does not actually represent a -10 region (section 3.3.6).

4.3.10 Closing remarks

The use of the FINDCODE system to analyse the promoter bearing fragments has demonstrated how useful this system can be. While only limited information was obtained with the shorter sequences, the potential reading frames could be determined and so database searches for homologous protein sequences were facilitated.

In the case of the longer sequences, from pGLC100 and pPS9, FINDCODE analysis was extremely useful. In both of these cases it would have been impossible to determine with any certainty which strand the protein coding sequence was likely to be on using CODONPREFERENCE alone (Fig. 4.10 A4 and E4).

The searches for possible promoter sequences were, however, disappointing. The lack of any such regions within the shorter sequences was surprising and may indicate that the rigorous selection requirements (phosphate repression and lack of expression in *bldA* mutants) resulted in the isolation of promoters that rely on as yet unidentified recognition sites.

The relatively large numbers of PPCS found within the longer sequences indicate that the system whereby promoter-like sequences are identified will have to be further refined before promoter regions can be identified with any confidence on the basis of sequence data alone.

The cloned fragments described here have been referred to as "promoters" or "promoter bearing fragments", however, a note of caution must be introduced. While the fragments certainly act as regulated promoters when cloned into the promoter probe this does not necessarily mean that they fulfil a similar or meaningful function when in their natural position in the chromosome. Firstly, it is possible that functionless promoters do exist in the chromosome (although, as argued in Chapter 3, they are probably rare). Secondly, the cloned DNA fragments might only function as promoters in the promoter probe due to the local sequence or topological characteristics of the vector itself. Such considerations are valid, and are often ignored. Recently it has been shown that certain reporter genes (notably the *lux* operon) affect expression of cloned promoters to the extent that cloned inserts that are inactive in a chromosomal location can promote transcription of reporter genes (Forsberg *et al.*, 1994). It is true that the cloned fragments in pPS9 and pGLC100 are both organised in a manner to be expected of promoter bearing fragments, in as much as they both appear to carry correctly positioned ORFS and presumptive intergenic regions. However, the fragments in pLP2, pLP3 and pLP13 are too short to derive such supporting evidence from.

4.4 Conclusion

The work presented here demonstrates that the promoter probe plasmid pWB151 is a useful addition to the tools available for use in research on *Streptomyces* genetics. It must be pointed out, however, that the use of the *mel* operon as a reporter gene does have certain limitations, primarily due to the fact that the indicator substrate can probably be utilised as a nutrient source.

The usefulness of the computer analysis systems developed previously has also been demonstrated. While the actual promoter recognition sequences and transcriptional start sites were not proven, potential promoter like sequences were determined.

The approach taken here has proved successful in isolating regions of the *S.coelicolor* chromosome potentially involved in *Streptomyces* development. However, further work is required to establish the true transcription start sites and to isolate and investigate the genes which the promoters control.

CHAPTER FIVE

CONCLUSION

5.1 Conclusion

The work described here adopted a two pronged approach to the isolation and analysis of *Streptomyces* promoters, with the focus primarily on developmentally regulated promoters.

Initially the promoter probes necessary for the isolation of regulatory regions were constructed. Secondly computer analysis was conducted in order to determine the sequence requirements for DNA to act as a promoter in streptomycetes. Finally promoters were isolated and the results of the computer analysis were used to try and identify likely regulatory sequences.

The *mel* based promoter probes constructed here will doubtless prove to be powerful tools in the further studies of *Streptomyces* genetics. They have considerable advantages over many of those promoter probes presently in use. These include, among others, the fact that promoter activity is detected chromogenically, the indicator substrate is cheap and the *mel* gene is free of TTA codons. Such characteristics make the promoter probes used here useful for large scale screening studies or in work directed at detecting differences in expression levels. Indeed, a recently constructed *mel* based reporter system has been used to measure the effect of mutations in promoter regions (Paget *et al.*, 1994). The studies on the effect of the medium on tyrosinase expression, as reported here, should also prove useful for those using the vectors constructed by Paget *et al.* (1994), who did not investigate this.

The *mel* operon does, however, have a number of weaknesses as a reporter system, and in hindsight it can be said that these make the choice of this operon to study developmental genes an unfortunate one. The most obvious weakness in the system is that the bacterial colony can probably utilise the indicator substrate as a nutrient source. This must ultimately complicate the analysis and limit the options of the researcher, as proved to be the case in this work. It is certainly true to say that the development and sporulation of a *Streptomyces* colony is intimately related to the growth conditions and nutrient levels of the medium. However, scant knowledge of this aspect of

development exists and such studies are unlikely to be simple, as the uptake of nutrients is probably regulated in a complex and integrated manner. In much of the investigation of *Streptomyces* development previously reported these considerations have been largely ignored, whereas in the work reported here a limited attempt to define growth conditions was conducted. It is crucial that in future work such studies be extended, as it is only in the context of survival in adverse conditions that *Streptomyces* sporulation will be fully understood.

The initial computer assisted approach towards determining potential regulatory sequences by testing for the avoidance of inappropriate protein-DNA binding sites was disappointing. However, it is quite possible that the original premise is sound and that further work coupled to novel methods of sequence data manipulation could yield important findings. It must be acknowledged, however, that considerable effort has been expended in this laboratory in order to achieve this objective (data not reported here) with no success.

The study of protein encoding DNA sequence bias did, however, reveal features of DNA sequence bias that are intriguing in themselves. Such features may have a bearing upon how proteins evolve and why codon bias takes the form it does. Finally, it was because of this study that it was possible to develop the novel FINDCODE method for the identification of protein encoding DNA. The method has proved extremely useful and was shown to have advantages over some systems that are currently used. A further potential use for the FINDCODE system is in the testing of DNA sequence for sequencing errors. The small search window used enables researchers to easily identify sequencing errors which are close to each other, even if they are complimentary.

The second approach to the identification of *Streptomyces* promoters using a promoter region sequence database was more successful. It was possible to tentatively identify and classify known *Streptomyces* promoters in a limited fashion. The study should prove useful to *Streptomyces* researchers in that it supplies a basis on which the promoters used for future work (for example *in vitro* transcription and mutagenesis studies)

can be selected. The classification system is certainly more extensive and less subjective than any other to date.

The cluster analysis method has potential for use not only in other organisms but also in the study of different regulatory sites. One obvious candidate is the recently recognised UP regulatory region of promoters (reviewed by Busby and Ebright, 1994) which is found between positions -65 and -45. It is also possible that the cluster analysis could be adapted to the investigation of ribosome binding sites. This is of particular interest to *Streptomyces* researchers, given the number of genes in which transcription and translation start at the same position and the consequent implication that ribosomes recognise sequences other than those upstream of the start codon. At present both of these approaches are being pursued in this laboratory.

The classification of promoters can also be utilised for the determination of potential regulatory sequences from sequence data alone. However, while this proved to be of some use in the work reported here, it was not possible to say with any certainty which of the regions within the cloned promoters were functionally significant. Future work in which more promoters (the sequences of which are reported with growing frequency) and other considerations, such as the role of the UP region, are included should enable promoters to be identified with increased accuracy.

The promoters isolated here, and the genes that they govern are also worth further investigation. One approach would be to map the actual promoter position by transcription mapping. This was the original intention when this study was initiated, however, the approach was abandoned for a variety of reasons. Firstly, such mapping studies do not, in fact, identify the promoter sequence but only a narrow area in which the sequence is likely to be found. Indeed this was the reason that the cluster analysis discussed above was undertaken. Secondly, only a limited number of promoters that displayed consistent temporal regulation and *bldA* dependence were isolated. Thus, too few were available to utilise in a cluster analysis, as originally intended. These considerations make it hard to justify the

considerable expenditure of resources that such transcription mapping studies entail.

A more fruitful line of future research is likely to be found in the cloning and analysis of the genes that these promoters regulate. Sequence data might ultimately enable the function of these genes to be identified by homology studies. Alternatively, directed mutation of these genes may yield novel developmental mutants. This type of information will doubtless prove to be of value to workers dissecting the complex process of *Streptomyces* development.

APPENDIX A ABREVIATIONS

A	adenine
aa	amino acid
ac.acid	acetic acid
A-factor	2-(6'-methylheptanoyl)-3R-hydroxymethyl-4-butanolide
amp	ampicillin
ATP	adenosine triphosphate
bp	base pair(s)
BSA	bovine serum albumin fraction V
c	cytosine
cfu	colony forming unit
CIAP	calf intestinal alkaline phoshatase
conc.	concentration
dATP	deoxyadenosine 5'triphosphate
dCTP	deoxycytidine 5'triphosphate
dGTP	deoxyguanosine 5'triphosphate
dTTP	deoxythreonine 5'triphosphate
DTE	dithioerythritol
DTT	dithiothreitol
DNA	deoxyribonucleic acid
DNase	deoxyribonuclease
DMSO	dimethylsulphoxide
EDTA	ethyline diamine tetra-acetic acid
EtBr	ethidium bromide
g	standard gravitational acceleration
G	guanine
h	hours
IPTG	isopropyl-thio-β-galactoside
kb	kilobase
LB	Luria-Bernardi
min	minutes
OD	Absorbance
ORF	open reading frame
PvP	polyvinyl-polypyrolidone
r	(superscript) resistance
RNase	ribonuclease
s	seconds
Sarkosyl	N-laurylsarcosine
SDS	sodium dodecyl sulphate
Str	streptomycin
T	thymine
thio	thiostrepton
Tris	tris(hydroxymethyl)aminoethane
v/v	volume/volume ratio
w/v	weight/volume ratio
x-gal	5-bromo-4-chloro-3-indolyl-β-galactoside

APPENDIX B BACTERIAL STRAINS AND PLASMIDS

Table B1 *Streptomyces* strains

Streptomyces strains.		
Strain.	Genotype.	Reference.
S.coelicolor M130	*hisA1 strA1 uraA1*	Hopwood *et al.*, 1985
S.coelicolor J1501	*hisA1 pgl strA1 uraA1*	Hopwood *et al.*, 1985
S.coelicolor J1700	*bldA39 hisA1 pgl strA1 uraA1*	Leskiw *et al.*, 1991
S.coelicolor J1820	*hisA1 pgl strA1 uraA1 whiG71*	Mendez and Chater, 1987
S.coelicolor J669	*agaA1 bldB43 cysD18 mthB2*	Merrick, 1976
S.coelicolor J660	*agaA1 bldC18 cysD18 mthB2*	Merrick, 1976
S.coelicolor 1169	*bldD53 hisA1 mthB2 pheA1 strA1*	Merrick, 1976
S.lividans TK24	*str-6*	Hopwood *et al.*, 1985
S.lividans TK21	(wild type)	Hopwood *et al.*, 1985

Table B2 *E.coli* strains

Escherichia coli strains.		
Strain.	Genotype.	Reference.
E.coli JM105	*supE4 endA sbcB15 hsdR4 rpsL thi Δ (lac-proAB)*	Yanish-Perron *et al.*, 1985
E.coli LK111	F^- *thi1 thr1 leuB6 lacI*$^-$ *tonA21 supE44* r_K^- m_K^-	Zabeau and Stanley, 1982
E.coli GM41	*dam*$^-$ *thi1 rel1* HfrH	D. Rawlings (pers. comm.)[a]

[a]From the strain collection of Doug Rawlings, University of Cape Town, South Africa. Originally constructed by M.G. Marinus, Rutgers Medical School, New Jersey.

Table B3 Plasmids and antibiotic resistance levels

Plasmids[a]	Antibiotic marker		Reference
	Type and working concentration[b]		
pIJ702	Thiostrepton	(l) 5μgml^{-1} (s) 50μgml^{-1}	Katz *et al.* (1983)
pIJ4642	Streptomycin	(l/s) 30μgml^{-1}	M.J. Bibb, unpublished
pUC18	Ampicillin	(l/s) 100μgml^{-1}	Yanisch-Perron *et al.* (1985)
pUC19	Ampicillin	(l/s) 100μgml^{-1}	Yanisch-Perron *et al.* (1985)
Bluescript KS	Ampicillin	(l/s) 100μgml^{-1}	Manual, Stratagene, San Diego, USA
pBlue	Streptomycin	(l/s) 30μgml^{-1}	Smith (1991)
pFull	Streptomycin	(l/s) 30μgml^{-1}	Smith (1991)
pMTL22	Ampicillin	(l/s) 100μgml^{-1}	Chambers *et al.* (1988)
OTHERS[c]			This work

[a]*Streptomyces* plasmids were maintained in *S.lividans* TK21, *E.coli* plasmids were maintained in *E.coli* LK111.

[b]Concentrations are given for liquid (l), solid (s) and liquid and solid (l/s) media.

[c]All plasmids constructed utilised the origins of replication and resistance markers of one of the previously described plasmids. Antibiotic selection conditions were therefore the same. The exceptions to this were plasmids that carried the pFULL origin of replication and the *tsr* thiostrepton resistance gene, and in these cases selection was as for pIJ702.

APPENDIX C GENERAL METHODS

Index of methods

Materials used are described in Appendix D.

C.1 Culture conditions

C.1.1 Culture conditions for *E.coli*

All *E.coli* strains were maintained in the long term on *E.coli* Minimal Medium containing suitable nutrient additives (Table A.1). Working stocks were grown in LB broth and stored frozen at -70°C in 15% glycerol. All plasmid carrying strains were grown in the presence of the appropriate antibiotic (Table A.2). LB broth and LB agar were used for all other culture purposes. Any additives are specified within the text. In every case growth was at 37°C.

C.1.2 Culture conditions for *Streptomyces*

All *Streptomyces* strains were maintained in the long term on *Streptomyces* Minimal Medium containing suitable nutrient additives for growth (see Table B.1). In the case of spore-producing strains, stocks of spores were scraped in 15% glycerol from M3/6.5 or M3/7.0 plates, filtered through cotton wool and stored at -70°C. Colony forming unit counts were conducted by dilution and plating on M3, M3/6.5 or M3/7.0 medium. For asporogenous strains, and some other cases (specified in text) stocks were obtained from mycelia that were grown in YEME broth, centrifuged in an Eppendorf microfuge, resuspended in 15% glycerol and stored at -70°C. In some cases (specified in text) spore and mycelium stocks used for inoculation were diluted with water, centrifuged, washed, recentrifuged and resuspended in distilled water. For spores, cfu counts were calculated assuming no loss occurred during the process. Mycelial samples were resuspended such that a packed volume of approximately 50 µl of mycelia was made up to 1 ml. All plasmid carrying strains were grown in the presence of the appropriate antibiotic (see Table B.3). For the purposes of DNA extraction and protoplast production mycelia were grown in YEME broth. Other media are specified within the text. Large cultures (100 ml or greater) were grown in flasks with vigorous

reciprocal shaking and were rotated daily to remove wall growth. Small cultures (10 ml) were grown in standard containers containing glass beads with reciprocal shaking and were briefly vortexed daily to remove wall growth and disperse mycelial clumps. In all cases growth was at 30°C.

C.2 Transformation procedures

C.2.1 Production of competent *E.coli* cells

This method is adapted from that described by Maniatis *et al.* (1982). A 100 ml prewarmed and preshaken LB broth was inoculated with 1 ml of an overnight culture of cells. The culture was grown at 37°C with vigorous shaking for 1 h. The culture was immediately chilled for 10 min in an ice/water bath. A 70 ml sample was centrifuged at 4000 rpm in a Sorval SS34 rotor at 4°C for 5 min. The pellet was resuspended in 20 ml ice cold 0.1 M $MgCl_2$ and centrifuged as before. The pellet was then resuspended in 20 ml of ice cold 0.1 M $CaCl_2$ and the sample was stored on ice for 20 min following which it was centrifuged as before. The cells were then suspended in 2-4 ml of ice cold 0.1 M $CaCl_2$ and stored on ice for 30 min before use.

C.2.2 Transformation of *E.coli*

This method is adapted from that described by Maniatis *et al.* (1982). DNA (up to 1 µg in up to 10 µl TE buffer) was added to 100 µl of competent cells (C.2.1) and the sample stored on ice for 30 min. The sample was then incubated at 37°C for 5 min, and 500 µl of LB Broth was added. Cells were plated after incubation at 37°C for 45 min. Plating was on LB agar or XGal plates with antibiotic selection (specified in text).

C.2.3 Production of *Streptomyces* protoplasts

This method is adapted from that of Hopwood *et al.*(1985). An equal volume of water was added to a 100 ml culture and the mycelia were pelleted by centrifugation at 15000-30000 x g for

10 min. The pellet was resuspended (after removal of all possible supernatant) in 10 ml P buffer containing 10 mg lysozyme powder (filter sterilised). The sample was incubated at 30°C with frequent mixing until the mycelial clumps were fully dispersed. The sample was centrifuged in a Hettich EBA 3S benchtop centrifuge at 40% of the maximum speed (the lowest speed at which a pellet will form) for 7 min. The pellet was resuspended in 1 ml P buffer and stored frozen at -70°C.

C.2.4 Transformation of *Streptomyces*

DNA (up to 1 µg in 10 µl TE buffer) was added to 50 µl of protoplasts (C.2.2) followed by 200 µl of T buffer. The sample was mixed by pipetting and immediately spread on two R2YE plates with the tip of a 1 ml pipette. R2YE medium contained Carbendazime (50 $\mu g ml^{-1}$) and ampicillin (100 $\mu g ml^{-1}$) as recommended by Melton and Kieser (1988). The plates were incubated until bacterial growth was just visible. Selection procedures were adapted from the methods of Landers (1987) and Portmore *et al.* (1987). For selection on R2YE, plates were flooded with 1-2 ml of an aqueous suspension of 200 $\mu g ml^{-1}$ thiostrepton before further incubation. For selection on other media plates were flooded with 1-2 ml of an aqueous suspension of 200 $\mu g ml^{-1}$ thiostrepton and a dry disc of Whatmans water hardened 540 filter paper was placed over the agar. Paper discs were previously sterilised by wetting and autoclaving. After 1 day at 30°C discs were removed and placed on the new medium which contained 50 $\mu g ml^{-1}$ thiostrepton. Discs were removed after 1 day of incubation. All *Streptomyces* transformation was by this procedure unless otherwise stated.

Small scale transformation was as described above, except that 2 µl of DNA was added to 10 µl of protoplasts followed by 40 µl of T buffer. The sample was spotted on R2YE medium with localised spreading.

C.3 DNA isolation

C.3.1 Small scale isolation of *E.coli* plasmid DNA

This method is adapted from that described by Maniatis *et al.* (1982). 400 µl of an overnight broth culture was centrifuged in an Eppendorf microfuge and the pellet was resuspended in 200 µl of solution 1. After 5 min, 400 µl of solution 2 was added and the sample was placed on ice for 10 min. 300 µl of solution 3 was then added and the sample was kept on ice for a further 10 min following which it was centrifuged at 4°C for 5 min in an Eppendorf microfuge. 600 µl of the supernatant was removed, and subjected to isopropanol precipitation (C.3.9). The pellet was resuspended in 50 µl of TE buffer and stored at 4°C.

C.3.2 Large scale isolation of *E.coli* plasmid DNA

This method is adapted from that described by Maniatis *et al.* (1982). 100 ml of an overnight broth culture was centrifuged at 15000-30000 x g and the pellet was resuspended in 5 ml of solution 1. After 5 min, 10 ml of solution 2 was added and the sample was placed on ice for 10 min. 7.5 ml of solution 3 was then added and the sample was kept on ice for 10 min following which it was centrifuged at 4°C at 15000-30000 x g. The supernatant was removed and subjected to isopropanol precipitation (C.3.12) and the pellet was resuspended in 4.5 ml of TE buffer. If the sample was to be used for sequencing purposes, 50 µl of RNaseA (4 $\mu g ml^{-1}$) was added. After 30 min the sample was subjected to one phenol/chloroform extraction (C.3.8), one chloroform wash (C.3.9), isopropanol precipitation (C.3.11) and the pellet was resuspended in 4.5 ml of TE buffer. If the DNA was not to be used for sequencing the RNaseA and phenol treatments were omitted. The sample was then subjected to isopycnic CsCl-EtBr density gradient ultracentrifugation (C.3.12).

C.3.3 Small scale isolation of *Streptomyces* plasmid DNA

This method is an adaptation of the neutral lysis procedure described by Hopwood *et al.* (1985). The DNA obtained is

suitable only for use in transformation. 1 ml of a 5-10 day old culture of *Streptomyces* was centrifuged in an Eppendorf microfuge and the pellet was resuspended in 400 µl of TE buffer containing 34% sucrose. To this was added 80 µl of 0.25 M EDTA and 1-2 mg of lysozyme powder. The sample was incubated at 30°C for 30 min, following which 35 µl of 10% SDS and 70 µl of 5 M NaCl was added. After being placed on ice for 2 h the sample was centrifuged in an Eppendorf microfuge for 5 min at 4°C. The supernatant was retained and subjected to isopropanol precipitation (C.3.11). The pellet was resuspended in 50 µl of TE buffer and stored at 4°C.

C.3.4 Large scale isolation of *Streptomyces* plasmid DNA (method 1)

This method is an adaptation of the neutral lysis procedure described by Hopwood *et al.* (1985) and is best suited for the isolation of high copy number plasmids. An equal volume of water was added to 100 ml of a 5-10 day old broth culture of *Streptomyces* and the sample was centrifuged at 15000-30000 x g for 10 min. The pellet was resuspended in 40 ml of TE buffer containing 34% sucrose. To this was added 8 ml of 0.25 M EDTA and 10-20 mg of lysozyme powder. The sample was incubated at 30°C for 30 min, following which 35 ml of 10% SDS and 70 ml of 5 M NaCl was added. After being placed on ice for 2-12 h the sample was centrifuged at 15000-30000 x g at 4°C for 10 min. The supernatant was retained and subjected to isopropanol precipitation (C.3.11). The pellet was resuspended in 4.5 ml of TE buffer and the sample was then subjected to isopycnic CsCl-EtBr density gradient ultracentrifugation (C.3.12).

C.3.5 Large scale isolation of *Streptomyces* plasmid DNA (method 2)

This method is identical to the method used to isolate *Streptomyces* chromosomal DNA (C.3.6) except that the plasmid band was isolated after CsCl-EtBr density gradient ultracentrifugation. This method is suitable for use with large and/or low copy number plasmids.

C.3.6 Isolation of *Streptomyces* chromosomal DNA.

This method is an adaptation of the isolation of *Streptomyces* "total" DNA (procedure 2) described by Hopwood *et al.* (1985). An equal volume of water was added to 100 ml of a 5-10 day old broth culture of *Streptomyces* and the sample was centrifuged at 15000-30000 x g for 10 min The pellet was resuspended in 3 ml of TE buffer. To this was added 1.3 ml of 0.25 M EDTA and 10-20 mg of lysozyme powder. The sample was incubated at 30°C for 1 h and occasionally mixed by pippetting. After this, 200 µl of 10% sarkosyl was added and the sample was shaken gently. The sample was then subjected to isopycnic CsCl-EtBr density gradient ultracentrifugation (C.3.12).

C.3.7 Isolation of *Streptomyces* "total" DNA

An equal volume of water was added to 100 ml of a 5-10 day old broth culture of *Streptomyces* and the sample was centrifuged at 15000-30000 x g for 10 min. The pellet was resuspended in 5 ml TE buffer to which was added 50 mg lysozyme powder. The sample was incubated at 30°C for 1 hr following which 1.2 ml 0.5 M EDTA, 0.13 ml 10 $mgml^{-1}$ pronase and 0.7 ml 10% SDS was added. After incubation at 37°C for 2 hr 6 ml phenol/chloroform was added and the sample shaken gently for 10 min. Another 6 ml phenol/chloroform was added and shaking resumed for 10 min. The sample was centrifuged at 15000-30000 x g for 15 min, the aqueous layer was removed. Phenol/chloroform extraction (C.3.8) was performed at least twice (or until there was no protein at the interface) followed by two chloroform washes (C.3.9). RNaseA (10 $mgml^{-1}$) was added to a final concentration of 40 μgml^{-1}, the sample was incubated for 1 hr at 37°C after which phenol/cloroform extraction was performed twice and the sample washed twice with chloroform as before. DNA was isolated by isopropanol precipitation (C.3.11) and the DNA pellet was resuspended in 2 ml TE buffer.

C.3.8 Phenol/chloroform extraction of protein from DNA samples

Samples of DNA in solution were mixed with an equal volume of phenol/chloroform, shaken for 5 min, centrifuged for 10 min at 15000-30000 x g, and the aqueous phase retained.

C.3.9 Chloroform washing of phenol extracted samples

DNA solutions were mixed by shaking for 1 min with an equal volume of chloroform, centrifuged for 1 min at 15000-30000 x g, and the aqueous phase retained.

C.3.10 Ethanol precipitation of DNA

DNA solutions were mixed with 3 M sodium acetate and ethanol in the ratio 10:1:22 and centrifuged for 10 min at 15000-30000 x g. The supernatant was removed and to the pellet was added 70% ethanol, of a volume equal to the original sample. The sample was centrifuged as before and the supernatant removed. The pellet was washed with 100% ethanol and air dried.

C.3.11 Isopropanol precipitation of DNA.

As for ethanol precipitation of DNA, except that solutions of DNA were originally mixed with 3 M sodium acetate and isopropanol in the ratio 10:1:11.

C.3.12 Isopycnic CsCl-EtBr density gradient centrifugation.

The DNA sample was dissolved in 4.5ml of buffer. To this was added 5.25 g of CsCl and 250 µl of EtBr ($10 mg ml^{-1}$). The refractive index was then adjusted to 3.962. The sample was loaded into a Beckman Quick-seal centrifuge tube and centrifuged at 55000 rpm for 10-16 h or at 60000 rpm for 6 h at 18°C in a Beckman VTi65 rotor. The plasmid and/or chromosomal DNA band was visualised under UV light (365 nm) and removed with a syringe and needle. The sample was shaken with NaCl-saturated isopropanol and the organic phase was removed after the phases had separated. This was repeated until all the

EtBr was removed from the aqueous phase. To the aqueous phase was added two volumes of TE buffer. The sample was subjected to isopropanol precipitation (C.3.11). The pellet was immediately resuspended (without air drying) in 500 µl of TE buffer and again subjected to isopropanol precipitation. The pellet was resuspended in 100 µl-2 ml of TE buffer.

C.3.13. Quantification of DNA solutions

The absorbance (OD_{210}-OD_{320}) of the DNA sample was recorded on a Beckman DU-64 spectrophotometer and the DNA concentration calculated according to Maniatis *et al.* (1982).

C.4 DNA manipulation

C.4.1 Restriction endonuclease digests

All restriction endonuclease digests were carried out using the Boehringer Mannheim A, B, L, M, and H restriction buffers as recommended by this company. Unless otherwise stated, all complete digestions used 4 units of enzyme for each 1 µg of DNA. Other variables were as recommended by the manufacturer of the enzyme. For partial digestions 2-10 µg of DNA was digested for 1 h with a serial dilution of the restriction enzyme concentrations, following which the reaction was stopped by the addition of 1/10 volume of 0.25 M EDTA. The degree of digestion was estimated by gel electrophoresis (C.4.2). For DNA to be used for cloning purposes restriction enzymes were inactivated by heat treatment where possible (Maniatis *et al.*, 1982) or by the addition of 1 µl of tRNA (1 $mgml^{-1}$) followed by phenol treatment, chloroform washing and ethanol precipitation (C.3.8, C.3.9 and C.3.10). For DNA to be treated with CIAP the tRNA was omitted.

C.4.2 Agarose gel electrophoresis of DNA.

This was performed as described by Maniatis *et al.* (1982).

TBE buffer was used and agarose gels were 0.7% or 1.4%. DNA bands were visualised under UV light (202 nm, 365 nm if DNA was

to be isolated from the gel). Restriction site mapping was as described in standard texts, using *PstI* digested lamda DNA as the size marker.

C.4.3 Electroelution of DNA from agarose gels

After agarose gel electrophoresis (C.4.2) DNA bands were excised from the gel and sealed in dialysis tubing with TBE buffer. The sample was subjected to electrophoresis until the DNA had left the gel slice. The TBE buffer was removed, 1 µl of tRNA (1 $mgml^{-1}$) was added and the sample was subjected to two phenol/chloroform extractions, two chloroform washes and isopropanol precipitation (C.3.8, C.3.9 and C.3.11). The pellet was resuspended in 20 µl of TE buffer and stored at -20°C.

C.4.4 Conversion of 5'overhanging DNA ends to blunt ends using the Klenow fragment

This was as described by Maniatis *et al.* (1982). This method was used if the DNA sample had 5'overhanging ends.

C.4.5 Conversion of 3'and 5'overhanging DNA ends to blunt ends using T4 DNA polymerase

This was as described by Maniatis *et al.* (1982). This method was used if the DNA sample had either 3'overhanging ends or 5' and 3'overhanging ends.

C.4.6 Partial fill-in of 5'overhanging DNA ends using the Klenow fragment

10 µg of insert DNA was digested with the appropriate restriction endonuclease (C.4.1). Seven 10 µg samples of vector DNA were digested with the appropriate restriction endonuclease. Each digest was at doubling dilutions of endonuclease (4.0-0.062 $U\mu g^{-1}$). Digestion was for 1 hr and reactions were stopped by adding 1/10 volume of 0.25 M EDTA. 1 µg samples of DNA from each digest were subjected to agarose gel electrophoresis (C.4.2). The digest that was completely cut with the minimum of restriction enzyme was identified. The

remainder of this sample was then subjected to two phenol/chloroform extractions, two chloroform washes and isopropanol precipitation (C.3.8, C.3.9 and C.3.11). The pellet was resuspended in 20 µl of TE buffer. The recessed ends of the vector and insert DNA were then each filled in using Klenow fragment (C.4.4) except that a limited selection of the four possible deoxynucleotides (dA, dC, dG and dT) were added to each reaction. The choice of deoxynucleotides used was such that the overhanging ends generated in each case were compatible for ligation between the vector and insert DNA. The samples were subjected to phenol/chloroform treatment, chloroform washing and ethanol precipitation (C.3.8, C.3.9 and C.3.10). The fill in reaction and subsequent steps were repeated and the DNA was resuspended in 50 µl TE buffer and stored at -20°C.

C.4.7 DNA ligations

Ligations were carried out as described previously (Hopwood *et al.*, 1985), with concentrations of DNA of 1 pM for self ligation recircularisation reactions and 5 pM vector and 5-25 pM insert for insertion reactions. All ligations were in a final volume of 20 µl. Where molar concentrations of DNA could not be determined, a range of dilutions and ratios of vector and insert were used. For ligations involving partially filled in DNA (C.4.6) the vector and insert DNA were mixed at a 1:1 mass ratio and doubling dilutions of this DNA were ligated. A doubling dilution ligation of vector DNA alone was also performed. Transformation of *E.coli* (C.2.2) using the ligation mixes revealed which dilutions gave the greatest number of transformants and also an insertion rate of above 95%.

C.4.8 Nick translation of DNA

For nick translation an Amersham N.5000 kit was used according to the manufacturers instructions. Unincorporated nucleotides were removed using a Sephadex G-50 spin column (Maniatis *et al.*, 1982).

C.4.9 Southern blotting

This was essentially as described by Maniatis *et al.* (1982) for capillary transfer except that the agarose gel was first soaked in 0.25 M HCl and washed with water before being used. Transfer was by 0.4 M NaOH to Amersham Hybond N+ membrane, so baking was omitted. The filter paper was washed with 2 x SSC and washed again with Prehybridisation buffer at 68°C for 2 h. The DNA probe, labeled with P^{32} by nick translation (C.4.8) was denatured by boiling for 5 min followed by chilling. The probe was hybridised with the filter in Hybridisation buffer and incubated at 68°C for 12-24 h. The filter was washed twice with 2 x SSPE, 0.5% SDS at room temperature for 15 min. It was then washed twice in 0.1 x SSPE, 0.5% SDS for one hour, once at 37°C and once at 68°C. The Bands were visualised by autoradiography using AGFA Curex X-ray film with enhancer screens unless otherwise specified.

C.4.10 Sequencing

Sequencing was conducted with a Sequenase 2.0 kit (United States Biochemical, Cleveland, Ohio) and was essentially according to the instructions of the manufacturers. Modifications are described by Schuurman and Keulen (1991) and included the use of DMSO in the reaction mixtures, which were conducted at 48°C, and the use of a chase reaction utilising the Klenow fragment.

C.4.11 CIAP treatment

CIAP (calf intestinal alkaline phospatase), molecular biology grade, was supplied by Boehringer Mannheim and treatment was as recommended by the manufacturers.

C.4.12 Nested deletion of DNA (method 1)

Plasmid DNA was isolated (C.3.2), 10 µg of this was partially digested with restriction endonuclease (specified in text) and the linear plasmid was isolated by gel purification (C.4.1 and C.4.3). Following this the DNA was digested with a second

restriction endonuclease and the overhanging ends converted to blunt ends (C.4.1, C.4.4 and C.4.5). The plasmids were recircularised by ligation and the ligation mix was used to transform *E.coli* LKIII, with ampicillin selection (C.4.7 and C.2.2). Plasmids with appropriate deletions were identified by isolation and restriction mapping (C.3.1).

C.4.13 Nested deletion of DNA (method 2)

Plasmid DNA was isolated (C.3.2), and 15 µg of this was digested with restriction endonuclease (specified in text; C.4.1). The DNA sample was cleaned by two phenol/chloroform extractions, two chloroform washes, precipitation with ethanol and resuspension in 30 µl TE buffer (C.3.8, C.3.9 and C.3.10). The DNA was then made up to 200 µl in Bal31 buffer and preheated to 37°C. Six units of Bal31 exonuclease was added and 25 µl aliquots were removed at intervals over a time course of 30 min. To each aliquot was added 275 µl of TE buffer and the sample was immediately subjected to two phenol/chloroform extractions, two chloroform washes, precipitation with ethanol and resuspension in 10 µl TE buffer (C.3.8, C.3.9 and C.3.10). Each DNA sample was then digested with restriction endonuclease (specified in text) following which a quarter of the sample was subjected to agarose gel electrophoresis to determine the extent of the deletions (C.4.2). The remaining DNA was subjected to two phenol/chloroform extractions, two chloroform washes, precipitation with ethanol and resuspension in 10 µl TE buffer (C.3.8, C.3.9 and C.3.10). Overhanging DNA ends were then converted to blunt ends using T4 polymerase (C.4.5). Each DNA sample was then ligated under conditions designed to promote recircularisation (C.4.7) and the ligation mix used to transform *E.coli* LKIII with ampicillin selection. (C.2.2). Plasmids with useful deletions were identified by isolation and restriction mapping (C.3.1).

APPENDIX D MATERIALS

D1 Media

Growth media are shown in Table D1

Standard media are described by Hopwood *et al.* (1985a)

D2 Buffers

Agarose gel loading buffer. EDTA 60 mM, Ficoll (400) 25%, Bromophenol Blue 0.25%.

Bal31 buffer. See Table D2.

CIAP buffer. See Table D2.

Enzyme buffers. See Table D2.

Hybridisation buffer. 6 x SSPE, 0,5% SDS, 100 $\mu g ml^{-1}$ sheared denatured herring testis DNA.

Klenow buffer. See Table D2.

P buffer. Sucrose 10.3%, K_2SO_4 0.025%, $MgCl_2.6H_2O$ 0.202%, Trace Solution 0.2%. After autoclaving: KH_2PO_4 0.005%, $CaCl_2.2H_2O$ 0.368%, Tes buffer (pH 7.2) 0.573%.

Prehybridisation buffer. 6 x SSPE, 6 x Denhardts Solution, 0,5% SDS, 100 $\mu g ml^{-1}$ sheared denatured herring testis DNA.

Phosphate buffer. 0.05 M. Stock solutions of 0.5 M Na_2HPO_4 and 1.0 M NaH_2PO_4 were mixed and made up to 1 l. Mixtures: pH8.0, 93.2 ml and 3.4 ml; pH7.5, 84.0 ml and 8.0 ml; pH7.0, 57.7 ml and 21.1 ml; pH6.5, 31.5 ml and 34.3 ml; pH6.0, 12.0 ml and 44.0 ml.

Restriction endonuclease buffers. See Table D2.

T4 ligase buffer. See Table D2.

T4 polymerse buffer. See Table D2.

T buffer. Part A. Sucrose 2.6%, Trace solution 0.2%, K_2SO_4 0.025%, $CaCl_2$ 0.1 M, Tris-maleic acid buffer 5%. 1 part PEG 1000 was added to 3 parts Part A.

TE buffer. Tris-Cl 10 mM (ph 7.5), EDTA 1 mM.

TBE buffer. Tris-borate 89 mM, boric acid 89 mM, EDTA 2 mM

Tris-maleic acid buffer. Tris 1 M, adjusted to pH 8 with maleic acid.

Tyrosine solution. 0.75%, 1 M NaOH was added dropwise to a double strength suspension until tyrosine was disolved. The solution was made up to the final concentration.

D3 Solutions

Chloroform. 24:1 chloroform:isoamyl alcohol.

Denhardts solution (x20). BSA 2%, Ficoll 2%, PvP 2%.

Phenol/chloroform. 24:1:25 chloroform:isoamyl alcohol:phenol. Equilibrated with 0.1 M Tris-Cl (pH 8) and containing 0.1% hydroxyquinoline.

Solution 1. Glucose 50 mM, Tris-Cl (pH 8) 25 mM, EDTA 10 mM.

Solution 2. NaOH 0.2 M, SDS 1%.

Solution 3. Potassium acetate 3 M, glacial acetic acid 11.5%.

SSC (x20). NaCl 3 M, Sodium citrate 0.3 M.

SSPE (x20). NaCl 3.6 M, NaH_2PO_4 0.2 M, EDTA 0.02 M.

Tiger Milk. L-arginine 1 gl^{-1}, L-cystine 0.75 gl^{-1}, L-histidine 0.75 gl^{-1}, DL-homoserine 0.75 gl^{-1}, L-leucine 0.75 gl^{-1}, L-phenylalaninine 0.75 gl^{-1}, L-proline 0.75 gl^{-1}, adenine 0.15 gl^{-1}, uracil 0.15 gl^{-1}, nicotinamide 0.01 gl^{-1}.

Trace solution. $ZnCl_2$ 40 mgl^{-1}, $FeCl_3.6H_2O$ 200 mgl^{-1}, $CuCl_2.2H_2O$ 10 mgl^{-1}, $MnCl_2.2H_2O$ 10 mgl^{-1}, $Na_2B_4O_7.10H_2O$ 10 mgl^{-1}, $(NH_4)_6Mo_7O_{24}.4H_2O$ 10 mgl^{-1}.

D4 Enzymes

Pronase. Pronase 20 $mgml^{-1}$ in H_2O. Preincubated at 37°C for 2 h.

RNase A. RNase A 10 $mgml^{-1}$ in Tris-Cl (pH 7.5) 10 mM, NaCl 15 mM. Boiled for 15 min before use.

D5 Antibiotics

Ampicillin. 100 $mgml^{-1}$, filter sterile.

Streptomycin. 50 $mgml^{-1}$, filter sterile.

Thiostrepton. 100 $mgml^{-1}$ in DMSO

D6 DNA primers

MEL primer	5'-GGTGAGTTCCGGCATGC-3'
ORI primer	5'-TTGCCGTCCGCCTTGAG-3'
1212 primer	5'-GTTTTCCCAGTCAGGAC-3'
1233 primer	5'-AGCGGATAACAATTTGAGACAGGA-3'

Table D1 Composition of growth media[a]

COMPONENTS[b]	MEDIUM NAME								
	LB[h] (XGal)	YEME	R2YE	M3 M3/6.5 M3/7.0	MM	MMT	MMtCu(-)	MMTCu	M3T/6.5 M3T/7.0
Tryptone	1%	-	-	-	-	-	-	-	-
Yeast extract	0.5%	0.3%	0.5%[c]	0.5%	-	-	-	-	0.5%
Malt extract		0.3%	-	2.4%	-	-	-	-	2.4%
Peptone	-	0.5%	-	-	-	-	-	-	-
Casamino acids	-	-	0.01%	-	-	0.6%[c]	-	0.6%[c]	0.6%[c]
L-asparagine	-	-	-	-	0.05%	0.05%	0.05%	0.05%	-
L-proline	-	-	0.3%[c]	-	-	-	-	-	-
L-tyrosine	-	-	-	-	-	0.0375%[f]	-	0.0375%[f]	0.0375%[f]
Glucose	-	1%	1%	-	1%[c]	1%[c]	1%[c]	1%[c]	-
Sucrose	-	34%	10.3%	-	-	-	-	-	-
Glycine	-	0.5%[c]	-	-	-	-	-	-	-
NaCl	0.5%	-	-	-	-	-	-	-	-
NaOH	-	-	0.005M[c]	-	-	-	-	-	-
$CaCl_2$	-	-	0.29%[c]	-	-	-	-	-	-
$MgCl_2.6H_2O$	-	0.005M[c]	1.012%	-	-	-	-	-	-
$MgSO_4.7H_2O$	-	-	-	-	0.02%	0.02%	0.02%	0.02%	-
$FeSO_4.7H_2O$	-	-	-	-	0.001%	0.001%	0.001%	0.001%	0.001%
K_2HPO_4	-	-	0.005%[c]	-	0.05%	0.05%	0.05%	0.05%	-
Na_2HPO_4	-	-	-	[g]	-	-	-	-	[g]
NaH_2PO_4	-	-	-	[g]	-	-	-	-	[g]
$CuCl_2.2H_2O$	-	-	-	-	-	-	0.01%[c]	0.01%[c]	0.01%[c]
TES (pH 7.2)	-	-	0.573%[c]	-	-	-	-	-	-
Trace solution	-	-	0.02%[c]	-	-	-	-	-	-
Tiger Milk	-	-	-	-	-	0.75%[c]	0.75%[c]	0.75%[c]	-
XGal	0.002%	-	-	-	-	-	-	-	-
IPTG	0.3 mM	-	-	-	-	-	-	-	-
Agar[d]	1.5%	-	2.2%	2%	2%	2%	2%	2%	2%
pH	-	-	7.2	[g]	7.0	7.0	7.0	7.0	[g]

[a] Final concentration of components are given.

[b] Stock solutions used:

[c] Added after autoclaving.

[d] Solid media only.

[e] Oxoid #1 Agar. Low quality agar gives a poor regeneration rate (personal observation).

[f] Added after autoclaving in a basic solution (see elsewhere in Appendix D).

[g] For pH 6.5, Na_2HPO_4 0.016 M and NaH_2PO_4 0.034 M; for pH 7.0, Na_2HPO_4 0.029 M and NaH_2PO_4 0.021 M, not included in M3 medium.

[h] XGal and IPTG were not included in standard LB medium.

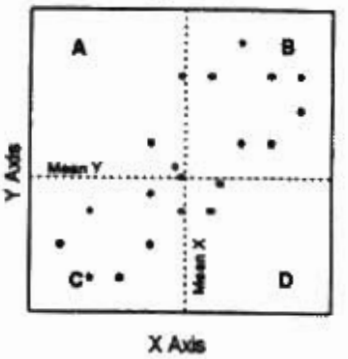
Table D2 Enzyme buffers

	BUFFER									
	RESTRICTION BUFFERS					T4	T4	KLENOW		
	L	M	H	A	B	LIGASE	POLYMERASE	FRAGMENT	CIAP	BAL31
CONCENTRATION	10x	10x	10x	10x	10x	5x	10x	10x	10x	5x
COMPONENT										
Tris	0.01 M	0.01 M	0.5 M	330 mM	0.1 M	250 mM	330 mM	500 mM	100 mM	250 mM
Potassium acetate	-	-	-	660 mM	-	-	660 mM	-	-	-
Magnesium acetate	-	-	-	100 mM	-	-	100 mM	-	-	-
$MgCl_2$	0.01 M	0.1 M	0.1 M	-	0.05 M	50 mM	-	-	10 mM	50 mM
$MgSO_4$	-	-	-	-	-	-	-	100 mM	-	-
$ZnCl_2$	-	-	-	-	-	-	-	-	10 mM	-
NaCl	-	0.05 M	1 M	-	1 M	-	-	-	-	3 M
$CaCl_2$	-	-	-	-	-	-	-	-	-	50 mM
DTT	-	-	-	5 mM	-	5 mM	5 mM	1 mM	-	-
DTE	0.01 M	0.01 M	0.01 M	-	-	-	-	-	-	-
ATP	-	-	-	-	-	5 mM	-	-	-	-
BSA	-	-	-	-	-	-	0.1%	0.05%	-	0.05%
PEG 6000	-	-	-	-	-	25%	-	-	-	-
2-mercaptoethanol	-	-	-	-	0.01 M	-	-	-	-	-
pH adjusted with	HCl	HCl	HCl	ac.acid	HCl	HCl	ac.acid	HCl	HCl	HCL
pH	7.5	7.5	7.9	8.0	7.6	7.9	7.2	8.3	8.0	8.0

APPENDIX E CORRELATION COEFFICIENT, AN EXPLANATION

$$r = \frac{\Sigma d_x d_y}{\sqrt{(\Sigma d_x^2)(\Sigma d_y^2)}}$$

Figure E



The correlation coefficient (r) measures the degree of correlation between a number of observations whose values are x and y. Such observations can also be plotted as points on a scatter graph, with the axes x and y, as shown in Fig. E.

In the above formula, deviation (d) is the difference between the value of an observation and the average value of all the observations. So

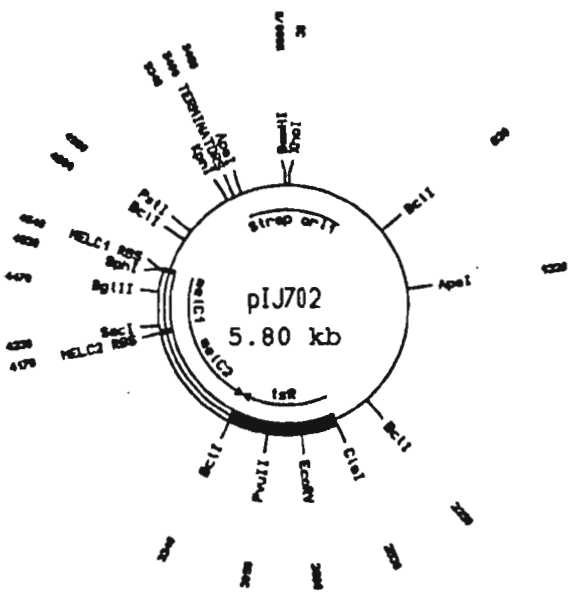
d_x=(x value of a single point)-(mean x value of all points)
d_y=(y value of a single point)-(mean y value of all points)
$d_x d_y$=product of the deviations of x and y for an observation
$\Sigma d_x d_y$=the sum of all the values of $d_x d_y$

In Fig. E it should be noted that all points that fall in the quadrants marked A or D give negative values of $d_x d_y$ (from the product of a negative and a positive pair of figures). Likewise all points that fall in the quadrants marked B or C give positive values of $d_x d_y$ (because d_x and d_y are either both positive or both negative). When there is no correlation between the values of the points, they are scattered in all four quadrants and $\Sigma d_x d_y$ is close to zero. If there is a positive correlation (as shown in Fig. E), more of the points are found in the B and C quadrants and $\Sigma d_x d_y$ is positive. A negative correlation gives a negative value for $\Sigma d_x d_y$.

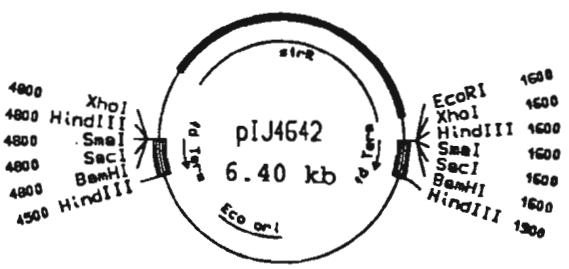
Dividing $\Sigma d_x d_y$ by $\sqrt{(\Sigma d_x^2)(\Sigma d_y^2)}$ makes r independent of the scale of measurement of x and y, and ensures that r has a maximum range of -1 to 1. Further information may be found in standard texts.

APPENDIX F PLASMID MAPS

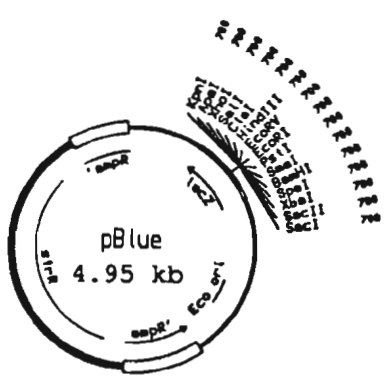
A.



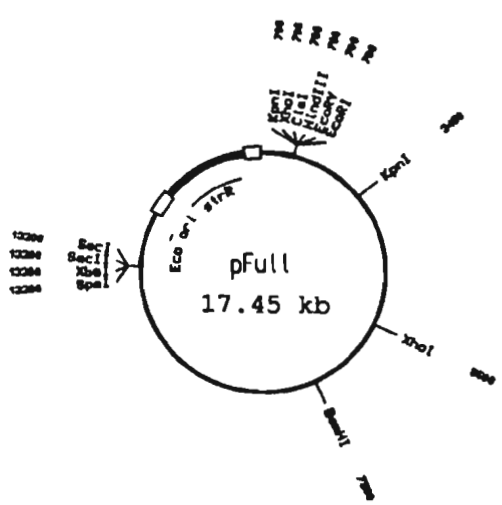
B.



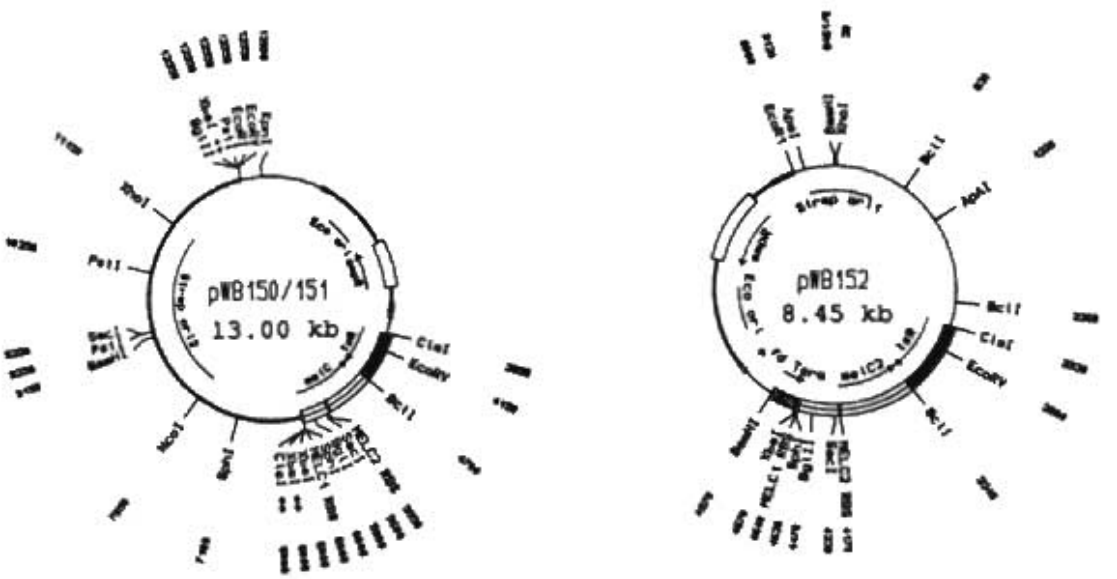
C.



D.



E. **F.**



A. pIJ702 was constructed by Katz *et al.* (1983) and carries markers for thiostrepton resistance (*tsr*) and melanin production (*melC*) and a high copy number origin of replication derived from pIJ101 (Kendal and Cohen, 1988).

B. pIJ4642 was a gift from M. Bibb (unpublished). It carries two copies of the coliphage fd transcription terminator, a streptomycin resistance marker (*str*R) and an *E.coli* origin of replication.

C. pBlue was constructed by Smith (1991). It consists of the streptomycin resistance marker from pIJ4642 cloned on a *Hin*dIII fragment into *Sca*I site of the plasmid Bluescript (Stratagene, San Diego, USA).

D. pFull was constructed by Smith (1991). It consists of a 12,5 kb origin of replication from plasmid pSPN1 cloned into the *Bam*HI site of pBlue.

E. and **F.** The construction of pWB150/151 and pWB152 is described in Chapter 2.

APPENDIX G COMPUTER PROGRAMS

The computer programs used were kindly written by Brendan Babb (Department of Medical Microbiology, University of Cape Town). The programs used, and information about them, are supplied on floppy disc. The disc can be found on the inside of the back cover. The programs are executable under MS DOS. All data used are also supplied on the disc. Instructions for use are under the MS Word5 file README.DOC.

BIBLIOGRAPHY

In cases which the reference below does not refer directly to a GenEmbl accession number but the DNA sequence was obtained from this source, accession numbers are given in brackets.

ADAMS, D.G. & CARR, N.G. (1989). Control of heterocyst development in the cyanobacterium *Anabaena cylindrica*. J. Gen. Microbiol. 135, 839-849.

AGNELL, S. et al. (1992). The glucose kinase gene of *Streptomyces coelicolor* A3(2): its nucleotide sequence, transcriptional analysis and role in glucose repression. Mol. Microbiol. 6, 2833-2844.

AGNELL, S. et al. (1994). Glucose repression in *Streptomyces coelicolor* A3(2): a likely role for glucose kinase. Mol. Gen. Genet. 244, 135-143.

ALTENBUCHNER, J. (1988). A new *E.coli* cloning vector containing a melanin marker for insertion screening. Nuc. Ac. Res. 16, 8710.

ALTSCHUL, S.F. et al., (1990). Basic local alignment search tool. J. Mol. Biol. 215, 403-410.

ALVAREZ, M.A. et al., (1994). Caracterization of yeast DNA sequences capable of directing transcription in *Streptomyces* and *E.coli*. FEMS Micrbiol. Lett. 115, 119-124.

ANDERSSON, S.G.E & KURLAND, C.G. (1990). Codon preferences in free-living organisms. Microbiol. Rev. 54, 198-210.

APELLIAN, D. & INOUYE, S. (1990). Development-specific σ-factor essential for late stage develpoment of *Myxococcus xanthus*. Genes Dev. 4, 1396-1403.

BABCOCK, M.J. & KENDRICK, K.E. (1990). Transcriptional and translational features of a sporulation gene of *Streptomyces griseus*. Gene 95, 57-63.

BASCARAN, V. et al. (1991). Stringent response and secondary metabolism in *Streptomyces clavuligerus*. J. Gen. Microbiol. 137, 1625-1634.

BEPPU, T. (1992). Secondary metabolites as chemical signals for cellular differentiation. Gene 115, 159-165.

BERNAN, V. et al. (1985). The nucleotide sequence of the tyrosinase gene from *Streptomyces antibioticus* and characterization of the gene product. Gene 37, 101-110.

BERG, O. & VON HIPPLE, P.H. (1987). Selection of DNA binding sites by regulatory proteins. J. Mol. Biol. 193, 723-750.

BETANCOURT, A.M. et al. (1992). Analysis of tyrosinase synthesis in *Streptomyces antibioticus*. J. Gen. Microbiol. 138, 787-794.

BIBB, M.J. & COHEN, S.N. (1982). Gene expression in *Streptomyces*: construction and application of promoter probe vectors. Mol. Gen. Genet. 187, 265-277.

BIBB, M.J. et al. (1984). The relationship between base composition and codon usage in bacterial genes and its use for the simple and reliable identification of protein-coding sequences. Gene, 30, 157-166.

BIBB, M.J. et al. (1985). Nucleotide sequence encoding and promoting expression of three antibiotic resistance genes indigenous to *Streptomyces*. Mol. Gen. Genet. 199, 26-36. (X02392).

BIBB, M.J. et al. (1987). The agarase gene (*dagA*) of *Streptomyces coelicolor* A3(2): affinity purification and characterization of the cloned gene product. J. Gen. Microbiol. 133, 2089-2096.

BLACK, T.A. & WOLK, C.A. (1994). Analysis of a Het$^-$ mutation in *Anabaena* sp. strain PCC 7120 implicates a secondary metabolite in the regulation of heterocyst spacing. J. Bacteriol. 176, 2282-2292.

BLANCO, G. *et al.* (1993). Hybridization and DNA sequence analyses suggest an early evolutionary divergence of related biosynthetic gene sets encoding polyketide antibiotics and spore pigments in *Streptomyces* spp. Gene 130, 107-116.

BOURRET, R.B. *et al.* (1991). Signal transduction pathways involving protein phosphorylation in prokaryotes. Ann. Rev. Biochem. 60, 401-441.

BOURN, W. (1986). Investigation of the DNA-binding proteins of *Streptomyces coelicolor* A3(2). Masters dissertation, University of Cape Town, Cape Town, South Africa.

BOYLAN, S.A. *et al.* (1993). Transcription factor σ^B of *Bacillus subtilis* controls a large stationary-phase regulon. J. Bacteriol. 175, 3957-3963.

BRAHAMSHA, B. & HASELKORN, R. (1992). Identification of multiple RNA polymerase sigma factor homologs in the cyanobacterium *Anabaena* sp. strain PCC 7120: cloning, expression, and inactivation of the *sig*B and *sig*C genes. J. Bacteriol. 174, 7273-7282.

BRANA, A.F. *et al.* (1982). Mode of cell wall growth of *Streptomyces antibioticus*. FEMS Microbiol. Letters 13, 231-235

BRANA, A.F. *et al.* (1986). Glycogen and trehalose accumulation during colony development in *Streptomyces antibioticus*. J. Gen. Microbiol. 132, 1319-1326.

BRASCH, M.A. *et al.* (1993). Activation of transcriptional fusions in *Streptomyces lividans* resulting from insertion of a 14-bp oligonucleotide. Nuc. Ac. Res. 21, 4515.

BRAWNER, M.E. *et al.* (1985). Characterization of *Streptomyces* promoter sequences using the *Escherichia coli* galactokinase gene. Gene 40, 191-201.

BROWN, K.L. *et al.* (1992). Isolation and characterization of the major vegetative RNA polymerase of *Streptomyces coelicolor* A3(2); renaturation of a sigma subunit using GroEL. Mol. Microbiol. 6, 1133-1139.

BRUTON, C.J. *et al.* (1991). Phage vectors that allow monitoring of transcription of secondary metabolism genes in *Streptomyces*. Biotechnology 9, 652-6564.

BRUTON, C.J. (1993). Unpublished. X73903.

BUCCA, G. *et al.* (1994). Unpublished. L27820.

BURTON, Z. *et al.* (1981). The nucleotide sequence of the cloned *rpoD* gene for the RNA polymerase sigma subunit from *E.coli* K12. Nuc. Ac. Res. 9, 2889-2903.

BUSBY, S. & EBRIGHT, R.H. (1994). Promoter structure, promoter recognition, and transcription activation in prokaryotes. Cell 79, 743-746.

BUTTNER, M.J. (1989). RNA polymerase heterogeneity in *Streptomyces coelicolor* A3(2). Molecular Microbiol. 3, 1653-1659.

BUTTNER, M.J. *et al.* (1990). Cloning, disruption, and transcriptional analysis of three RNA polymerase sigma factor genes of *Streptomyces coelicolor* A3(2). J. Bacteriol. 172, 3367-3378.

BUTTNER, M.J. & LEWIS, C.G. (1992). Construction and characterization of *Streptomyces coelicolor* A3(2) mutants that are multiply deficient in the nonessential *hrd*-encoded RNA polymerase sigma factors. J. Bacteriol. 174, 5165-5167.

BUTTNER, M.J. *et al.* (1987). The agarase gene (*dagA*) of *Streptomyces coelicolor* A3(2): nucleotide sequence and trascriptional analysis. Mol. Gen Genet. 209, 101-109. (X05811)

BUTTNER, M.J. *et al.* (1988). At least three different RNA polymerase holoenzymes direct transcription of the agarase gene (*dagA*) of *Streptomyces coelicolor* A3(2). Cell 52, 599-607.

BUTTNER, M.J. *et al.* (1990). Cloning, disruption, and transcriptional analysis of three RNA polymerase sigma factor genes of *Streptomyces coelicolor* A3(2). J. Bacteriol. 172, 3367-3378.

CASHEL, M. & RUDD, K. (1987). The stringent response. In *Escherichia coli* and *Salmonella typhimurium*, (Ed. F.C. Niedhart) 1410-1438. American Society for Microbiology, Washington D.C.

CHAMBERS, S.P. *et al.* (1988). The pMTL *nic*⁻ cloning vectors. 1. Improved pUC polylinker regions to facilitate the use of sonicated DNA for nucleotide sequencing. Gene 68, 139-149.

CHAMBLISS, G.H. (1993). Carbon source-mediated catabolite repression. *Bacillus subtilis* and other Gram-positive bacteria. (ed. A.L. Sonenshein, J.A. Hoch & R. Losick) 213-219, American Society for Microbiology, Washington DC.

CHAMPNESS, W.C. (1988). New loci required for *Streptomyces coelicolor* morphological and physiological development. J. Bacteriol. 170, 1168-1174.

CHAMPNESS, W.C. *et al.* (1992). Identification of *Streptomyces coelicolor* genes involved in regulation of antibiotic synthesis. Gene 115, 55-60.

CHATER, K.F. (1972). A morphological and genetic mapping study of white colony mutants of *Streptomyces coelicolor*. J. Gen. Microbiol. 72, 9-28.

CHATER, K.F. (1975). Construction and phenotypes of double sporulation deficient mutants in *Streptomyces coelicolor* A3(2). J. Gen. Microbiol. 87, 312-325.

CHATER, K.F. (1984). Morphological and phisiological differentiation in *Streptomyces*. Microbial Development. (ed. R. Losick & L. Shapiro) 89-115, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

CHATER, K.F. (1989a). Sporulation in *Streptomyces*. Regulation of Procaryotic Development (ed. I. Smith, R.A. Slepecky & P. Setlow) 277-299, American Society for Microbiology, Washington DC, USA.

CHATER, K.F. (1989b). Multilevel regulation of *Streptomyces* differentiation. Trends in Genetics 5, 372-377.

CHATER, K.F. (1990). The improving prospects for yield increase by genetic engineering in antibiotic-producing streptomcetes. Biotechnology 8, 115-121.

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

CHATER, K.F. & BRUTON C.J. (1985). Resistance, regulatory and production genes for the antibiotic methylenomycin are clustered. EMBO 4, 1893-1892.

CHATER, K.F. & HOPWOOD, D.A. (1993). *Streptomyces*. In *Bacilus subtilis* and other Gram-positive bacteria. (Ed. A.L. Sonenshein, J.A. Hoch & R. Losick) 83-99, American Society for Microbiology, Washington DC.

CHATER, K.F. *et al.* (1989). The developmental fate of *S.coelicolor* hyphae depends on a gene product homologous with the motility sigma factor of *B.subtilis*. Cell 59, 133-143. (J03169).

CHEN, C.W. *et al.* (1992). Discovery and characterization of a new transposable element, Tn4811, in *Streptomyces lividans* 66. J. Bacteriol. 174, 7762-7769.

CHEN, L-Y. *et al.* (1992). Copper transfer and activation of the *Streptomyces* apotyrosinase are mediated through a complex formation between apotyrosinase and its *trans*-activator MelC1. J. Biol. Chem. 267, 20100-20107.

CHEN, L-Y. *et al.* (1993). Mutational study of *Streptomyces* tyrosinase *trans*-activator MelC1. J. Biol. Chem. 268, 18710-18716.

CLARKE, G.M. & COOKE, D. (1983). A basic course in statistics. 2nd editition. Edward Arnold Publishers Ltd., London U.K.

COCO, E.A. *et al.* (1991). New classes of *Streptomyces coelicolor* A3(2) mutants blocked in undecylprodigiosin (Red) biosynthesis. Mol. Gen. Genet. 227, 28-32.

COHN, D.H. *et al.* (1985). Nucleotide sequence of the *luxA* gene of *Vibrio harveyi* and the complete amino acid sequence of the α subunit of bacterial luciferase. J. Biol. Chem. 260, 6139-6146.

COLLADO-VIDES, J. *et al.* (1991). Control site location and transcriptional regulation in *Escherichia coli*. Microbiol. Rev. 55, 371-394.

CRAMERI, R. *et al.* (1982). Secretion of tyrosinase in *Streptomyce glaucescens*. J. Gen. Microbiol. 128, 371-379.

DAVIS, N.K. & CHATER, K.F. (1990). Spore colour in *Streptomyces coelicolor* A3(2) involves the developmentally regulated synthesis of a compound biosynthetically related to polyketide antibiotics. Molecular Microbiol. 4, 1679-1691. (X55942).

DAVIS, N.K. & CHATER, K.F. (1992). The *Streptomyces coelicolor whiB* gene encodes a small transcription factor-like protein dispensible for growth but essential for sporulation. Mol. Gen. Genet. 232, 351-358. (X62287).

DAZA, A. *et al.* (1989). Sporulation of several species of *Streptomyces* in submerged cultures after nutritional shiftdown, J. Gen. Microbiol. 135, 2483-2491.

DELIC, I. *et al.* (1992). Direct repeat sequences are implicated in the regulation of two *Streptomyces* chitinase promoters that are subject to carbon catabolite control. Proc. Natl. Accad Sci. 89, 1885-1889.

DONADIO, S. *et al.* (1991). Modular organisation of genes required for complex polyketide biosynthesis. Science 252, 675-679.

DUNCAN, M.L. *et al.* (1987). Gene encoding the 37,000-Dalton minor sigma factor of *Bacillus subtilis* RNA polymerase: isolation, nucleotide sequence, chromosomal locus, and cryptic function. J.Bacteriol. 169, 771-778.

DWORKIN, M. (1985). Developmental Biology of the Bacteria. Benjamin/Cummings Publishing Company, Inc.

ELY, B. & SHAPIRO, L. (1984). Regulation of cell differentiation in *Caulobacter crescentus*. Microbial Development. (ed. R. Losick & L. Shapiro) 89-115, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

ERRINGTON, J. (1993). *Bacillus subtilis* sporulation: regulation of gene expression and control of morphogenesis. Micrbiol. Rev. 57, 1-33.

FEITELSON, J.S. (1988). An improved vector for the isolation and analysis of *Streptomyces* promoters. Gene 66, 159-162.

FEITELSON, J.S. *et al.* (1985). Genetic and biochemical characterisation of the *red* gene cluster of *Streptomyces coelicolor* A3(2). J. Gen. Microbiol. 131, 2431-2441.

FEITELSON, J.S. *et al.* (1986). Molecular genetics of *red* biosynthesis in *Streptomyces*. J. Natural Products 49, 988-994.

FERNANDEZ-MORENO, M.A. *et al.* (1991). The *act* cluster contains regulatory and antibiotic export genes, direct targets for translational control by the *bldA* tRNA gene of Streptomyces. Cell 66, 769-780. (M64683).

FORNWALD, J.A. *et al.* (1987). Two promoters, one inducible and one constitutive, control transcription of the *Streptomyces lividans* galactose operon. Proc. Natl. Accad. Sci. 84, 2130-2134.

FORSBERG, A..J. *et al* (1994). Use of transcriptional fusions to monitor gene expression: a cautionary tale. J. Bacteriol. 176, 2128-2132.

FORSMAN, M. & GRANSTROM, M. (1992). Mutagenic analysis of the promoter of *Streptomyces fradiae* β-lactamase-encoding gene. Gene 121, 87-94.

FORSMAN, M. & JAURIN, J. (1987). Chromogenic identification of promoters in *Streptomyces lividans* by using an *ampC* β-lactamase promoter-probe vector. Mol. Gen. Genet. 210, 23-32.

GEISTLICH, M. *et al.* (1989). Localization and functional analysis of the regulated promoter from the *Streptomyces glaucescens mel* operon. Molecular Microbiol. 3, 1061-1069.

GEISTLICH, M. *et al.* (1992). Characterization of a novel regulatory gene governing the expression of a polyketide synthase gene in *Streptomyces* ambofaciens. Molecular Microbiol. 6, 2019-2029.

GLAZEBROOK, M.A. *et al.* (1990). Sporulation of *Streptomyces venezuelae* in submerged cultures. J. Gen. Microbiol. 136, 581-588.

GOODFELLOW, M. & WILLIAMS, S.T. (1983). Ecology of the actinomycetes. Ann. Rev. Microbiol. 37, 189-216.

GRAMAJO, H.C. *et al.* (1993). Stationary phase production of the antibiotic actinorhodin in *Streptomyces coelicolor* A3(2) is transcriptionally regulated. Molecular Microbiol. 7, 837-845.

GRANOZZI, C. *et al.* (1990). A breakdown in macromolecular synthesis preceeding differentiation in *Streptomyces coelicolor* A3(2). J. Gen. Microbiol. 136, 713-716.

GRAY, D.I., *et al.* (1990). Apical hyphal extension in *Streptomyces coelicolor* A3(2). J. Gen. Microbiol. 136, 1077-1084.

GRIBSKOV, M. *et al.*, (1984). The codon preference plot: graphic anlysis of protein coding sequences and prediction of gene expression. Nuc. Ac. Res. 12, 539-549.

GROSSMAN, A.D. *et al.* (1984). The *htpR* gene product of E.coli is a sigma factor for heat-shock promoters. Cell 38, 383-390.

GUIJARRO, J. *et al.* (1988). Promoter determining the timing and spatial localization of transcription of a cloned *Streptomyces coelicolor* gene encoding a spore-associated polypeptide. J. Bacteriol. 170, 1895-1901.

GUILFOILE, P.G. & HUTCHINSON C.R. (1992). Sequence and transcriptional analysis of the *Streptomyces glaucesens tcmAR* tetracenomycin C resistance repressor gene loci. J. Bacteriol. 174, 3651-3658.

GUNTER, K. *et al.* (1992). Characterization of an iron-regulated promoter involved in desferrioxamone B synthesis in *Streptomyces pilosus*: repressor-binding site and homology to the diptheria toxin gene promoter. J. Bacteriol. 175, 3295-3302.

GUTHRIE, E.P. & CHATER, K.F. (1990). The level of a transcript required for production of a *Streptomyces coelicolor* antibiotic is conditionally dependent on a tRNA gene. J. Bacteriol. 172, 6189-6193.

HARA O. *et al.* (1983). Genetic analysis of A-factor synthesis in *Streptomyces coelicolor* A3(2) and *Streptomyces griseus*. J. Gen. Microbiol. 129, 2939-2944.

HARASYM, M. *et al.* (1990). The *Streptomyces coelicolor* A3(2) *bldB* region contains at least two genes involved in morphological development. J. Gen. Microbiol. 136, 1543-1550.

HARDISSON, C. & MANZANAL, M-B. (1976). Ultrastructural studies of sporulation in *Streptomyces*. J. Bacteriol. 127, 1443-1454

HARDISSON, C. *et al.* (1978). Fine structure, physiology and biochemistry of arthrospore germination in *Streptomyces antibioticus*. J. Gen. Microbiol. 105, 203-214.

HARLEY, C.B. & REYNOLDS, R.P. (1987). Analysis of *E.coli* promoters. Nuc. Ac. Res. 15, 2343-2361.

HASELKORN, R. (1992). Developmentally regulated gene rearrangements in prokaryotes. Annu. Rev. Genet. 26, 113-130.

HASELKORN, R. *et al.* (1986). Developmental rearrangement of cyanobacterial nitrogen-fixation genes. Trends Genet. 2, 255-259.

HAWLEY, D.K. & McCLURE W. (1983). Compilation and analysis of *Escherichia coli* promoter DNA sequences. Nuc. Ac. Res. 11, 2237-2255.

HELD, T. & KUTZNER, H.J. (1990). Transcription of the tyrosinase gene of *Streptomyces michiganensis* DSM 40015 is induced by copper and repressed by ammonium. J. Gen. Microbiol. 136, 2413-2419.

HELMANN, J.D. (1991). Alternative sigma factors and the regulation of flagellar gene expression. Mol. Microbiol. 5, 2875-2882.

HENGGE-ARONIS, R. (1993). Survival of hunger and stress: the role of *rpoS* in early stationary phase gene regulation in *E.coli*. Cell 72, 165-168.

HINTERMANN, G. *et al.* (1985). Cloning and expression of the genetically unstable tyrosinase structural gene from *Streptomyces glaucescens*. Mol. Gen. Genet. 200, 422-432.

HIRSCHMAN, J. *et al.* (1985). Products of the nitrogen regulatory genes *ntrA* and *ntrC* of enteric bacteria activate *glnA* transcription *in vitro*: evidence that the *ntrA* product is a σ factor. Proc. Natl. Acad. Sci. 82, 7525-7529.

HODGSON, D.A. (1982). Glucose repression of carbon source uptake and metabolism in *Streptomyces coelicolor* A3(2) and its perturbation in mutants resistant to 2-deoxyglucose. J. Gen. Microbiol. 128, 2417-2430.

HODGSON, D.A. (1992). Differentiation in actinomycetes. Prokaryotic Structure and Function: A New Perspective. (ed. S. Mohan, C. Dow & J.A. Cole) 407-440, Cambridge University Press, Cambridge.

HOLMES, D.J. *et al.* (1993). Autogenous transcriptional activation of a thiostrepton induced gene in *Streptomyces lividans*. EMBO 12, 318-319.

HONG, S-K. *et al.* (1991). Phosphorylation of the AfsR product, a global regulatory protein for secondary-metabolite formation in *Streptomyces coelicolor* A3(2). J. Bacteriol. 173, 2311-2318.

HOPWOOD, D.A. & MERRICK, M.J. (1977). Genetics of antibiotic production. Bacteriological Reviews. 41, 595-635.

HOPWOOD, D.A. & WRIGHT, H.M. (1983). CDA is a new chromosomally determined antibiotic in *Streptomyces coelicolor*. J. Gen. Microbiol. 129, 3575-3579.

HOPWOOD, D.A. *et al.* (1970). Mutants of *Streptomyces coelicolor* defective in sporulation. J. Gen. Microbiol. 61, 397-408.

HOPWOOD, D.A. *et al.* (1985a). Genetic manipulation of *Streptomyces*: A laboratory manual. The John Innes Foundation, Norwich, England.

HOPWOOD, D.A. *et al.* (1993). The chromosome map of *Streptomyces coelicolor* A3(2). In *Bacilus subtilis* and other Gram-positive bacteria. (ed. A.L. Sonenshein, J.A. Hoch & R. Losick) 497-504, American Society for Microbiology, Washington DC.

HOPWOOD, D.A. *et al.*. (1985b). Production of "hybrid" antibiotics by genetic engineering. Nature 134, 642-644.

HORINOUCHI, S. & BEPPU, T. (1984). Production in large quantities of actinorhodin and undecylprodigiosin induced by *afsB* in *Streptomyces lividans*. Agric. Biol. Chem. 48, 2131-2133.

HORINOUCHI, S. & BEPPU, T. (1985). Construction and application of a promoter-probe plasmid that allows chromogenic identifcation in *Streptomyces lividans*. J. Bacteriol. 162, 406-412.

HORINOUCHI, S. & BEPPU, T. (1992a). Autoregulatory factors and communication in Actinomycetes. Annu. Rev. Microbiol. 46, 377-398.

HORINOUCHI, S. & BEPPU, T. (1992b). Regulation of secondary metabolism and cell differentiation in *Streptomyces*: A-factor as a microbial hormone and the AfsR protein as a component of a two component regulatory system. Gene 115 167-172.

HORINOUCHI, S. & BEPPU, T. (1994). A-factor as a microbial hormone that controls cellular differntiation and secondary metabolism in *Streptomyces griseus*. Mol. Microbiol. 12, 859-864.

HORINOUCHI, S. *et al.* (1983). Cloning of a pleiotrropic gene that positively controls biosynthesis of A-factor, actinorhodin and prodigiosin in *Streptomyces lividans*. J. Bacteriol. 155, 1238-1248.

HORINOUCHI, S. *et al.* (1984). Unstable genetic determinant of A-factor biosynthesis in streptomycin-producing organisms: Cloning and characterisation. J. Bacteriol. 158, 481-487.

HORINOUCHI, S. *et al.* (1986). Nucleotide sequence of *afsB*, a pleiotropic gene involved in secondary metabolism in *Streptomyces coelicolor* A3(2) and "*Streptomyces lividans*". J. Bacteriol. 168, 257-269. (M16011).

HORINOUCHI, S. *et al.* (1989a). *afsB* stimulates transcription of the actinorhodin biosynthetic pathway in *Streptomyces coelicolor* A3(2) and *Streptomyces lividans*. Mol. Gen Genet. 215, 355-357.

HORINOUCHI, S. *et al.* (1989b). Nucleotide sequence and transcriptional analysis of the *Streptomyces griseus* gene (*afsA*) responsible for A-factor biosynthesis. J. Bacteriol. 171, 1206-1210.

HORINOUCHI, S. *et al.* (1990). Primary structure of AfsR, a global regulatory protein for secondary metabolite formation in *Streptomyces coelicolor* A3(2). Gene 95, 45-46.

HUBER, M. *et al.* (1985). Primary structure of tyrosinase from *Streptomyces glaucescens*. Biochemistry 24, 6038-6044.

HUBER, M. *et al.* (1987). The promoter of the *Streptomyces glaucescens mel* operon. Nuc. Ac. Res. 19, 8106.

ISHIKAWA, J. & HOTTA, K. (1991). Nucleotide sequence and transcriptional start point of the *kan* gene encoding an aminoglycoside

3-*N*-acetyltransferase from *Streptomyces griseus* ss-1198PR. Gene 108, 127-132.

ISHIZUKA, H. *et al.* (1992). A putative two-component regulatory system involved in secondary metabolism in *Streptomyces* spp. J. Bacteriol. 174, 7585-7594.

INGRAM, C. *et al.* (1989). *xylE* functions as an efficient reporter gene in *Streptomyces* spp.: Use for the study of *galP*, a catabolite-controlled promoter. J. Bacteriol. 171, 6617-6624.

JOHNSTON, T.C. *et al.* (1986). Nucleotide sequence of the *luxB* gene of *Vibrio harveyi* and the complete amino acid sequence of the β subunit of bacterial luciferase. J. Biol. Chem. 261, 4805-4811.

KATO, F. *et al.* (1995). Carotenoid synthesis in *Streptomyces setonii* ISP5395 is induced by *crtS*, whose product is similar to a sigma factor. Mol. Gen. Genet. 247, 387-390.

KATZ, E. & BETANCOURT, A.M. (1988). Induction of tyrosinase by L-methionine in *Streptomyces antibioticus*. Can. J. Micribiol. 34, 1297-1303.

KATZ, E. *et al.* (1983). Cloning and expression of the tyrosinase gene from *Streptomyces antibioticus* in *Streptomyces lividans*. J. Gen. Microbiol. 129, 2703-2714.

KENDALL, K.J. & COHEN, S.N. (1988). Complete nucleotide sequence of the *Streptomyces lividans* plasmid pIJ101 and correlation of the sequence with genetic properties. J. Bacteriol. 170, 4634-5651.

KENDRICK, K.E. & ENSIGN, J.C. (1983). Sporulation of *Streptomyces griseus* in submerged culture. J. Bacteriol. 155, 357-366.

KHOSLA, C. *et al.* (1993). Genetic construction and functional analysis of hybrid polyketide synthases containing heterologous acyl carrier proteins. J. Bacteriol. 175, 2197-2204.

KIM, S.K. & KAISER, D. (1992). Control of cell density and pattern by intercellular signaling in *Myxococcus* development. Annu. Rev. Microbiol. 46, 117-139.

KING, A. & CHATER, K. (1986). The expression of the *Escherichia coli lacZ* gene in *Streptomyces*. J. Gen. Microbiol. 132, 1739-1752.

KIRBY, R. *et al.* (1975). Plasmid determined antibiotic synthesis and resistance in *Stretomyces coelicolor*. Nature 254, 265-267.

KOBAYASHI, T. *et al.* (1989). Molecular characterisation of a gene encoding a photolyase from *Streptomyces griseus*. Nuc. Ac. Res. 17, 4731-4744. (X15060).

KOEPSEL, R. & ENSIGN, J.C. (1984). Microcycle sporulation of *Streptomyces viridochromogenes*. Arch. Microbiol. 140, 9-14.

KORMANEC, J. & FARKOSOVSKY, M. (1993). Differential expression of principle sigma factor homologues of *Streptomyces aureofaciens* correlates with developmental stage. Nuc. Ac. Res. 21, 3647-3652.

KORMANEC, J. & HOMEROVA, D. (1993). *Streptomyces aureofaciens whiB* gene encoding putative transcription factor essential for differentiation. Nuc. Ac. Res. 21, 2515.

KORMANEC, J. *et al.* (1992). Four genes in *Streptomyces aureofaciens* containing a domain characteristic of principle sigma factors. Gene 122, 63-70.

KOZAK, M. (1983). Comparison of initiation of protein synthesis in procaryotes, eucaryotes and organelles. Microbiol. Rev. 47, 1-45.

KRETSCHMER, S. (1987). Nucleoid segregation pattern during branching in *Streptomyces granaticolor* mycelia. J. Basic Microbiol. 27, 203-206.

KRETSCHMER, S. & KUMMER, C. (1987). Increase of nucleoid size with increasing age of hyphal region in vegetative mycelia of *Streptomyces granaticolor*. J. Basic Microbiol. 27, 23-27.

KWAKMAN, J.J.J.M & POSTMA, P.W. (1994). Glucose kinase has a regulatory role in carbon catabolite repression in *Streptomyces coelicolor*. J. Bacteriol. 176, 2694-2698.

KUMAR, A. *et al.* (1993). The minus 35-recognition region of *Escherichia coli* sigma 70 is inessential for initiation of transcription at an "extended minus 10" promoter. J. Mol. Biol. 232, 406-418.

LAKEY, J.H. *et al.* (1983). A new channel-forming antibiotic from *Streptomyces coelicolor* A3(2) requires calcium for its activity. J. Gen. Microbiol. 129, 3565-3573.

LANDERS, T.A. (1987). Cloning in *Streptomyces lividans*: Colony lifts and replica plates. Focus 9, 1.

LAWLOR, E.J. *et al.* (1987). Pleiotropic morphological and antibiotic defficiencies result from mutations in a gene encoding a tRNA-like product in *Streptomyces coelicolor* A3(2). Genes and Development 1, 1305-1310.

LEBLOND, P. *et al.*, (1991). Genetic instability and associated genome plasticity in *Streptomyces ambofaciens*: pulsed-field gel electrophoresis evidence for large DNA alterations in a limited genomic region. J. Bacteriol. 173, 4229-4233.

LEE, Y-H.W. *et al.* (1988). A *trans*-acting gene is required for the phenotypic expression of tyrosinase in *Streptomyces*. Gene 65, 71-81.

LESKIW, B.K. *et al.* (1991a). The use of a rare codon specifically during development? Molecular Microbiol. 5, 2861-2867.

LESKIW, B.K. *et al.* (1991b). TTA codons in some genes prevent their expression in a class of developmental, antbiotic-negative *Streptomyces* mutants. Proc. Natl. Acad. Sci. USA. 88, 2461-2465.

LESKIW, B.K. *et al.* (1993). Accumulation of *BldA*-specified tRNA is temporally regulated in *Streptomyces coelicolor* A3(2). J. Bacteriol. 175, 1995-2005.

LEU, M-L. *et al.* (1992). Secretion of the *Streptomyces* tyrosinase is mediated through its *trans*-activator protein, MelC1. J. Biol. Chem. 267, 20108-20113.

LEU, W-M. *et al.* (1989). Analysis of the promoter region of the melanin locus from *Streptomyces antibioticus*. Gene 84, 267-277.

LIN, Y-S. *et al.* (1993). The chromosomal DNA of *Streptomyces lividans* 66 is linear. Mol Microbiol. 10, 923-933.

LONETTO, M.A. *et al.* (1994). Analysis of the *Streptomyces coelicolor sigE* gene reveals the existence of a subfamily of eubacterial RNA polymerase σ factors involved in the regulation of extracytoplasmic functions. Proc. Natl. Accad Sci. USA 91, 7573-7577.

LOPEZ, J.M. *et al.* (1979). The decrease of guanine nucleotides initiates sporulation of *Bacillus subtilis*. Biochim. Biophys. Acta 587, 238-252.

LOPEZ, J.M. *et al.* (1981). Response of guanosine 5'-triphosphate concentration and its significance for *Bacillus subtilis*. J. Bacteriol. 146, 605-613.

LOSICK, R. & STRAGIER, P. (1992). Crisscross regulation of cell-type-specific gene expression during development in *B.subtilis*. Nature 355, 601-604.

LUDOVICE, M. *et al.* (1992). Characterization of the *Streptomyces clavuligerus argC* gene encoding N-acetylglutamyl-phosphate reductase: Expression in *Streptomyces lividans* and effect on clavulanic acid production. J. Bacteriol. 174, 4606-4613. (M83659).

MALPARTIDA, F. & HOPWOOD, D.A. (1984). Molecular cloning of the whole biosynthetic pathway of a *Streptomyces* antibiotic and its expression in a heterologous host. Nature 309, 462-464.

MALPARTIDA, F. & HOPWOOD, D.A. (1986). Physical and genetic characterisation of the gene cluster for the antibiotic actinorhodin in *Streptomyces coelicolor* A3(2). Mol. Gen. Genet. 205, 66-73.

MANIATIS, T. *et al.* (1982). Molecular cloning: A laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

MANZANAL, M-B. & HARDISSON, C. (1978). Early stages of arthrospore maturation in *Streptomyces*. J. Bacteriol. 133, 293-297.

MARTIN, J.F. & LIRAS, P. (1989). Organisation and expression of genes involved in the biosynthesis of antibiotics and other secondary metabolites. Annu. Rev. Microbiol. 43, 173-206.

MASUDA, E.S. *et al.* **(1988).** Two developmental genes encoding σ factor homologs are arranged in tandem in *Bacillus subtilis*. Proc. Natl. Acad. Sci. USA 85, 7637-7641.

McBRIDE, M.J. & ENSIGN, J.C. (1987). Metabolism of endogenous trehalose by *Streptomyces griseus* spores and by spores or cells of other actinomycetes. J. Bacteriol. 169, 5002-5007.

McCARTHY, A.J. & WILLIAMS, S.T. (1992). Actinomycetes as agents of biodegradation in the environment - a review. Gene 115, 189-192.

McCUE, L.A. *et al.* (1992). Molecular analysis of sporulation in *Streptomyces griseus*. Gene 115, 173-179.

McLAUGHLIN, J.R. *et al.* (1981). Unique features in the ribosome binding site sequence of the Gram-positive *Staphylococcus aureus* β-lactamase gene. J. Biol. Chem. 256, 11283-11291.

McVITTIE, A. (1974). Ultrastructural studies on sporulation in wild type and white colony mutants of *Streptomyces coelicolor*. J. Gen. Microbiol. 81, 291-302.

MELTON, R.E. & KIESER, T. (1988). Regeneration of *Streptomyces lividans* protoplasts on media containing antibacterial and antifungal agents. Trends in Genetics 4, 325.

MENDEZ, C. & CHATER, K. (1987). Cloning of *whiG*, a gene critical for sporulation of *Streptomyces coelicolor* A3(2). J. Bacteriol. 169, 5715-5720.

MENDEZ, C. *et al.* (1985). Role of the substrate mycelium in colony development in *Streptomyces*. Can J. Microbiol. 31, 446-450.

MERRICK, M.J. (1976). A morpological and genetic mapping study of bald colony mutants of *Streptomyces coelicolor*. J. Gen. Microbiol. 96, 299-315.

MIYAKE, K. *et al.* (1989). Detection and properties of A-factor-binding protein from *Streptomyces griseus*. J. Bacteriol. 171, 4298-4302.

MIYAKE, K. *et al.* (1990). The A-factor-binding protein of *Streptomyces griseus* negatively controls streptomycin production and sporulation. J. Bacteriol. 172, 3003-3008.

MIYASHITE, K. & FUJII, T. (1993). Nucleotide sequence and analysis of a gene (*chiA*) for a chitinase from *Streptomyces lividans* 66. Biosci. Biotech. Biochem. 57, 1691-1698. (D13775).

NEESEN, K. & VOLCKAERT, G. (1989). Construction and shuttling of novel bifunctional vectors for *Streptomyces* spp. and *Escherichia coil*. J. Bacteriol. 171, 15691573.

NEWTON, A. & OHATA, N. (1990). Regulation of the cell division cycle and differentiation in bacteria. Annu. Rev. Microbiol. 44, 689-719.

OCHI, K. (1986a). A decrease in GTP content is associated with aerial mycelium formation in *Streptomyces* MA406-A-1. J. Gen. Microbiol. 132, 299-305.

OCHI, K. (1986b). Occurence of the stringent response in *Streptomyces* sp. and its significance for the initiation of morphological and physiological differentiation. J. Gen. Microbiol. 132, 2621-2631.

OCHI, K. (1987a). Changes in nucleotide pools during sporulation of *Streptomyces griseus* in submerged culture. J. Gen. Microbiol. 133, 2787-2795.

OCHI, K. (1987b). Metabolic initiation of differentiation and secondary metabolism by *Streptomyces griseus*: significance of the stringent response (ppGpp) and GTP content in relation to A factor. J. Bacteriol. 169, 3608-3616.

OCHI, K. (1987c). A *rel* mutation abolishes the enzyme induction needed for actinomycin synthesis by *Streptomyces antbioticus*. Agric. Biol. Chem. 51, 829-835.

OGAWARA, H. *et al.* (1993). Cloning, sequence and expression of the *argG* gene from *Streptomyces lavendulae*. Gene 125, 91-96

OKAMOTO, S. *et al.* (1992). Purification and molecular cloning of a butyryolactone autoregulator receptor from *Streptomyces virginiae*. J. Biol. Chem. 267, 1093-1098.

PAGET, M.S.B. *et al.* (1994). Construction and application of streptomycete promoter probe vectors which employ the *Streptomyces* glaucescens tyrosinase-encoding gene as a reporter. Gene 146, 105-110.

PARADAKAR, A.S. *et al.* (1994). Transcriptional analysis and heterologous expression of a gene encoding β-lactamase inhibitor protein (BLIP) from *Streptomyces clavuligerus*. Gene 144, 31-36.

PARKINSON, J.S. (1993). Signal transduction schemes of bacteria. Cell 73, 857-871.

PEARSON, W.R. & LIPMAN, D.J. (1988). Improved tools for biological sequence comparison. Proc. Natl. Accad. Sci. USA 85, 24442448.

PEREZ, C. (1993). Cloning, characterization, and expression in *Streptomyces lividans* 66 of an extracellular lipase-encoding gene from *Streptomyces* sp. M11. Gene 123, 109-114.

PIRET, J.M. & CHATER, K.F. (1985). Phage-mediated cloning of *bldA*, a region involved in *Streptomyces coelicolor* morpological development, and its analysis by genetic complementation. J. Bacteriol. 163, 965-972.

PORTMORE, J.J. *et al.* (1987). Bifunctional cosmid vector for *Streptomyces*. Biotechnology Lett. 9, 7-12.

POTUCKOVA, L. *et al.* (1995). A new RNA polymerase sigma factor, σ^F, is required for the late stages of morphological differentiation in *Streptomyces* spp. Mol. Microbiol. 17, 3748.

PREDICH, M. *et al.* (1995). Characterization of RNA polymerase and two sigma-factor genes from *Mycobacterium smegmatis*. Mol. Microbiol. 15, 355-366.

PUGLIA, A.M. & CAPPELLETTI, E. (1984). A bald ultrafertile UV-resistantstrain in *Streptomyces coelicolor* A3(2). Microbiologica 7, 263-266.

RIESENBERG, D. *et al.* (1984). Effect of serine hydroxamate and methyl α-D-glucopyranoside treatment on nucleoside polyphosphate pools, RNA and protein accumulation in *Streptomyces hygroscopicus*. J. Gen. Microbiol. 130, 2549-2558.

ROGERSON, A.C. (1989). The sequence assymetry of the *Escherichia coli* chromosome appears to be independent of strand or function and may be evolutionarily conserved. Nuc. AC. Res. 17, 5547-5563.

RUDD, B.A.M. & HOPWOOD, D.A. (1979). Genetics of actinorhodin biosynthesis by *Streptomyces coelicolor* A3(2). J. Gen. Microbiol. 114, 35-43.

RUDD, B.A.M. & HOPWOOD, D.A. (1980). A pigmented mycelial antibiotic in *Streptomyces coelicolor*: Control by a chromosomal gene cluster. J. Gen. Microbiol. 119, 333-340.

SAIER, M.H. (1991). A multiplicity of potential carbon catabolite repression mechanisms in prokaryotic and eukaryotic microorganisms. The New Biologist 3, 1137-1147.

SAUER, U. & DURRE, P. (1992). Possible function of $tRNA_{AGG}$ in regulation of solvent formation in *Clostidium acetobutylicum*. FEMS Microbiol. Lett. 100, 147-154.

SCHAUER, A. *et al.* (1988). Vizualizing gene expression in time and space in the filamentous *Streptomyces coelicolor*. Science 240, 768-772.

SCHAUER, A. *et al.* (1991). Tn*4563* transposition in *Streptomyces coelicolor* and its application to isolation of new morphological mutants. J. Bacteriol. 173, 5060-5067.

SCHUURMAN, R. & KEULEN, W. (1991). Modified protocol for DNA sequence analysis using Sequenase 2.0. BioTechniques 10, 185.

SENO, E.T. & CHATER, K.F. (1983). Glycerol catabolic enzymes and their regulation in wild-type and mutant strains of *Streptomyces coelicolor* A3(2). J. Gen. Micrbiol. 129, 1403-1413.

SHERMAN, D.H. *et al.* (1992). Functional replacement of genes for individual polyketide synthase components in *Streptomyces coelicolor* A3(2) by heterologous genes from a different polyketide pathway. J. Bacteriol. 174, 6184-6190.

SHIINA, T. *et al.* (1991). Sequence of *hrdB*, an essential gene encoding sigma-like transcription factor of *Streptomyces coelicolor* A3(2): homology to principal sigma factors. Gene 107, 145-148.

SHIMKETS, L.J. (1990). Social and developmental biology of the myxobacteria. Microbiol. Rev. 54, 473-501.

SIMONET, J-M. *et al.*, (1992). Genetic instability in *Streptomyces ambofaciens*: inducibility and associated genome plasticity. Gene 115, 49-54.

SMITH, A. (1991). An investigation of the region of DNA required for *Streptomyces pememafaciens* plasmid pSPN1 replication. Masters

dissertation, University of Cape Town, Cape Town, Republic of South Africa.

SMITH, C.P. & CHATER, K.F. (1988). Structure and regulation of controlling sequences for the *Streptomyces coelicolor* glycerol operon. J. Mol. Biol. 204, 569-580.

SOHASKEY, C.D. *et al.* (1992). Construction and application of plasmid- and transposon-based promoter-probe vectors for *Streptomyces* spp. that employ a *Vibrio harveyi* luciferase reporter cassette. J. Bacteriol. 174, 367-376.

SOLIVERI, J. *et al.* (1992). Two promoters for the whiB sporulation gene of *Streptomyces coelicolor* and their activities in relation to development. J. Bacteriol. 174, 6215-6220.

SOLIVERI, J. *et al.* (1993). Functional and evolutionary implications of a survey of various actinomycetes for homologues of two *Streptomyces coelicolor* sporulation genes. J. Gen. Microbiol. 139, 2569-2578.

STEIN, D. & COHEN, S.N. (1989). A cloned regulatory gene of *Streptomyces lividans* can supress the pigment deficiency phenotype of different developmental mutants. J. Bacteriol. 171, 2258-2261.

STRAUCH, E. *et al.* (1991). The stringent response in *Streptomyces coelicolor* A3(2). Mol. Microbiol. 5, 289-298.

STRICKLER, J.E. *et al.* (1992). Two novel *Streptomyces* protein protease inhibitors: Purification, activity, cloning and expression. J. Biol. Chem. 267, 3236-3241. (M80576).

STROHL, W.R. (1992). Compilation and analysis of DNA sequences associated with apparent streptomycete promoters. Nucleic Acids Res. 20, 961-974.

SUGIYAMA, M. *et al.* (1990). The use of the tyrosinase gene to probe promoter sequences for *Escherichia coli*. Plasmid 23, 237-241.

SZABO, G. & VITALIS, S. (1992). Sporulation without aerial mycelium formation on agar medium by *Streptomyces bikiniensis* HH1, an A-factor-deficient mutant. J. Gen. Microbiol. 138, 1887-1892.

TAN, H. & CHATER, K.F. (1993). Two developmentally controlled promoters of *Streptomyces coelicolor* A3(2) that resemble the major class of motility related promoters in other bacteria. J. Bacteriol. 175, 933-940.

TANKA, K. *et al.* (1988). Multiple principle sigma factor homologs in eubacteria: Identification of the *rpoD* box. Science 242, 1040-1042.

TANKA, K. *et al.* (1991). Nucleotide sequence of genes *hrdA*, *hrdC* and *hrdD* from *Streptomyces coelicolor* A3(2) having similarity to *rpoD* genes. Mol. Gen. Genet. 229, 334-340.

THOMPSON, C.J. *et al.* (1987). Characterization of the herbicide-resistance gene *bar* from *Streptomyces hygroscopicus*. EMBO 6, 2519-2523. (X05822).

TSAO, S-W. *et al.* (1985). Identification of a red pigment from *Streptomyces coelicolor* A3(2) as a mixture of prodigiosin derivatives. J. Antibiotics 38, 128-131.

VIROLLE, M-J. & BIBB, M.J. (1988). Cloning, characterization and regulation of an α-amylase gene from *Streptomyces limosus*. Mol. Microbiol. 2, 197-208.

VOEGTLI, M. & HUETTER, R. (1987). Characterization of the hydroxystreptomycin phosphotransferase gene (*sph*) of *Streptomyces glaucescens*: Nucleotide sequence and promoter analysis. Mol. Gen. Genet. 208, 195-203. (X05648).

VON HIPPLE P.H. & BERG O.G. (1986). On the specificity of DNA-protein interactions. Proc. Natl. Accad. Sci. USA 83, 1608-1612.

VUJAKLIJA, D. *et al.* (1991). Identification of an A-factor-dependent promoter in the streptomycin biosynthetic gene cluster of *Streptomyces griseus*. Mol. Gen Genet. 229, 119-128.

VUJAKLIJA, D. *et al.* (1993). Detection of an A-factor-responsive protein that binds to the upstream activation sequence of *strR*, a regulatory gene for streptomycin biosynthesis in *Streptomyces griseus*. J. Bacteriol. 175, 2652-2661.

WANG, J-Y. & SYVANEN, M. (1992). DNA twist as a transcriptional sensor for environmental changes. Mol. Microbiol. 6, 1861-1866.

WARD, J.M. *et al.* (1986). Construction and characterisation of a series of multi-copy promoter-probe plasmid vectors for *Streptomyces* using the aminoglycoside phosphotransferase gene from Tn5 as indicator. Mol. Gen. Genet. 203, 468-478.

WESTPHELING, J. & BRAWNER, M. (1989). Two transcribing activities are involved in expression of the *Streptomyces* galactose operon. J. Bacteriol. 171, 1355-1361.

WESTPHELING, J. *et al.* (1985). RNA polymerase heterogeneity in *Streptomyces coelicolor*. Nature 313, 22-27.

WILDERMUTH, H. (1970). Development and orgaisation of the aerial mycelium in *Streptomyces coelicolor* J. Gen. Microbiol. 60, 43-50.

WILDERMUTH, H. & HOPWOOD, D.A. (1970). Septation during sporulation in *Streptomyces coelicolor*. J. Gen. Microbiol. 60, 51-59.

WILLEY, J. *et al.* (1991). Extracellular complementation of a developmental mutation implicates a small sporulation protein in aerial mycelium formation by *S.coelicolor*. Cell 65, 641-650.

WILLEY, J. *et al.* (1993). Multiple extracellular signals govern the production of a morphogenetic protein involved in aerial mycelium formation by *Streptomyces coelicolor*. Genes and Development 7, 895-903.

WILLIAMS, S.T. *et al.* (1972). Fine stucture of the spore sheath of some *Streptomyces* species. J. Gen. Microbiol. 72, 249-258.

WILLIAMS, S.T. *et al.* (1983). Numerical classification of *Streptomyces* and related genera. J. Gen. Microbiol. 129, 1743-1813.

WRAY, L.V. & FISHER, S.H. (1988). Cloning and nucleotided sequence of the *Streptomyces coelicolor* gene encoding glutamine synthetase. Gene 71, 247-256. (M23172).

YANISCH-PERRON, C. *et al.* (1985). Improved M13 phage cloning vectors and host strains: Nucleotide sequences of the M13mp18 and pUC19 vectors. Gene 33, 103-119.

YOMO, T. & URABE, I. (1994). A frame-specific symmetry of complementary strands of DNA suggests the existence of genes on the antisense strand. J. Mol. Evol. 38, 113-120.

YUDKIN, M.D. (1987). Structure and function in a ***Bacillus subtilis*** sporulation-specific sigma factor: molecular nature of mutations in *spo*IIAC. J. Gen. Microbiol. 133, 475-481.

YU, T-W. & CHEN, C.W. (1993). The unstable *mel*C operon of *Streptomyces antibioticus* is codeleted with a Tn4811 homologous locus. J. Bacteriol. 175. 1847-1852.

ZAKARZEWSKA-CZERWINSKA, J. *et al.* (1994). Inducible transcription from the *dnaA* gene from *Streptomyces lividans* 66. Mol. Gen. Genet. 242, 440-447.