Cytosine DNA methyltransferases and CWG islands in *Pisum sativum*

by

Sriharsa Pradhan

Thesis submitted to the University of Glasgow
for the degree of Doctor of Philosophy

Division of Biochemistry and Molecular Biology
Institute of Biomedical and Lifesciences
University of Glasgow
Glasgow, United Kingdom

October, 1995

Acknowledgements

I would like to express my gratitude to Dr. Roger Adams for accepting me in his laboratory, for continuous discussion and flow of new ideas during the past three years and more recently for his advice and help to write this thesis. I would like to thank Professor J. G. Lindsay and Professor C. A. Fewson for making available the facilities of the Divn. of Biochemistry and Molecular Biology, I.B.L.S for this research. I would also like to thank my auditor Dr. G. I. Jenkins for advice during the course of my Ph.D work. I am very thankful to the following professors, researchers and colleagues: Professor. J. R. Coggins, Dr. Willam J. Caspary, Dr. J. P. Goddard, Dr. S U. Kass, Dr. N Urwin, C.A. Johnson, M. Cummings, H. Lindsay, Tom Carr and other past and present members of the lab for discussion, help and advice during various stages of my work.

I am very grateful to a few of my former teachers and friends who helped me to stay in science they are: Professr H. N. Singh, Professor Sudhir K. Sopory, Professor S. P. Modak, Professor Shipra Guha Mukhrjee and Dr Arun K. Sharma.

I acknowledge financial support from the Association of Commonwealth University and British Council.

Finally, I would like to thank my parents and in-laws especially Mr. N. Jha for tremendous support during my last three years of work. I am indebted to my wife, Lipica, and my daughter Mihika who have been supportive and understanding to my devotion to science and have tolerated my late nights in the laboratory.

Abbreviations

The abbreviation used in this thesis are in agreement with the recomendations of the editors of the Biochemical Journal (Biochem.J (1994) 297, 1-15) with the following additions:

ATP	adenosine triphosphate
BSA	bovine serum albumin
cDNA	complementary DNA
CAT	chloramphenicol acetyltransferase
CTAB	Hexadecyltrimethyl-ammonium bromide
DMSO	dimethyl sulphoxide
DNA	deoxyribonucleic acid
DNase	deoxyribonuclease
DTT	dithiothreitol
EDTA	ethylene diamine tetra-acetic acid
IPTG	isopropyl-β-D-galactopyranoside
MOPS	morpholinopropane-sulphonic acid
mRNA	messenger RNA
PEG	polyethylene glycol
PMSF	phenylmethylsulphonylflouride
PVP	poly-N-vinyl pyrrolidone
RNA	ribonucleic acid
RNase A	ribonuclease A
SDS	sodium dodecyl sulphate

SSC	saline sodium citrate
SSPE	saline sodium phosphate/EDTA
TEMED	N, N, N', N'-tetramethylethylenediamine
X-Gal	5-bromo-4-chloro-3-indolyl-B-D-galactopyranoside

Contents

List of tables and figures

Summary

The aim of this project was to characterise the DNA methyltransferases in plants using *Pisum sativum* as a model system. In order to purify and assess the target specificity of various forms, I applied the partially purified methyltransferase to a MonoQ column, and fractionated two different forms of the DNA methyltransferase. One, a CG methyltransferase, was 140 kDa in size and appeared to be unstable. This protein only methylates CG and CI oligonucleotide duplexes. The other enzyme, a 110 kDa protein, methylates 5′-CAG-3′ and 5′-CTG-3′trinucleotide targets but not 5′-CCG-3′ or 5′-CGG-3′ sequences, as revealed by *in vitro* methylation followed by Maxam-Gilbert chemical sequencing of the methylated strand in the model substrates. I fully purified this second methyltransferase using a CNG hemimethylated DNA affinity chromatography column. The purified protein gave two distinct bands on acrylamide gel, one at 110 and the other 100 kDa. This exciting observation opened up two different avenues in my research project, one to understand more about the targets of CWG methylation and the other to understand more about the enzyme itself.

CsCl purified nuclear DNA was digested with several restriction endonucleases, which were sensitive to cytosine methylation such as *Pst*I, *Pvu*II and *EcoR*II. After end labelling two to three size range of fragments were produced. I gel-purified and cloned these fragments, which arose from the unmethylated region of the genome. The sequence analysis of fifty such *Pst*I tiny fragment (PTF) clones shows that these DNAs are rich in unmethylated CWGs and are deficient in CG dinucleotides indicating the

occurrence of CWG islands in pea. Occasional unmethylated CGsare found in CWG islands

Analysis of the 5´ region of 40 published plant gene sequences from the GenEMBL database does not shows any particular trend between CG and CWG content of the plant promoters. It is estimated that 5% of the plant genes are associated with CWG islands, in contrast to most of them being associated with CG islands. To address the biological role of CG vs CWG methylation, I performed transient gene expression studies in collaboration with Dr. Nigel Urwin. Plasmid constructs containing promoters of either *CAMV35S* or *rbcs2* fused to a CAT reporter gene were *in vitro* methylated and electroporated in to bean protoplasts. Methylation of CWG sequences in the *rbcs2* promoter leads to a dramatic inhibition of CAT gene expression suggesting the possible involvement of CNG methyltransferase (MTase) in gene regulation.

In order to understand the structure, function and pattern of methyltransferase gene expression during plant development, I have isolated a series of overlapping clones from pea cDNA libraries. The assembled nucleic acid sequence is 5 kb. The deduced amino acid sequence has an open reading frame of 1560 amino acids with a predicted protein of 180 kDa. Like other higher eukaryotic DNA MTase this protein has an N terminal domain fused with a catalytic domain with a short linker which, in case of pea, is RKKKG. The pea enzyme has eight of the ten conserved motifs found in prokaryotic enzymes. The C terminal (catalytic domain) of the protein is homologous to both vertebrate and plant enzymes. However, the N-terminal has poor homology with the vertebrate enzymes. The unique feature of the N terminal region is the presence of several, putative nuclear localisation signals that may also be the sites of

action for proteases. The presence of a Zn binding domain on this protein makes it different from the *Arabidopsis* protein. An uncharacterised acidic domain is also present in the protein. By Southern analysis I concluded that the MTase gene is a single copy gene and Northern analysis shows it to be mostly expressed in the meristematic tissue. Amino acid sequence alignment with the N terminal sequence of the CWG MTase suggests that this protein could be a processed product of the 180 kDa protein.

CHAPTER ONE

Introduction

Some of the cytosine and adenine bases in the DNA of living organisms are methylated. This modification involves the transfer of a methyl group from S-adenosyl methionine to the nucleotide base, resulting in formation of N-6-methyladenine or 5-methylcytosine (sometimes N-4-methylcytosine). This reaction is catalysed by two distinct classes of enzymes: adenine and cytosine methyltransferases. The resultant 5-methylcytosine (5-mC) is unstable and can be converted to thymine by oxidative deamination. Thus, methylation of cytosine generates hot spots for mutation. However, for the survival of the organism it is important that the benefits of methylation should outweigh the mutational burden. Methylation of DNA is believed to be a key player in gene regulation and expression. In this chapter various aspects of DNA methylation and the DNA methyltransferase enzymes are discussed.

1.1 Various methods to study DNA methylation.

Much before the discovery of the structure of DNA, Rollin Hotchkiss separated 5-methylcytosine from calf thymus DNA using paper chromatography techniques (Hotchkiss, 1948). Since then, with the advance of science and molecular biological techniques, it has been possible to analyse and characterise the modified bases in DNA. Some of these methods are described in the following sections.

1.1.1 Sequence unspecific methods

These methods have been used successfully to estimate the quantity of methylated bases in DNA. The key steps are either acid or enzymatic hydrolysis of DNA followed by chromatographic separation and quantification of individual bases (Eick *et al.*, 1983; Ford *et al.*, 1980). Pollack *et al.* (1984) reported an improved method, whereby nicks were introduced to DNA by means of DNaseI treatment and the nucleotide 3′ to the nick was labelled using a ^{32}P-labelled 5′ triphosphate. On hydrolysis to the 3′ monophosphate, and separation by TLC, this modified nearest neighbour analysis indicated the proportion of the four CpN nucleotides that are methylated. Subsequently, mass spectroscopy (Razin and Ceder, 1977) and amplified ELISA coupled with photoacoustic spectroscopy (Achwal *et al.*, 1984) were used to detect picomole quantity of methylated cytosine.

1.1.2 Sequence specific methods

With the discovery of isoschizomers it was possible to study the differential methylation pattern throughout the genome and at specific genes. Several of these enzymes are available commercially e.g. *HpaII*, an enzyme that cleaves unmethylated 5′-CCGG-3′ sequences, while the isoschizomer *MspI* does not cleave if the internal cytosine is methylated (reviewed by Adams and Burdon, 1985; Saluz and Jost, 1993). An improvement on this method was done by employing a PCR assay. Undigested DNA was amplified after digestion with methylation sensitive restriction endonuclease. Unmethylated DNA gets digested hence no PCR product was formed (Singer-Sam *et al.*, 1990). However, these methods cannot detect methylated bases other than those present in the recognition sequences.

Church and Gilbert (1984) employed a genomic sequencing protocol to detect the presence of methylated cytosine in the genomic DNA. The DNA was subjected to chemical sequencing chemistry (Maxam and Gilbert, 1980) and separated on a sequencing gel. The separated DNA is transferred to a nylon membrane that is hybridised to radiolabelled complementary DNA probes. Methylated cytosine appears as a gap in the C-reaction lane, as hydrazine does not react with 5-methylcytosine during the C specific modification reaction. Several improvements have been made on this method. The introduction of ligation mediated PCR (LMPCR) by Pfeifer *et al.* (1989) enhanced the sensitivity of detection. More recently a positive display of 5-methylcytosine using the modified genomic sequencing technique has been introduced (Frommer *et al.*, 1993; Clark *et al.*, 1995). In this technique bisulphite is used to deaminate cytosine (but not 5-methylcytosine) to uracil. Strand-specific primers to bisulphite deaminated DNA are used to amplify the target DNA. In the amplified products, all the uracil and thymine residues are detected as thymine whereas, 5-methylcytosine remain as cytosine. The amplified PCR fragments can be cloned and sequenced to obtain the methylation status of each cytosine.

1.1.3 The use of methylase inhibitors to study DNA methylation *in vivo*

Inhibition of DNA methylation in the organism is often used to study the role of demethylation on the expression of a particular gene. The most widely used chemical compounds are the base analogues 5-azacytidine (5-AZC) and 5-azadeoxycytidine. On incorporation into DNA, 5-azacytosine brings about an inhibition of DNA methylation (Jones and Taylor, 1980). The triazine ring is not able to accept the methyl group and in an *in vitro* assay it has been shown that DNA methyltransferase binds irreversibly to

the DNA containing 5-azacytosine (Adams *et al.*, 1984). This phenomenon was also observed in plants. Treatment with 5-azacytidine results in the activation of previously inactive transgenes in plants (Weber *et al.*, 1990). In cultured cells demethylation of cellular DNA is also observed after 5-azacytidine treatment. However, these analogues are toxic to cells and could lead to a marked change in the metabolism and development of the cell (Jones, 1984). Recently it has been shown that 5-azadeoxycytidine toxicity in mammalian cells is mediated through covalent trapping of DNA Mtase (Juttermann *et al.*, 1994).

1.2 Methylated and unmethylated fraction of DNA

1.2.1 Distribution of methylated bases

In the most studied prokaryote *E. coli* K12, the minor modified nucleobases are 5-methylcytosine, N6-methyladenine and N4-methylcytosine (Dunn and Smith, 1955; Doskocil and Sormova, 1965). However, in the vertebrate genome the only modified base is 5-methylcytosine (5mC). In mammals ~5% of the cytosines are modified to 5mC (Ehrlich *et al.*, 1982; Gama-Sosa *et al.*, 1983a,b) in contrast to plants, where a third of the total cytosines are methylated (Adams and Burdon, 1985). Evidence also has been obtained for the presence of other minor modified bases in the plant genome. HPLC analysis of heterotropic cell line of sycamore indicated that amyloplast DNA contains a range of other different modified bases in addition to 5% 5mC (Ngernprasirtsiri *et al.*, 1989).

Among the well studied organisms such as *Drosophila melanogaster* (Urieli-Shoval *et al.*, 1981; Pollack *et al.*, 1984) and *Saccharomyces cerevisiae* (Proffitt *et al.*, 1984) methylated bases have not yet been detected.

There seems to be a correlation between genome size and DNA methylation. It has been proposed that methylation inactivates the nontranscribed DNA. The satellite DNA of plants and vertebrates are enriched with 5mC. As high as 50% of all 5mC are found in the satellite DNA of the mouse genome (Millar *et al.*, 1974). It is believed that methylation was first associated with satellite DNA and, with evolution, has become a component of previouslyunmethylated DNA in higher plants and animals. This evolutionary process may have allowed the involvement of DNA methylation with various biological processes.

1.2.2 Sequence specificity of cytosine methylation

In vertebrate DNA symmetrical CpG dinucleotides are methylated and, more recently, the occurrence of CpNpG methylation has been reported (Clark *et al.*, 1995). In the plant genome methylation at the 5-carbon of the first C in both symmetrical CG and CNG exists. (Gruenbaum *et al.*, 1981). In transgenic plants and fungi asymmetric methylation has been reported (Meyer *et al.*, 1994; Selker *et al.*, 1993).

1.2.3 The CpG and CpNpG Islands

A fraction of vertebrate DNA has an elevated CpG frequency and G+C rich base composition. This fraction contains the so-called CG islands (Bird, 1986). The ratio of dinucleotides CpG/GpC in the islands is about 1 whereas in the rest of the

genome this ratio is down to 0.2. The possible mechanism of CpG depletion from the bulk genome is as a result of deamination of 5mC to thymine. Thus, a 5mCpG is converted to TpG, which is maintained through successive round of DNA replication. In the case of deamination of cytosine to uracil, the mismatch is corrected by the action of uracil glycosylase and excision repair mechanism of the cell (Lindahl, 1974). Both higher plants and vertebrates have CpG islands (Bird, 1986; Antequera and Bird, 1988). These islands are present in the 5´ region of most of the housekeeping genes whereas genes known to be expressed in a tissue-specific manner are usually not associated with CpG islands.

As mentioned earlier, CpG islands are usually methylation free in all tissue with one exception: the inactive mammalian X chromosome. One of the two X chromosomes in eutherian females is methylated during the early stage of development, and is cytologically distinguished as the Barr body. Its characteristic features are heterochromatisation, hypermethylation and lack of acetylated histones (Tribioli *et al.*, 1992; Jeppesen and Turner, 1993).

How do these islands keep themselves methylation free in spite of having many DNA methylase target sites? CpG islands introduced into mice remain methylation free irrespective of their expression (Kølsto *et al* ., 1986; Shemer *et al.*, 1990). It is speculated that an active demethylation event could be responsible for the methylation free status of the island DNA. This hypothesis is based on the work of Frank *et al.* (1991) who demonstrated that an *in vitro* methylated *aprt* (adeninephosphoribosyl-transferase) gene becomes demethylated in transgenic mice. One of the possible mechanism could be excision and repair of 5mC to C (Razin *et al* ., 1986). The other

could be an active replication dependent demethylase activity as demonstrated by Jost (1993).

In plants the presence of both CpG and CpNpG islands have been postulated (Messeguer *et al.*, 1991). However the distribution and function of plant islands are not well studied.

1.3 DNA methylation and gene expression

The frequency of the occurrence of 5mC and the conservation of methylated bases during DNA replication has led to the suggestion that interference of binding of regulatory proteins to the transcriptionally active DNA may be affected by DNA methylation. The role of methylation and gene expression in plants (review by Finnegan *et al.*, 1993) and animals (review by Razin and Cedar, 1991) is well documented. Two models for the involvement of DNA methylation in gene expression are hypothesised, and are described below in detail.

1.3.1 Methylated DNA prevents transcription factor binding

One of the obvious effects of addition of methyl groups to DNA is the structural change in the sequence of the nucleic acid, which in turn could affect the binding of a number of transcription factors. The net result could be inhibition of transcription. Several transcription factors of this class have been reported (reviewed by Tate and Bird, 1993) The prominent candidates are the cAMP responsive element binding protein (Iguchi-Ariga and Schaffner, 1989), the activating protein 2 (Comb and Goodman, 1990), c-Myc/Myn (Prendergast *et al.*, 1991), NF-kB (Bednarick *et al.*,

1991) in animal cells. In maize TnpA (Gierl *et al.*, 1988) cannot bind to methylated *Spm* elements. One of the tobacco nuclear proteins, CG-1 also does not bind to the methylated CACGTG motif in the chalcone synthatase promoter of *Antirrhinum* (Staiger *et al.*, 1989). However, most of these binding assays were performed *in vitro*. In case of the rat tyrosine aminotransferase (TAT) gene, Becker *et al.* (1987) demonstrated the binding of ubiquitous factors to the TAT promoter *in vivo* . This promoter was unmethylated in expressing cells. In the nonexpressing cells methylation of the promoter was observed. These cells also did not have the characteristic DNA footprint pattern. Demethylation of DNA using 5-AZC in the non expressing cells did not result in binding of the factors. Some of the transcription factors such as SP1, are insensitive to CpG methylation in their target sites (Hollar *et al.*, 1988), but transcription from their promoter can be inhibited by DNA methylation (Ben Hattar *et al.*, 1989; Bryans *et al.*, 1992). Thus the inability of the transcription factors to bind to the methylated DNA is not the only mechanism for methylation-mediated gene repression.

1.3.2 Specific proteins bind to methylated DNA

The indirect mechanism of DNA methylation and transcriptional inhibition is mediated by the inaccessibility of transcription factors to the methylated promoter. This is a two step process. Soon after the DNA methylation, binding of one of a group of methylated DNA binding proteins (MDBP) takes place to the DNA which prevents the binding of the essential regulatory factors. This could induce formation of inactive chromatin. These MDBPs have been isolated and characterised in mouse and pea. The proteins charactrised in mouse are called methylated cytosine binding protein: MeCP1 and MeCP2 (Meehan *et al.*, 1992). The pea methylated DNA binding protein, MDBP is

isolated and partially characterised by Ehrlich (1993). MeCP1 is known to repress methylated genes *in vitro*, as well as *in vivo,* and it requires at least 12 closely linked methylated CpGs (Boyes and Bird, 1991; Boyes and Bird, 1992), whereas, MeCP2 requires only one methyl CpG to bind and is associated with pericentromeric heterochromatin (Lewis *et al.*, 1992). MeCP2 does not inhibit transcription *in vitro* from methylated genes (Meehan *et al.*, 1992). Taking into account the characteristics of MeCP1 and 2, Bird and co-workers have suggested a possible model, whereby MeCP1 competes with transcription factors to bind to methylated DNA and guides the DNA to form a heterochromatic structure which, in turn, is maintained by MeCP2.

1.3.3 DNA methylation and chromatin structure

In the nucleus of the cell most of the DNA is packaged into chromatin. It has been shown that the transcriptionally active chromatin is more accessible to DNaseI and micrococcal nucleases. This indicates that the structure of transcriptionally active chromatin is in some way more open thus allowing the nuclease to act upon. Razin and Cedar (1977) demonstrated that after limited micrococcal nuclease digestion the unreleased fraction of the DNA contains 75% of 5mC. Subsequent work by Adams *et al.* (1984) indicates that linker DNA is more methylated than the core DNA. Keshet *et al.* (1986) have shown that DNA methylation has an effect on chromatin formation. They used a methylated construct which was integrated into the genome of L cells and observed a DNaseI insensitive chromatin structure, characteristic of inactive chromatin. Nuclease insensitivity is observed in several other methylated constructs (Antequera *et al.*, 1989; Levine *et al.*, 1991; Sasaki *et al.*, 1992). There was more evidence from the CpG island methylation in inactive X chromosomes in mammals, where nuclease

resistant for the methylated genes has been reported (Wolf *et al.*, 1984; Hansen *et al.*, 1988).

Microinjection and transfection experiments with methylated constructs in animals and plant cells have demonstrated that methylation leads to gene repression. In rat tk^- cells, reporter gene expression drops sharply 8 hours post-injection. *In vitro* assembled chromatin for the same gene construct had no lag period before repression. This represents the indirect mechanism of gene repression through chromatin structure. Further evidence came from methylated γ-globin genes. The authors (Murray and Grosveld, 1987) showed 5mC are required on the construct for the inactivation phenomenon to occur. This led to the belief that the factors and proteins associated with chromatin could be playing a major role.

Histone proteins are a major component of chromatin . Hence, there could be some influence of histones on methylated DNA templates. Felsenfeld *et al.* (1983) reported that the methylated CpG polymer has two-fold greater affinity to bind histones as compared with unmethylated polymers. Later studies on reporter gene constructs indicated no significant difference between chromatin formation on either methylated or unmethylated template. This was further confirmed by electron microscopy and micrococcal nuclease digestion (Buschhausen *et al.*, 1987). However, the recent focus is on histone H1, long suspected to be involved with transcriptionally inactive chromatin. Higurashi and Cole (1991) demonstrated that *in vitro* assembled histone H1-DNA complexes could be digested by restriction enzymes but not with *MspI*. They also showed that hisone H1 has the same affinity for methylated and unmethylated DNA. Thus, the inability of the *Msp*I restriction endonucleases to digest the methylated DNA-histone H1 complex could be the result of changes in DNA conformation. In contrast, Levine *et al.* (1993) did observed preferential binding of histone H1 to the

DNA. In a series of *in vitro* transcription assays Johnson *et al.* (1995) demonstrated the methylated template could be transcriptionally inactivated with much lower amounts of histone H1, when compared with the same amount of unmethylated DNA. And among the histone variants (Santoro *et al.*, 1995) H1c showed greater preferential inhibition of DNA methyltransferase. However, all these *in vitro* assay experiments have to be interpreted with great care since they do not mimic the exact intracellular conditions.

1.3.4 DNA methylation may spread inactive chromatin

Though the exact mechanism behind DNA methylation and inactive chromatin formation is still unknown, there is growing speculation about the involvement of a number of factors responsible for the spreading of inactive chromatin. Kass *et al.* (1993) used a 'patch methylation' technique to generate a series of methylated reporter gene constructs in predetermined sequences. Upon transfection it was observed that transcriptional inhibition had occurred independently of the position of the methylated patch. Both unmethylated and methylated regions are protected from limited *Msp*I digestion. This shows that the methylated region could spread inactive chromatin to unmethylated parts of the DNA.

1.4 Other biological significance of DNA methylation

Apart from participation in gene regulation and expression, DNA methylation is also involved in developmental regulation (Antequera *et al.*, 1989), genomic imprinting (Reik *et al.*, 1987; Swain *et al.*, 1987) and X chromosome inactivation. Aberrant DNA methylation may be mutagenic in mammals (Cooper and Youssoufian, 1988; Rideout

et al., 1990; Shen *et al.*, 1992) and often plays a role in development of certain human diseases (Oberle *et al.*, 1992)

1.5 DNA methyltransferase

As described in the beginning of this chapter, DNA methyltransferases transfer methyl groups from S-adenosyl-L-methionine (AdoMet) to certain cytosine and adenine nucleobases in the DNA. In prokaryotes, most of the methylases are associated with host restriction-modification systems. In the past few years a number of such prokaryotic and eukaryotic enzymes have been studied.

1.5.1 Bacterial DNA methyltransferases

Bacterial restriction endonucleases are classified as type I, II and III based on their structure, cofactor requirement, nature of the recognition site and the spatial relationship between recognition and catalytic domains. In the type II and IIS-system the endonuclease and corresponding methylase system are conferred by two independent polypeptides, both acting independently of the other. They require Mg^{2+} for the endonuclease action and AdoMet for modification. These are plasmid encoded enzymes and the genes are often adjacent to each other. In the case of Type I and III both the activities resides in the same enzyme complex. They require Mg^{2+} and ATP and are stimulated by AdoMet. Different type II MTases have the capacity to methylate cytosine and adenine residues on DNA to produce 5-methylcytosine (5mC), N-4-methylcytosine (N4mC) or N-6-methyladenine (N6mA). The type II enzymes are active as monomers. They methylate the target duplex DNA in two stages. In the first step the enzyme converts an unmethylated site to a hemimethylated one. The second

step generates a fully methylated target sequence (Rubin and Modrich, 1977; Herman and Modrich, 1982). S-adenosyl-L-methionine (AdoMet) acts as a cofactor for the donation of methyl group to the nucleobase.

Type I restriction-modification systems are found in *E. coli* B, *E. coli* K12 and different salmonella species. This type is coded by three contiguous genes known as *hsd*R, *hsd*M, and *hsd*S (host specificity for DNA Restriction, Modification, and Specificity). All the three gene products are required for restriction, and the products of genes *hsd*M and *hsd*S are required for modification (Yuan and Hamilton, 1984). The *hsd*S gene product is required to recognise the specific sequence on DNA which is modified by the *hsd*M gene product.

Both *Eco*B and *Eco*K methylase methylates adenine residues. Restriction takes place several thousand bases away from the recognition and methylation site. The complete restriction-modification enzyme binds five or more molecules of AdoMet, and is converted into an active form that can bind to any DNA molecule. The enzyme may than slide on the DNA to locate the recognition sequence

Type III methyltransferases are very similar to type I except that the *hsd*M and *hsd*S gene products are combined into a single subunit as they are in type II methylases. ATP is probably not an absolute requirement, though it does stimulate the activity of the enzyme slightly. Both the methyltransferase and nuclease activity require the presence of Mg^{2+} ions. *Eco*P1 and *Eco*P15 restriction enzymes are encoded by almost identical sequences carried by P1 phage and the P15B plasmid respectively. Each is encoded by two genes: *res*, required for restriction only, and *mod*, which is required for restriction and modification. These two methyltransferases are transcribed

from two different promoter and are different at the N-terminus (for a detailed review on DNA methylation in prokaryotes refer Noyer-Weidner and Trautner, 1993).

1.5.2 Conserved motifs in m5C methyltransferases

The DNA and predicted protein sequences are available for at least fifty m5C methyltransferase (m5C MTase) in the GenEMBL data bank. Comparative analysis has shown that these protein sequences are an ordered set of conserved sequence motifs with non-conserved sequences between motifs (Som *et al.*, 1987; Lauster *et al.*, 1989; Posfai *et al.*, 1989; Klimasauskas *et al.*, 1989., Kumar *et al.*, 1994). Depending on the criteria used to define the conserved motifs (conserved blocks), ten conserved blocks have been identified (Som *et al.*, 1987; Lauster *et al.*, 1989; Posfai *et al.*, 1989; Kumar *et al.*, 1994). These blocks are also present in the catalytic domain of eukaryotic MTases. Out of the ten conserved blocks six motifs (I, II, VI, VIII, IX and X) are highly homologous. The largest nonconserved or variable region lies between VIII and IX. All the motifs are arranged in a sequential order (Fig. 1.1). The highly conserved motifs and common architectural organisation indicates that the function could be very similar. Similarities include the mechanisms of recognition of a specific DNA sequence and the catalysis of methyl group transfer

To date, two conserved motifs have been assigned functional roles based on the common chemistry of these enzymes. Motif I (FxGxG) is shared commonly with other AdoMet dependant MTases, and is known as the cofactor S-adenosyl-L-methionine binding domain. The other conserved Motif IV with Pro-Cys dipeptide, known to be a part of the catalytic site. This contains the nucleophilic thiol as proposed by Wu and Santi (1987). Mutagenesis of Cys in the dipeptide leads to the loss of catalytic function

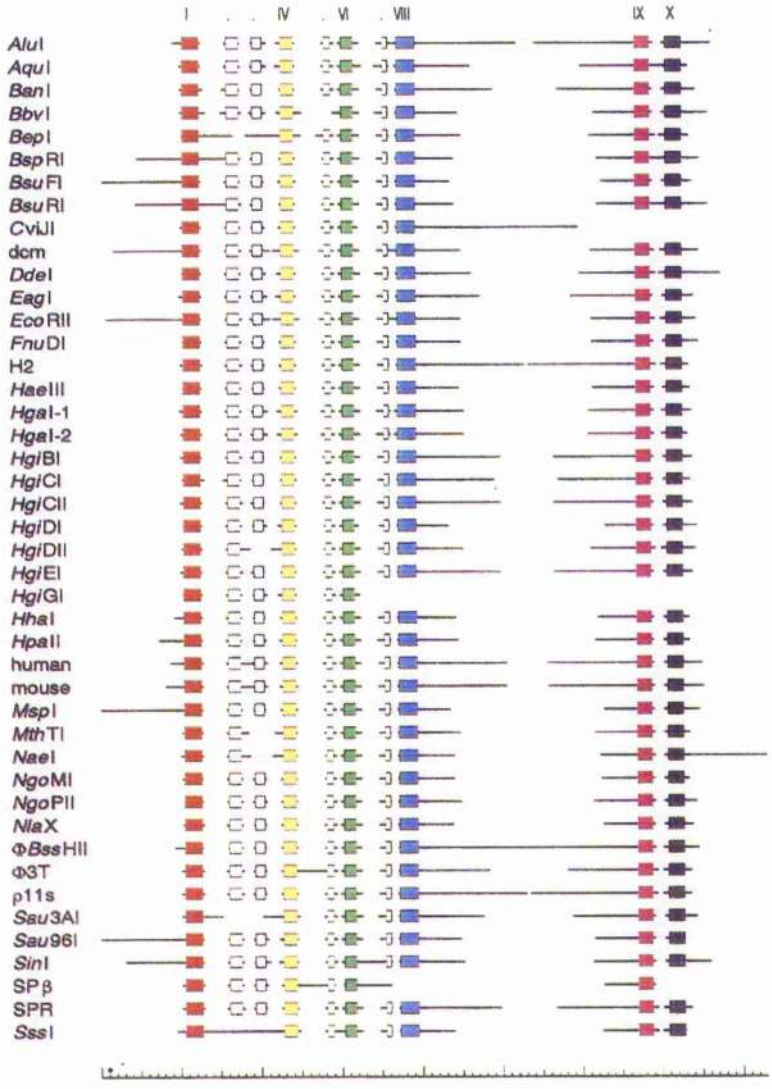


Figure 1.1
Schematic showing the alignment of 45 m5C-MTases. The motifs are coloured: red - motif I, F-G-G; yellow - motif IV, PC; green - motif VI, ENV; cyan - motif VIII, Q-R-R; magenta - motif IX, RE; dark blue - motif X, GN.

Kumar et al., (1994) NAR, 22. pp4

in M. *Eco*RII (Wyszynski *et al.*, 1993), M. *Hha*I (Mi and Roberts, 1993), M. *Hae*III (Chen *et al.*, 1993) and dcm (Hanck *et al.*, 1993). The involvement of Cys in catalysis was also confirmed by trapping the enzyme with 5-fluorodeoxycytosine substituted suicide substrate and subsequent analysis of the trapped intermediate (Wyszynski *et al.*, 1993; Wyszynski *et al.*, 1992). Recently Roberts' group (Klimasauskas *et al.*, 1994) has managed to obtain co-crystals of the M.*Hha*I-DNA complex and demonstrated the covalent bond between Cys81 and carbon 6 of cytosine.

The variable region between motif VIII and IX is known to be the TRD (Target recognising domain). In the multi-specific enzymes, a point mutation in the variable region is capable of abolishing one target specificity leaving the others intact. By mapping the mutation and determining the target specificity, Trautner's group has managed to swap TRDs in the variable region. For a thorough review, see Noyer-Weidner and Trautner (1993). These experiments suggest not only that the variable region determines the sequence specificity, but also the choice of the specific base that is to be methylated (Klimasauskas *et al.*, 1991; Mi and Roberts, 1992).

1.5.3 Mechanism of methyl group transfer in m5C methyltransferases

*Hha*I MTase recognises 5´-GCGC-3´ tetranucleotide sequences and transfers the methyl group from AdoMet to the inner cytosine. The kinetic and the catalytic mechanisms have been studied in detail by Wu and Santi (1987), later modified by Chen *et al.* (1993). The reaction is initiated by covalent attack of cysteine thiolate (Chen *et al.*, 1991) on C6 of the substrate cytosine (Fig. 1.2, step 1). To avoid the formation of a high energy carbanion, the thiolate attack on C6 is accompanied by protonation at N3 by the enzyme driven acid (Chen *et al.*, 1993; Erlanson *et al.*, 1993).

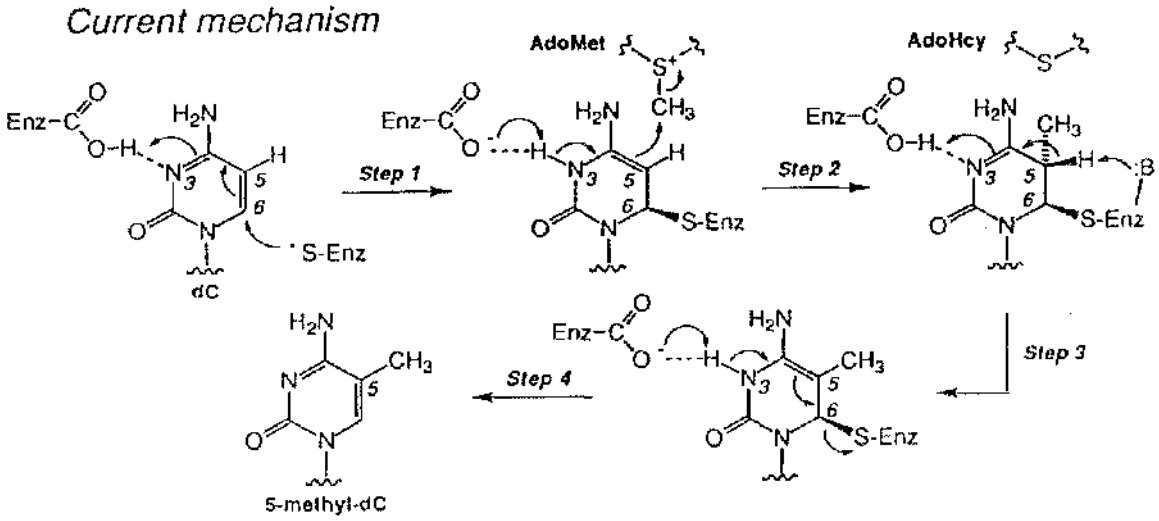


Figure 1.2 Current mechanism of action of DNA (cytosine-5) methyltransferase

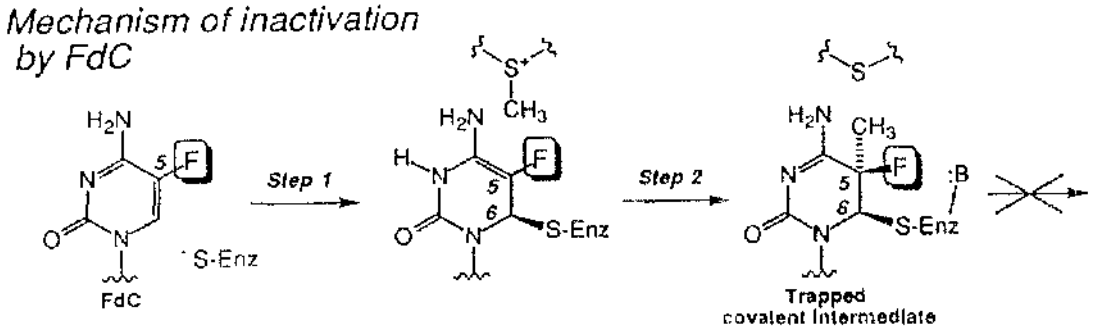


Figure 1.3 Mechanism of inactivation of DNA (cytosine-5) methyltransferase by FdC

The resulting structure is termed an enamine and this attacks the methyl group of AdoMet resulting in its transfer to C5 (Fig. 1.2, step 2). Abstraction of the proton at C5 again yields an enamine (Fig.1.2, step 3), which undergoes conjugate elimination to yield 5methyl-dC. For all these events to occur a substantial structural distortion of the DNA is a prerequisite. This is because the trajectory of the attack of the thiolate on C6 is blocked by the DNA backbone, and the methyl group delivery will also be hindered by neighbouring bases. This mechanism also led to speculation of the need for the strand separation event to occur in order to permit access of the enzyme derived acid to N3.

1.5.4 Enzyme and suicide substrate

Based on the covalent catalysis mechanism, Santi and co-worker designed the inhibitor 5-fluoro-2'-deoxycytidine (FdC), in which the 5H of cytosine is replaced by a fluorine (Osterman *et al.*,1988). This inhibitor blocks the progression through step 3 of the mechanism (Fig.1.3), presumably because the abstraction of F^{+} under physiological conditions is impossible. The FdC compound has been shown to form covalent complexes *in vitro* with cytosine methyltransferase from bacterial and mammalian origin (Osterman et *al.*, 1988; Hank *et al.*, 1993; Friedman and Ansari, 1992; Smith *et al.*, 1992). Substituting a fluorine in place of a hydrogen does give a similar van der Walls radius, so that the trapped intermediate resembles the normal intermediate.

Another mechanism-based inhibitor of m5C MTase is formed by the incorporation of pyrimidinone (2-pyrimidinone-1-β-D-2′-deoxyribofuranoside) into the DNA duplex containing the recognition site for the enzyme. Covalent complexes are formed between M.*Msp*I, M.*Hga*I-2 and pyrimidinone containing substrate (Taylor *et*

al., 1993; Ford *et al.*, 1993; Baldwin *et al.*, 1994), even if the active site cysteine is modified to serine. The amino group of cytosine normally at position 4 reduces the reactivity of C6 position, making it susceptible to thiolate attack. Replacing the four amino group with a hydrogen, results in the enhancement of C6 reactivity. The lack of the 4-amino group of cytosine in pyrimidinone-DNA reduces the number of inter strand hydrogen bonds from three to two, and this alone may stimulate the enzyme-DNA covalent formation.

1.5.5 Structural studies of M.*Hha*I-DNA covalent ternary complex

Klimasauskas *et al.* (1994) crystallised M.*Hha*I and a 13 mer oligonucleotide containing its recognition sequence, with 5-fluorocytosine at the target site for methylation (DNA-M.*Hha*I). The two form a 1:1 complex, with the enzyme linked with a thioether bond to the FdC residue on the strand. The product of the reaction, S-adenosyl-L-homocysteine, remains bound to the active site. The DNA lies in a large cleft formed at the junction of the two domains of the protein. The smaller domain makes sequence specific contacts in the major groove and the large domain interacts mainly with target cytosine in the minor groove. In the native DNA duplex the carbon-5 of cytosine is in the major groove and is considered to be inaccessible. In order to access the cytosine the distortion that occurs is very spectacular (Fig. 1.4a). The enzyme MTase clearly extends the target cytosine out of the helix into the catalytic site, without disturbing the rest of the DNA helix. In the process, the protein undergoes a major conformational change upon binding to the DNA; the tip of the catalytic loop containing motif IV moves nearly 25Å towards the cleft into the minor groove of the DNA, at the same time pulling Cys81 to the region that will become an active site (Fig. 1.4). The gap left by the evicted base is filled by Gln237 from one of the recognition

Figure 1.4
The structure of M.*Hha*I in a ternary complexed with a substrate duplex DNA oligonucleotide and the end product of the reaction AdoHcy. The motifs are shown in colour: AdoHcy is in white, the DNA bases are in orange, the deoxyribose is in purple and the phosphates are green.

Roberts R.J. (1993), Lex Prix Nobel, p156

loops in the small domain. This maintains the base stacking and provides hydrogen bonding to the orphan guanine. Ser87 interacts and stabilises the Gln. Once the target cytosine is buried deep in the cleft, it is held in place by four residues Phe79, Cys81, Arg165 and Gln119 (Fig1.5). Thus, residues from motif IV, VI, and VIII take active parts in these processes.

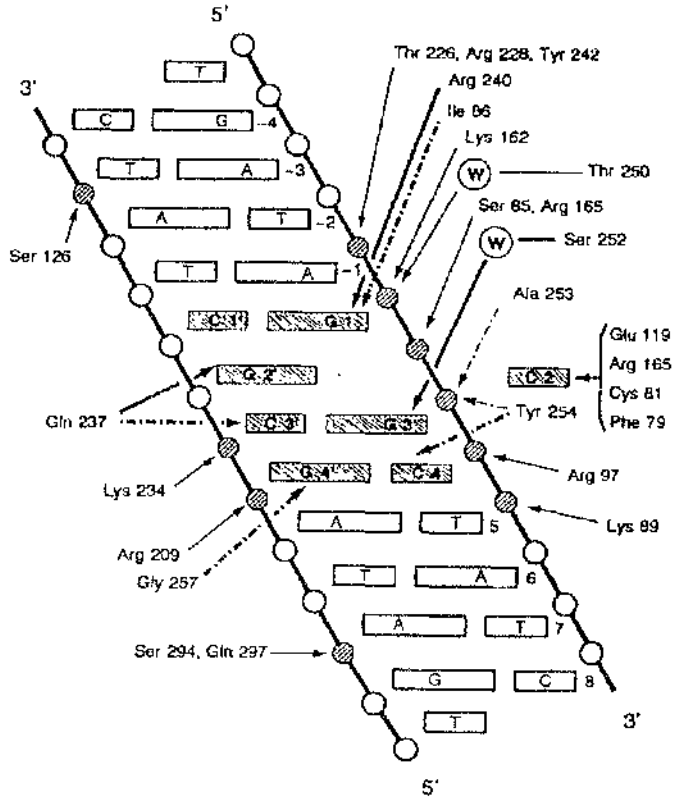


Figure 1.5 Diagram showing the specific base and phosphate contacts between M.*Hha*I and DNA

The DNA is represented as a cylindrical projection. The recognition bases and contacted phosphates are shaded. Base contacts are shown as thick lines, phosphate contacts are thin lines, and the contacts with main chain atoms with dashed lines. The symbol (W) indicates the water - mediated contacts. Klimasauskas *et al.*, (1994) *Cell* **76**, pp 365.

The AdoMet binds to the large domain of M-*Hha*I in a pocket facing the cleft. Residues from motif I to V and X take part in this process. Motif I forms a tight loop in the first turn of a $\beta_1\alpha_A\beta_2$ structural unit (Cheng *et al.*, 1993). The conserved Gly in the FxGxG allows the tight turn to occur in the structure. The conserved Phe interacts with the adenosyl moiety of the cofactor, with the aromatic ring perpendicular to the plane of the purine ring. The interaction of motif X with AdoMet or AdoHcy is mediated through Asn304 of the conserved Gly-Asn dipeptide.

Most of the prokaryotic enzymes recognise a particular DNA sequence. This recognition of the target sequence resides in the variable region between motif VIII and IX. Work from Trautner's group on multispecific m5C-MTases indicates a length of 40 amino acids each; and individuals TRDs do not overlap and are separated by a single amino acid (Noyer-Weidner and Trautner., 1993; Balganesh *et al.*, 1987). In the case of M-*Hha*I, DNA binds in the cleft between the domain with its major groove facing the variable region in the small domain, and than minor groove faces towards the catalytic site in the large domain. Three loops contact the target sequence:one catalytic loop and two Gly-rich recognition loops (I: aa 233-240, GKGGQGER; II :aa 250-257, TLSAYGGG) from the small domain. Loop II interacts extensively with the backbone of the DNA strand containing the target cytosine, and contacts three bases in the 3′ half of the double-stranded recognition sequence. When DNA is bound in the cleft, the Thr-Leu dipeptide (aa 250-251) from the TRD consensus appears to be important in positioning the recognition loop II relative to the target cytosine. The variable region within the methyltransferases recognising internal C of CCNGG (M.*Scr*FI, M.*Sso*II and M.*Dsa*V) and CCWGG (M.*Nla*X, M.*Eco*RII and Dcm) can be aligned. The two segments within that region also have similarity with the region of M.*Hha*I that contacts DNA bases (Gopal *et al.*, 1994).

1.5.6 Eukaryotic DNA methyltransferase

The eukaryotic DNA methyltransferases have been purified from a number of animal and plant sources. In proliferating mouse tissue a protein of 190 kDa is found (Turnbull and Adams, 1976; Bestor and Ingram, 1985; Pfeifer and Drahovsky, 1986). In plants, several sizes of methyltransferases are reported. In wheat proteins of 55 and 35 kDa are believed to be associated with the methyltransferase activity (Theiss *et al.*, 1987). A similar molecular weight protein (54 kDa) is also found in rice cell culture (Giordano *et al.*, 1991). However, in pea a 160 kDa protein is identified as a DNA methyltransferase (Yesufu *et al.*, (1990). In the higher plant genome two different sequences are methylated. It has been shown in pea that such processes are carried out by at least two different species of enzymes: one methylating the CG and the other CNG sequences (Pradhan and Adams, 1995).

The first cDNA for DNA MTase to be cloned was from murine erythroleukemia cells (Bestor *et al.*, 1988). The sequences of the cDNA for human (Yen *et al.*, 1992) and *Arabidopsis* (Finnegan and Dennis, 1993) MTases are also available in the database. The translated proteins in all the above eukaryotic sequences contain eight out of the ten conserved motifs. The most studied MTase is the murine enzyme. This has a 1000 amino-acid long N terminal fused with a 500 amino-acid long C terminal. It is hypothesised that mammalian MTase is a fusion between genes for a prokaryotic-like methyltransferase and an unrelated DNA binding protein (Bestor, 1990). The C terminal domain has significant homology to bacterial type II DNA C5-methyltransferases. Both the domains fold independently and are joined by a run of 13 alternating glycyl and lysyl residues (Fig. 1.6). In the centre of the N terminal region there is a putative zinc binding domain (Bestor, 1992). The N-terminal domain is involved in distinguishing between methylated and hemimethylated DNA.

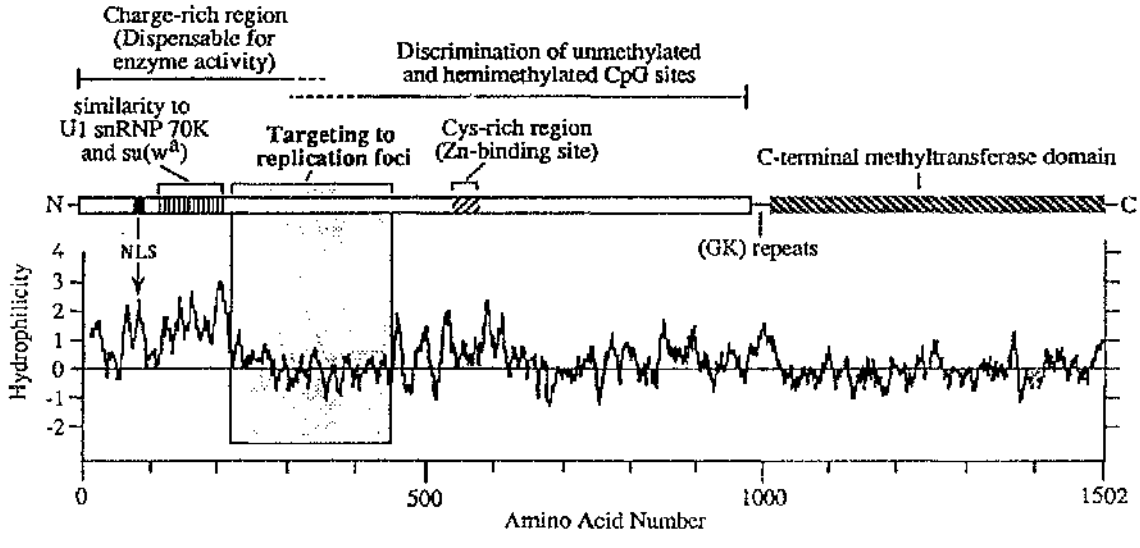


Figure 1.6 Diagram showing the sequence elements and functional domains in mammalian (cytosine-5) methyltransferase

The striped region represents the sequences required for targeting the enzyme to the replication foci during S phase. NLS is a nuclear localisation signal located between residues 72 and 92, and (GK) repeats are the run of alternating lysyl and glycyl residues that join the N- and C-terminal domains of the DNA MTase. The hydrophilic sequences have positive values in the hydrophilicity plot. Leonhardt *et al.*, (1992) *Cell* **71**, pp 870.

The first two hundred amino acids are polar in nature and their deletion from the protein does not affect the enzyme function (Bestor and Ingram, 1985), but cleavage between the N and C terminal region stimulates *de novo* methylation. Bestor and Ingram (1983) showed the *de novo* and maintenance activities of the

methyltransferase reside on the same protein and that the preference for the hemimethylated sites are at least 30-40 fold higher. This phenomenon was also observed in human placental MTase (Pfefier *et al.*, 1983). So what makes the enzyme discriminate between hemimethylated and unmethylated CpG sites? Adams *et al* (1983) observed an increased rate of *de novo* methylation after limited trypsin treatment of a crude DNA MTase preparation from mouse ascites cells. This indicated that there could be a protease-sensitive domain that makes contacts with the C5 methyl group of the hemimethylated site. Protease V8 has been shown to cleave DNA MTase between the N- and C- terminal domain. This cleavage increases the *de novo* activity encoded by the C-terminal MTase (Bestor, 1992).

To find out the precise role of the N-terminus, Leonhardt *et al.* (1992) constructed N terminal deletion mutants fused with β- galactosidase. These proteins were targeted to the cell nucleus. A construct that encoded all the amino acids of DNA MTase and had 630 amino acids of β-galactosidase fused to the C-terminus was found to be localised to the replication foci. An internal deletion of amino acids 110-307 caused the chimeric protein to take on a diffuse nucleoplasmic distribution indicating the existence of a signal in this region that directs the protein to replication foci.

The regulatory role of DNA methylation in eukaryotes still remains controversial since fundamentally important eukaryotes (mammals and plants) have a large genome and the genetic approach to solve the problems are limited. With the availability of the technique of gene targeting to embryonic stem (ES) cells (Mansour *et al.*, 1988), it has been possible to introduce predetermined mutations in any mouse gene for which the clones are available. Using this technique Li *et al.* (1992) disrupted both the alleles of DNA MTase in the ES cells with a construct which introduces short

replacements at the translation start site. These cells have a third of the wild type 5mC level and only 5% of the MTase activity in an *in vitro* assay. Embryos homozygous for the mutation complete gastulation and the early stages of organogenesis but fail to develop past the 20 somite stage. The mutant embryo contains pycnotic nuclei which are the result of apoptosis. This observation confirms that methylation is an essential feature for mammalian development.

1.5.7 Promoter of DNA methyltransferase

Rouleau *et al* . (1992) characterised the promoter of the murine MTase gene a unique housekeeping gene. This promoter has neither TATTA or CAAT boxes nor the CG rich sequences unique of housekeeping gene promoters (Bird, 1986). The 5′ region of the gene has AP-1, AP-2 and glucocorticoid response elements, which suggests that MTase gene expression could be a result of a signal transduction pathway (Rouleau *et al* . (1992)

1.6 Aim of the project

At the start of this project there were very conflicting reports on the size of plant DNA methyltransferases and no reports on either their structure or their target specificity. This project was initiated (a) to throw some light on different types or forms of DNA methyltransferase in plant (b) to analyse their recognition sequences. As the project developed so did the aims of (c) investigating the effect of different sequence specific DNA methylation on reporter gene expression; (d) cloning of methyltransferase cDNA; and identification and analysis of CNG islands in the pea genome.

CHAPTER TWO

Materials and methods

2.1 Materials

2.1.1 List of suppliers

All the chemicals used in this work were Analar grade, supplied by BDH Chemicals, Poole, Dorset, or Fisons Scientific, Loughborough, Leics., unless otherwise stated . Radiochemicals were procured from Amersham International plc., Aylesbury, Bucks. Growth media for bacterial culture were purchased from Difco Laboratories, Detroit USA and the cell culture media were from GIBCO/BRL Ltd., Paisley, Scotland. The prokaryotic methyltransferases were supplied by New England Biolabs, Beverly, MA, USA. Restriction and modifying enzymes were from Promega Ltd., Southampton, Pharmacia Ltd., Milton Keynes and Boehringer Mannheim Ltd., Lewes, UK. Pea seeds (Feltham first) were obtained from Brooker seeds ltd, Sleaford, Lincs, UK. Apart from the above certain special chemicals, reagents, equipments and kits were used from other sources which are indicated in the text.

2.1.2 Bacterial strains

Escherichia coli XL1-Blue MRF′ (Stratagene Ltd., Cambridge) was the host strain used for the growth of all plasmid DNA as well as for the plating and titering of

λZAP cDNA libraries. It had the following genotype : Δ(*mcrA*) *183*, Δ(*mcrCB-hsdSMR-mrr*) *173*, *endA1*, *supE44*, *thi-1 recA1*, *gyrA96*, *relA1*, Lac [F′ *proAB*, *lacI*^{q}ZΔ*M15*, Tn*10* (tet^{r})].

For genomic DNA library plating, titering and amplification, the following genotypes were used. XL1-Blue MRA : Δ*(mcrA)183*, Δ*(mcrCB-hsdSMR-mrr)173*, *end A1*, *supE44*, *thi-1*,*gyr A96*,*relA1*,*Lac*. and XL1-Blue MRA (P2) : XL1-Blue MRA (P2 lysogen).

For λgt 11 library plating, *E. coli* strain Y1090 was used with the following genotype : Δ(*lac)*U169, *sup*E, *sup*F, *met*B, (r_{k-}, m_{k+}), *trp*R, *ton*A21, *pro*C::Tn5, [pMC9-pBR322, *lacI*^{q}].

2.1.3 Bacteriophage strains

cDNA libraries were constructed in λgt 11 or λZAPXR (Stratagene) vectors. The genomic library was constructed in *Xho* I digested partially filled λFIX (Stratagene) vectors.

2.1.4 Plant materials

Pea seeds were soaked in water for four hours. The imbibed seeds were plated densely on a thin layer of compost on germination trays. Then the seeds were covered with a thin layer of compost and watered. The trays were kept in a growth room set at 27°C. The compost was kept moist by watering regularly. The apical shoots were harvested on the fifth day after planting and were frozen at -20°C until use.

2.1.5 Media and solutions for growth of bacteria and bacteriophage

Bacterial cultures were grown at 37°C in Luria broth (LB) medium supplemented with appropriate antibiotics in an orbital shaker at 225 rpm. Bacteria required to produce a lawn for subsequent bacteriophage infection were grown on L maltose/NZY agar plates. Strains for short term storage were maintained on LB plates at 4°C and those for long term storage were mixed with glycerol to a final concentration of 15% v/v and stored at -70°C. Bacteriophage were stored as phage suspension in phage buffer supplemented with 0.01% of chloroform at 4°C.
Media compositions are as follows:

LB medium: 1% (w/v)bactotryptone, 0.5% (w/v) yeast extract, 0.5% (w/v) NaCl, pH 7.5 (adjusted with NaOH).

LB Agar medium: As per LB Broth, except the addition of 15 g of agar/l.

LB Maltose medium: As per LB Broth, except the addition of 0.4% (w/v) maltose.

SOC medium: 1% (w/v)bactotryptone, 0.5% (w/v) yeast extract, 0.5% (w/v) NaCl, 10 mM $MgSO_4.7H_2O$, 10mM $MgCl_2$ pH 7.0 (adjusted with NaOH). Add filter-sterilised glucose to a final conc. of 20 mM.

NZY Broth: 1% (w/v) casein hydrolysate, 0.5% (w/v) NaCl, 0.5% (w/v) yeast extract, 0.2% (w/v) $MgSO_4.7H_2O$, pH 7.5 (adjusted with NaOH).

NZY Bottom agar: As per NZY Broth, except the addition of 15 g of agar/l.

Top agarose: As per NZY Broth, except the addition of 7 g of agarose/l.

Phage Buffer: 100 mM NaCl, 10 mM $MgSO_4.H_2O$, 50 mM Tris HCl at pH 7.5 and 0.01% (w/v) gelatine.

2.1.6 Buffers and solutions

TE buffer

Tris-HCl	10 mM
EDTA (pH 8.0)	1 mM

TAE buffer, 50X

Tris	242 g
glacial acetic acid	57.1 ml
EDTA (0.5 M, pH 8.0)	100 ml

Amounts per litre of 50X TAE.

1x: 40 mM Tris-acetate (pH 7.9), 1 mM EDTA.

TBE buffer (pH 8.3)

Tris	89 mM
Boric acid	89 mM
EDTA	2 mM

Alkaline buffer, 10X

NaOH (5.0 M)	5 ml
EDTA (0.5 M)	2 ml
Water	45 ml

SSPE buffer, 20X

NaCl	3.6 M
Na-phosphate	0.2 M
EDTA	0.02M

pH 7.5 with HCl

TEN buffer

Tris-HCl (pH 7.8)	20 mM
EDTA	10 mM
NaCl	12 mM

FSB buffer

K-acetate	10 mM
$MnCl_2 \cdot 4H_2O$	45 mM
$CaCl_2 \cdot 2H_2O$	10 mM
KCl	100 mM
$(Co[NH_3]_6)Cl_3$	3 mM
glycerol	10%

Buffer M+

MOPS (pH 7.0)	50 mM
EDTA	1 mM
NaN_3	0.02%
DTT	1 mM
glycerol	10%
PMSF	0.06%

Stopper

SDS	1%
EDTA	2 mM
Butanol	5%
p-Amino salicylate	3%
Salmon testis DNA	0.25mg/ml
NaCl	125 mM

Phenol mix

Phenol	88%
m-Cresol	12%
8 Hydroxyquinoline	0.1%

Denhardt's reagent, 50X

Ficoll (Type 400, Pharmacia)	1%
BSA (fraction V, Sigma)	1%
polyvinylpyrrolidone	1%

Hybridization buffer

SSPE	5X
Denhardt's	5X
SDS	0.5%

Oligohybridization solution

Tetramethylammonium Chloride	3.0 M
sodium phosphate (pH 6.8)	0.01 M
EDTA	1 mM
SDS	0.5%
salmon sperm DNA	100 µg/ml
BSA	0.1%

oligonucleotide elution buffer

NH_4-acetate	0.5 M
Mg-acetate	10 mM
EDTA	1 mM
SDS	0.1%

***Taq* DNA polymerase buffer**

KCl	50 mM
Tris-HCl (pH 9.0 at 25°C)	10 mM
Triton X-100	0.1%

T4 polynucleotide kinase buffer, 10X

Tris-HCl (pH 7.6)	0.5 M
$MgCl_2$	0.1 M
DTT	50 mM
spermidine	1 mM
EDTA	1 mM

ExoIII buffer, 10X

Tris-HCl (pH 8.0)	660 mM
$MgCl_2$	6.6 mM

S1 nuclease buffer, 7.4X

K-acetate (pH 4.6)	0.3 M
NaCl	2.5 M
$ZnSO_4$	10 mM
glycerol	50%

S1 stop buffer

Tris base	0.3 M
EDTA (pH 8.0)	0.05 mM

T4 ligase buffer, 10X

Tris-HCl (pH 7.4)	200 mM
$MgCl_2$	50 mM
DTT	50 mM
BSA	500μg/ml

CIP buffer, 10X

Tris-HCl (pH 8.4)	500 mM
$MgCl_2$	10 mM
$ZnCl_2$	10 mM
spermidine	10 mM

Annealing buffer, 10X

Tris-HCl (pH 7.5)	200 mM
$MgCl_2$	100 mM
NaCl	500 mM
DTT	10 mM

Klenow buffer, 10X

Tris-HCl (pH 7.6)	500 mM
$MgCl_2$	100 mM
DTT	1 mM

Phage precipitant

PEG-8000	33%
NaCl	3.3 M

Phage buffer

NaCl	150 mM
Tris-HCl	40 mM
$MgSO_4$	10 mM

Dideoxy reaction premix, 20X, ABI DNA sequencing kit

5X TACS buffer	80 µl
dNTP mix	16 µl
Dideoxy A terminator	20 µl
Dideoxy T terminator	20 µl
Dideoxy G terminator	20 µl
Dideoxy C terminator	20 µl

Note: The dideoxy terminators in the kit is named as 'Dyedeoxy terminator'

CAPS buffer

CAPS (pH 11.0)	10 mM
methanol	10%

SDS sample loading dye for proteins to be sequenced

glycerol	1 ml
β mercaptoetanol	0.5 ml
SDS	0.03 %
Upper tris (4X) sec 2.2.12.1	1.25 ml

final vol made to 10 ml

Tris-glycine buffer

Tris	25 mM
glycine	250 mM
SDS	0.1%

Solutions for plasmid preparations

Solution I:	glucose	50 mM
	Tris-HCl (pH 8.0)	25 mM
	EDTA	10 mM
Solution II:	NaOH	0.2 M
	SDS	1%
Solution III:	K-acetate (5 M)	60 ml
	glacial acetic acid	11.5 ml
	dH_2O	28.5 ml

Solutions for Maxam-Gilbert DNA sequencing

DMS buffer:

sodium cacodylate (pH 7.0)	50 mM
EDTA	1 mM

DMS stop solution

sodium acetate (pH 6.5)	1.5 M
β-mercaptoethanol	1 M
yeast tRNA	250 µg/ml

Hydrazine stop solution

sodium acetate (pH 6.5)	0.3 M
EDTA	0.1 mM
yeast tRNA	100 µg/ml

Sequencing gel loading buffer

formamide (deionized)	98%
EDTA	10 mM
xylene cyanol FF	0.025%
bromophenol blue	0.025%

Solution D (for RNA isolation)

guanidinium-isothiocyanate	4 M
sodium citrate, pH 7.0	25 mM
sarcosyl	0.5%
2-mercaptoethanol	0.1 M

Poly A+ mRNA purification buffers

Column loading buffer, 2X

Tris-HCl, pH 7.6	40 mM
NaCl	1.0 M
EDTA	2 mM
sarkosyl	0.2%

Elution buffer

Tris-HCl, pH 7.6	10 mM
EDTA	1 mM
SDS	0.05%

Plant DNA isolation buffers

CTAB buffer, 2X

CTAB (w/v)	2%
Tris-HCl, pH 8.0	100 mM
EDTA	20 mM
NaCl	1.4 M
PVP (40,000)	1%

10% CTAB soln.

CTAB	10%
NaCl	0.7 M

CTAB precipitation buffer

CTAB	1%
Tris-HCl, pH 8.0	50 mM
EDTA	10 mM

High salt TE

Tris-HCl, pH 8.0	10 mM
EDTA	1 mM
NaCl	1 M

2.1.7 Plasmid vectors

All the recombinant DNA constructs were made either in pBluescript II KS (+) (Stratagene, Cambridge) or pGEM 7Zf (+) (Promega, Southampton). Vector pGEM 7Zf (+) was the choice for Exo III deletion.

2.1.8 Synthetic oligonucleotides

Synthetic oligonucleotides for PCR or primer extension assays were made in the Department of Biochemistry by Dr. V. Math, using an Applied Biosystems 381A DNA synthesiser for the phosphite-triester method.

2.1.9 Molecular size and weight standards

(a) Nucleic acid molecular size standards

The following standards were used as size markers for the analysis of nucleic acids gel electrophoresis.

λ DNA/*Hind*III+*Eco*RI (λHE): 21.22, 5.14, 4.97, 4.26, 3.53, 2.02, 1.90, 1.58, 1.37, 0.94, 0.83, 0.56, 0.12 kb

100 bp ladder (GIBCO/BRL): 2072, 1500, 1400, 1300, 1200, 1100, 1000, 900, 800, 700, 600, 500, 400, 300, 200, 100 bp.

1 kb ladder (GIBCO/BRL): 12.21, 11.19, 10.18, 9.16, 8.14, 7.12, 6.10, 5.09, 4.07, 3.05, 2.03, 1.63, 1.01, 0.50, 0.51, 0.39, 0.34, 0.29, 0.22 kb

RNA ladder (GIBCO/BRL): 9.49, 7.46, 4.40, 2.37, 1.35, 0.24 kb

RNA ladder (Promega): 6.58, 4.98, 3.63, 2.60, 1.90, 1.38, 0.95, 0.62, 0.28 kb

(b) Protein molecular weight standards

Rainbow protein marker (Amersham):

Myosin (200), Phosphorylase b (97.4), BSA (69), Ovalbumin (49), Carbonic anhydrase (30), Trypsin inhibitor (21.5), Lysozyme (14.3). Size in kDa in parentheses.

2.2 Methods

2.2.1 General methods

The following methods are standard procedures in molecular biology which essentially follow protocols described by Sambrook *et al.* (1989), adapted to the equipment available in the laboratory.

2.2.1.1 Phenol/chloroform extraction

Phenol, equilibrated with TE buffer and containing 8-hydroxyquinoline at 0.1% was mixed with chloroform/isoamylalcohol (24:1) at a ratio of 1:1. Extraction of an aqueous DNA solution was performed as follows: an equal volume of phenol/chloroform was added to the DNA solution, vortexed and centrifuged for 5 minutes (min) in a microfuge; the upper aqueous phase was transferred to a fresh tube and the extraction repeated if necessary.

2.2.1.2 Ethanol / isopropanol precipitation

A salt solution was added to the DNA solution (either Na-acetate to 0.3 M or NaCl to 0.2 M) and the DNA precipitated with 2.5 volumes of ethanol (98%) at -70°C for 15 min. The DNA was recovered by centrifugation at 10,000 rpm for 15 min at 4°C. Subsequent washes with cold 70% ethanol removed any salt which had coprecipitated with the DNA. For isopropanol precipitation 0.7 volume of isopropanol was added instead of ethanol.

2.2.1.3 Spin column chromatography

This technique was used routinely to separate the unincorporated free nucleotides from the products. Chromaspin columns (Clonetech, USA) were shaken to mix the resin and were spun at 1700 x **g** for 3 min to pack the resin. The sample of 40 µl was applied in the centre of the bed and was spun at 1700 x **g** for 4 min to collect the product.

2.2.1.4 Restriction mapping and agarose gel electrophoresis

Restriction digests were routinely carried out for 1-16 hours at 37°C (or the temperature recommended by the manufacturer if it was different) in 20-100 µl aliquots which contained a suitable amount of DNA, 1-10 units per µg DNA of the appropriate restriction enzyme and 1 X restriction enzyme buffer. The restriction buffers used were those supplied with the restriction enzymes.

DNA fragments were analysed on 0·7-2·0% agarose slab gels. Gels were prepared with 1 X TBE Buffer (Section 2.1.6) which contained 0·1 $\mu g ml^{-1}$ ethidium bromide. Samples were prepared for loading by the addition of 0·2 volumes of 6 X Loading Buffer (0·25% (w/v) bromophenol blue, 40% sucrose and 50 mM EDTA). The gel was placed in a tank which contained 1 X TBE buffer and the samples were loaded into the submerged wells. Electrophoresis was done at 2-5 volts/cm. The length of time that electrophoresis was carried out depended upon the length of the gel and the size of fragments which were to be resolved. DNA was visualised under 320 nm UV light.

2.2.1.5 Isolation of nucleic acids from agarose gels

Two methods were used to remove DNA separated by electrophoresis from agarose gels.

(a) Electroelution onto DE-81 paper

This method was based on that of Dretzen *et al.* (1981). DE-81 Whatman paper was used without any initial priming. DNA was electrophoresed directly onto DE-81 paper, which had been placed in a slot cut immediately in front of the band of interest. Then, the DE-81 paper was transferred to a 1·5 ml microfuge tube and the bound DNA eluted off the paper into 200 µl of 1·5 M NaCl in TE by incubation at 65°C for 60 minutes. After incubation, the paper was washed once again with the high salt solution and the pooled washes were extracted with an equal volume of phenol:chloroform. This was followed by a chloroform extraction and precipitation of the DNA with three volumes of ethanol. The DNA was resuspended in a minimal volume of TE and passed through a spin column to remove any small particles of paper and remaining salt.

(b) Gene clean of DNA

Methods were followed as per manufacturer BIO101 (USA) instruction manual. This method gives clean DNA for ligation and has a working range between 1 to 5 µg.

2.2.1.6 Small scale preparation of plasmid DNA

A 3 ml overnight culture was prepared from a single colony of transformed bacteria in LB medium supplemented with ampicillin (100 µg/ml). 1.5 ml of bacterial suspension was transferred to a microcentrifuge tube and the bacteria pelleted at 5,000 rpm for 5 min. The pellet was suspended in 100 µl solution I (see Section 2.1.8). 200 µl solution II was added, mixed and incubated for 5 min on ice. Subsequently 150 µl of solution III was added, mixed and incubated again for 5 min on ice. The bacterial debris was pelleted at 12,000 rpm for 10 min at 4°C. To the supernatant was added 460 µl phenol/chloroform mix; the mixture was vortexed and the phases separated at 10,000 rpm for 10 min at 4°C. To the aqueous phase was added 1 ml ethanol and the sample was incubated for 15 min at RT. The DNA was pelleted at 12,000 rpm for 10 min at 4°C and the pellet was washed in 70% ethanol and air-dried. The pellet was resuspended in 10 µl of TE buffer containing RNase A at 1 µg/µl. 1 µl was used for a restriction digest as per the enzyme manufacturer's instructions.

2.2.1.7 Large scale plasmid preparation

Transformed cells were grown overnight in 200 ml LB medium supplemented with ampicillin (100 µg/ml). The cells were harvested by centrifugation for 15 min at

6,000 rpm in a Beckman J2-21 centrifuge using a JA 14 rotor. The pellet was resuspended in 10 ml solution I (Section 2.1.8). Lysis of the cells was achieved by addition of 10 ml solution II and incubation on ice for 5 min. Addition of 7.5 ml solution III precipitated out the cell debris which was sedimented at 10,000 rpm at 4°C (Beckman J2-21 centrifuge, JA 20 rotor). The supernatant was filtered through Whatman 3MM paper and 0.6 volumes of isopropanol were added to precipitate the nucleic acids for 15 min at room temperature. The nucleic acids were pelleted at 3,000 rpm for 15 min at 4°C using a Beckman CS-6R bench top centrifuge and resuspended in 2 ml TE buffer. RNAs were removed by addition of 30 µl DNase-free RNase (10 mg/ml) and incubation for 1 hour at 37°C. The DNA was ethanol precipitated after a phenol, a phenol/chloroform and a chloroform extraction and finally dissolved in 500 µl TE buffer. Alternatively, before RNAse treatment the TE plasmid solution is processed for CsCl EtBr centrifugation (Section 2.2.1.11) to obtain pure plasmid DNA.

2.2.1.8 Quantitation of nucleic acids

The absorption of various dilutions of DNA samples was measured using quartz 1ml cuvettes in a spectrophotometer at 260 nm. An OD of 1 corresponds to approximately 50 µg/ml for double-stranded DNA, 40 µg/ml for single-stranded DNA and RNA and 30 µg/ml for oligonucleotides. For quantitative assay of more dilute samples of DNA (e.g. cDNA) a dot assay technique on an Et Br plate was done. Known standards (0.5 µl) between the range of 10 to 500 ng/µl was spotted on a 1% agarose TAE-EtBr plate alongside the samples. An empirical quantitative estimation is done by visually correlating the amount of fluorescence.

2.2.1.9 Protein quantification

The protein concentration in all samples were assayed by Bradford's method (Bradford, 1976). Bovine serum albumin (BSA) was used as a standard protein. BSA (0-20 µg) was taken in 50 µl of buffer M^+ and a ml of Bradford's reagent was added, mixed and left to stand for 5 min and the absorbance is measured at 595 nm. The protein concentration was estimated from the standard curve.

2.2.1.10 SDS/PAGE of proteins

SDS polyacrylamide gel electrophoresis was performed according to the method of Laemmli (1970). A 7.0% resolving gel was prepared by adding 2.5 ml of acrylamide soln (29:1; Sambrook *et al.*, 1989), 2.5 ml of 1.5M Tris-HCl pH 8.8, 0.1 ml of 10% SDS, 0.1 ml of 10% APS, 10 µl of TEMED and water made up to 10 ml. The stacking gel (3.6%) was prepared by adding 0.6 ml of acrylamide soln (29:1), 0.6 ml of 1.5M Tris-HCl pH 6.6, 0.05 ml of 10% SDS, 0.05 ml of 10% APS, 5 µl of TEMED and water made up to 5 ml. The resolving gel was poured first and allowed to polymerise at RT for 30 min with a thin layer of 0.01% SDS on the top of the gel, thus facilitating the polymerisation giving a flat top. After the polymerisation, the SDS solution was removed and the stacking gel solution was poured. A well-forming comb was inserted. The samples were boiled with equal volume of SDS sample loading buffer (2.1.6) for 3 min and loaded on to the gel. The electrode buffer used was Tris-glycine (2.1.6) and Rainbow protein molecular wt markers (2.1.6) were used. The gel was run at 15-45 milliamp constant current. After the run was completed the gel was removed and incubated in 50% methanol overnight for silver staining.

2.2.1.11 Silver staining

Solution A was made by dissolving 0.8 grams of silver nitrate in 4 ml distilled water. Solution B was made by mixing 21 ml of 0.36% NaOH with 1.4 ml of 14.8 M ammonia. Solution C was freshly prepared by dropwise addition of Solution A into Solution B with constant steering. The final volume of solution C was made up to 100 ml with addition of water. The gel was stained with freshly prepared solution C for 15 min with constant gentle agitation. Two distilled water rinses of 5´ were given to remove residual solution C. The gel was developed by adding solution D (2.5 ml 1% citric acid, 0.25 ml of 38% formaldehyde and 500 ml of water) to it and agitating till the colour appeared. After the colour development the gel was quickly rinsed with water three times and photographed. Later the gel was stored in 50% methanol.

2.2.1.12 Purification of DNA by equilibrium centrifugation in CsCl-ethidium bromide gradients

If large amounts of pure DNA were required, the miniprep alkaline lysis method was scaled up to accommodate a 400 ml bacterial culture and the resulting DNA solution was passed through a CsCl-ethidium bromide gradient.

The exact volume of the DNA solution was determined and for every 1 ml of solution 1 g of CsCl was added and mixed until it had dissolved. Then, 0·8 ml of ethidium bromide (10 $mgml^{-1}$) was added for every 10 ml of the DNA/CsCl solution. The final density of the solution was 1·55 gml^{-1}

This solution was centrifuged at 12000 x g for 5 min at RT. The scum which floated to the top or adhered to the sides of the tubes represented complexes formed by ethidium bromide and bacterial protein; care was taken to avoid the scum in the proceeding steps.

The clear red solution was transferred into 5.1 ml polypropylene ultracentrifuge tubes (Beckman, UK) and heat sealed. Next, the tubes were balanced and then centrifuged at 45000 rpm in a VTi 65 rotor (Beckman) overnight at 22°C. After centrifugation, two bands of DNA located near the centre of the gradient were visible. The upper band consisted of chromosomal and linear DNA; the lower band of the desired closed circular plasmid DNA.

This lower band was 'pulled off' the gradient under 320 nm UV light, by the insertion of a 21-gauge hypodermic needle into the top of the tube, which allowed for the entry of air, and the insertion of an 18-gauge hypodermic needle just below the lower band which was to be collected. The DNA was collected with the aid of a sterile 1 ml syringe.

The same purification technique was followed for obtaining pure plant DNA (2.2.5.1). The banded plant DNA was collected and further purified as described below.

2.2.1.13 Removal of ethidium bromide from DNAs purified by equilibrium centrifugation

To the solution of DNA an equal volume of water or 20 X SSC saturated butan-1-ol was added. The two phases were mixed and left to separate. The organic

phase was removed and the extraction repeated a further 4-6 times until all the pink colour, due to ethidium bromide, had disappeared from both phases.

Then, caesium chloride in the sample was diluted with the addition of three volumes of sterile deionised water and the DNA precipitated by the addition of six volumes of ethanol. Genomic DNA floats out as a white lump. For the preparation of high molecular weight plant DNA the banded DNA-CsCl solution was dialysed (2.2.5.1) against several change of TE to dialyse out the CsCl salt. However, for plasmids, a 30 min incubation at 4°C followed by centrifugation at 1000x*g* for 15 minutes at 4°C was required. The resulting DNA pellet was washed with 70% ethanol, air dried and resuspended in a small volume of TE or autoclaved water.

2.2.1.14 Purification of oligonucleotides (crush and soak method)

This method was used for obtaining oligonucleotides of uniform length for *in vitro* methylation. These *in vitro* methylated products were used for Maxam-Gilbert DNA sequencing. 60 µg (approx. 2 OD_{260}) of the oligonucleotide were lyophilised to dryness in a rotary vacuum evaporater to remove the NH_4OH. The pellet was then dissolved in 20 µl H_2O and 20 µl formamide was added. Before loading onto the gel the DNA was heat denatured for 5 min at 55°C. Separation was on 1 mm thick, 19% denaturing polyacrylamide (acrylamide : N,N'-methylenebisacrylamide 19:1) gels in 1X TBE buffer (Section 2.1.6). The gel was pre-run for 30 min at 30 mA before the oligonucleotide was loaded in 8 slots. To monitor the electrophoresis 5 µl of a 50% formamide solution of 0.05% xylene cyanol FF and 0.05% bromophenol blue was run alongside the oligonucleotide samples. After completion of the electrophoresis, the apparatus was dismantled and

the gel transferred to a piece of Saranwrap. The gel was then transferred onto a fluorescent thin-layer chromatography plate (Merck, Silica gel F_{254}). The gel was taken to the dark room and illuminated with a hand-held, long-wavelength UV lamp. The oligonucleotide absorbed the UV-light and appeared as dark bands on the fluorescing background of the chromatography plate. The top bands of the oligonucleotide (= full size oligonucleotide) were marked, cut out of the gel, transferred to microfuge tubes and crushed using a yellow pipette tip. 300 µl of oligonucleotide elution buffer (Section 2.1.6) per band were added and incubated for 16 hours at 37°C. The gel pieces were pelleted by centrifugation at 12,000 rpm for 10 min in a microcentrifuge. The supernatants were pooled, ethanol precipitated and dissolved in 50 µl TE buffer. Finally, to remove any remaining gel particles the oligonucleotides were purified using a spin column as described in Section 2.2.1.13.

2.2.1.15 End labelling of oligonucleotides

Synthetic oligonucleotides were end labelled by transfer of the γ-^{32}P from [γ-^{32}P]ATP using bacteriophage T4 polynucleotide kinase (PNK). 10 pmol of oligonucleotide were incubated in T4 PNK buffer (Section 2.1.6) with 5 µl of [γ-^{32}P]ATP (5,000 Ci/mmole, 10 mCi/ml) and 10 units of T4 PNK in a total volume of 20 µl. Unincorporated radiolabel was removed using Chromaspin columns (Section 2.2.1.3).

2.2.1.16 Dephosphorylation of plasmid DNA

The 5´-phosphate groups of a linear plasmid were removed by calf intestinal alkaline phosphatase (CIP) to avoid self-ligation of the vector. 1 µg of linearised vector

DNA was incubated in CIP buffer (Section 2.1.8) with 1 unit of calf intestinal alkaline phosphatase (Boehringer Mannheim) in a total volume of 50 µl at 37°C for 30 min. SDS and EDTA were added to final concentrations of 0.5% and 5 mM, respectively. Proteinase K (Boehringer Mannheim) was added to a final concentration of 100 µg/ml and incubated for 30 min. at 37°C. The reaction was subjected to a phenol and phenol/chloroform extraction and the DNA recovered by an ethanol precipitation.

2.2.1.17 Ligations

Routinely, 200 ng of linearised vector DNA were used in a ligation in a total volume of 10 µl. The vector:insert ratio was 1:3. 1 µl of 10X ligation buffer (Section 2.1.8) and 1 µl of 10 mM ATP were added to the DNA and the ligation initiated by addition of 1 Weiss unit of bacteriophage T4 DNA ligase. Incubation was for 8 to 16 hours at 16°C.

2.2.1.18 Preparation of competent cells

For the preparation of competent cells five well separated colonies of *E. coli* XL1-Blue MRF´ (from a fresh overnight grown plate) were transferred to 100 ml of SOC medium and incubated 3 to 4 h at 37°C, 225 rpm, until an OD_{600} of 0.45 - 0.60 was reached. The cells were then cooled on ice for 10 min and harvested by centrifugation at 4,000 rpm for 10 min at 4°C using a Beckman J2-21 centrifuge and JA14 rotor. The pellet was carefully resuspended in 40 ml of ice cold FSB buffer (Section 2.1.6) and stored for 10 min on ice. The cells were recovered as above and resuspended in 8 ml of FSB buffer. 400 µl of DMSO was added to the cells, mixed and left on ice for 15 min Another 400 µl of DMSO was added, mixed and cells were

dispensed into aliquots of 100-200 µl and snap-freozen in liquid nitrogen. The tubes were stored at -70°C until needed.

2.2.1.19 Transformation of bacteria

Competent cells (100 µl, Section 2.2.1.16.) were allowed to thaw on ice and plasmid DNA (no more than 50 ng) was added. After a 30 min incubation on ice, the tubes were transferred to a 42°C water bath, incubated for 90 seconds and rapidly transferred to an ice bath. After 5 min, 800 µl of SOC was added and incubated for 45 min at 37°C to allow the bacteria to express the antibiotic resistance. Dilutions of the bacterial suspension were made and plated out on LB plates containing appropriate antibiotics.

2.2.1.20 DNA methyltransferase assay

10 µl of dialysed fraction was assayed using either (poly dI-dC).(poly dI-dC) or hemimethylated CNG substrate (1 µg) and 1 nmole of ^{3}H AdoMet in a total reaction volume of 60 µl. The incubation was for 1 hour at 30°C. Further steps were carried out as previously described (Yesufu *et al*, 1988).

2.2.2 Methyltransferase/(s) purification from pea shoot apices

2.2.2.1 Preparation of total soluble pea methyltransferase

Pea seeds (*Pisum sativium*, variety : Feltham First) were obtained from Brooker Seeds, Sleaford, Lincs., U.K. and were grown under continuous white light

illumination at 26°C. Typically, nuclei were prepared from 4–5 kg of five day old shoot tips as described by Yesufu *et al.* (1991), using buffer M (50 mM Tris-HCl, pH 7.3; 1 mM EDTA; 0.01% (w/v) NaN_3; 1 mM DTT; 70 µg PMSF/ml and 10% (v/v) glycerol. All steps were carried out at 0–4°C. Purified nuclei were mixed with buffer M containing 0.2 M NaCl and kept on ice for 30 min with occasional shaking. The extract was spun for 25 min at 4500 rpm in a JA14 rotor (Beckman) and the supernant was collected.

2.2.2.2 Heparin Sepharose chromatography

Heparin-Sepharose (Pharmacia), equilibrated with buffer M containing 0.2 M NaCl was shaken with the nuclei extract for 30 min; centrifuged at 4000 rpm for 20 min and the supernatant discarded. The pellet was washed twice with buffer M containing 0.2 M NaCl and the bound proteins eluted with buffer M containing 0.6 M NaCl. Buffer M was then added to the elutant to reduce the final concentration of NaCl to 0.15 M.

2.2.2.3 Q-Sepharose chromatography

A 1.5 x 4.0 cm column of Q Sepharose (Pharmacia) was washed with five column volumes of buffer M, followed by buffer M containing 0.6 M NaCl and was finally equilibrated with buffer M containing 0.15 M NaCl. The heparin-Sepharose elutant was loaded onto the column at a flow rate of 1.0 ml/min. Unbound proteins were washed away with five column volumes of buffer M containing 0.15 M NaCl and bound proteins eluted, at a flow rate of 0.5 ml/min, with buffer M containing 0.6 M NaCl. Fractions of 2.0 ml were collected and dialysed for 6 hours against buffer M at

0°C. The conductivity of the protein solutions were measured to ensure the removal of salt from these Q fractions prior to assay for methyltransferases activity.

2.2.2.4 Pea DNA affinity chromatography

The peak methyltransferases fractions eluted from Q-Sepharose were pooled and applied to a pea DNA affinity column (1.5 x 4.0 cm) previously washed with buffer M containing 1.0 M NaCl and equilibrated with buffer M. The sample was allowed to stand in the column for 30 min and the resin was then washed four times with 4 ml of buffer M by gravity flow. The passage of buffer through the column was then stopped, 4 ml buffer M containing 1.0 M NaCl was added to the column and the resin was thoroughly mixed by gently shaking. The mix was allowed to stand for 10 min, and the protein was eluted as one peak. The same process was performed once again with 3 ml of buffer containing 1.0 M NaCl to remove the residual proteins. The two eluants were mixed and dialysed against buffer M for 9 h.

2.2.2.5 MonoQ fractionation of CG and CNG methyltransferase

A Mono Q HR5/5 (Pharmacia) column was equilibrated with buffer M and the sample was applied to the column by multiple injection using a 2 ml sample loop. Unbound proteins were washed away with three column volumes of buffer M. A linear gradient was run from 0 to 1.0 M NaCl at a flow rate of 0.5 ml/min. Fractions of 1.0 ml were collected and dialysed against buffer M containing 50% glycerol. Table 1 shows a typical purification.

2.2.2.6 CNG DNA affinity purification of CNG methyltransferase

CNG oligonucleotide affinity column was prepared as per the method of Kadonaga and Tijan (1986). 1.5 gram of CNBr activated Sepharose (Pharmacia) was washed with 100 ml of 1 mM HCl. The HCl was removed from the matrix by repeated washing with sterile water. 100 ml of 10 mM potassium phosphate, pH 8.0 was mixed with the matrix and washed. The activated matrix was resuspended in 10 ml of 10 mM potassium phosphate and 50 µg of duplex oligonucleotides. The mixture was rotated in a rotary wheel for 16 hrs. at room temp. The resin was collected and washed with sterile water first then followed by ethanolamine-HCl, pH 8.0. The final matrix was resuspended in 7 ml of ethanolamine-HCl, pH 8.0 and left on a rotary wheel for 6 h. The resin was collected in a sintered funnel and washed with (1). 50 ml of 10 mM potassium phosphate, pH 8.0, (2). 50 ml of 1 M potassium phosphate, pH 8.0, (3). 50 ml of 1 M KCl, (4). 100 ml of sterile water. The final affinity matrix was resuspended in resuspention buffer (2.1.6) and stored at 4 °C.

2.2.3 Characterisation of methyltransferases

2.2.3.1 Native molecular weight determination

A Sephacryl S-300 (Pharmacia) gel filtration column (5 mm x 25 cm) was equilibrated with buffer M and calibrated with Pharmacia gel filtration protein standards (catalase, 232 kDa; aldolase, 158 kDa and bovine serum albumin, 67 kDa). 25–50 µl of the enzyme from the Mono Q peak fractions were applied to the column and fractions were collected dropwise. Each drop was assayed, with either the CNG or the CG substrate, for methyltransferase activity.

2.2.3.2 Oligonucleotide synthesis and annealing

Oligonucleotides were synthesised using an Applied Biosystems automated synthesiser. The phoshporamidite derivatives were obtained from British Drug House or Cruachem chemicals, Glasgow, with the exception of the methylcytosine which was obtained from Glen Research Corp, Hernden, VA, USA. Hemimethylated substrates were prepared by annealing equimolar concentrations of methylated strand and the corresponding complementary unmethylated strand as described by Kadonaga and Tijan (1986).

The oligonucleotides had the following composition:

1. CNG 21mer hemimethylated substrate:

5'-CAG CAG CAG CAG CAG CAG CAG -3'

3'-GTM GTM GTM GTM GTM GTM GTM -3'

2. 41mer hemimethylated substrate

5'-TGC GCC CCG TGC CTG CCA GCT GAG CGT TGT GGT GGA TGG GG-3'

3'-ACG MGG GGM ACG GAM GGT MGA MTC GMA ACA CCA CCT ACC CC-5'

3. 42-mer hemimethylated substrate

5'-TGT GTC CAG CGC CCG TCA CGG GAC CGG CAG CTG CTG GAA TGG-3'

3'-ACA CAG GTM GCG GGM AGT GMC CTG GCM GTC GAM GAC CTT ACC-5'

4. (Poly dI-dC).(poly dI-dC) (Sigma) was used as the CG-specific enzyme substrate.

2.2.3.3 Gel electrophoretic analysis of crosslinked DNA/protein complexes

1 ng of (γ^{32}P)-5'- end labelled CNG hemimethylated 21-mer duplex DNA (50,000 cpm) was incubated with 2.5 µg of protein (fractions across the MonoQ gradient), 2 µg (poly dI-dC).(poly dI-dC) and 10 µl of buffer M^+ in a total volume of 30 µl and kept on ice for 10 min. The mixture was transferred onto a piece of Parafilm kept on ice and irradiated for 20 min with a 15 watt UVGL 58 (UVP) germicidal lamp at 254 nm (short wavelength) held at a distance of 4 cm. 2 µl of 100 mM $MgCl_2$ and 2 µl of 100 mM $CaCl_2$ and 1unit of DNase I were added, mixed thoroughly and incubated at 37°C for 30 min. and the reaction terminated by addition of SDS sample buffer. After boiling for 2 m., the mixture was loaded onto a 7.5% SDS/polyacrylamide gel and electrophoresed at 40 mA constant current for 4.5 hours. The gel was fixed with 7% acetic acid for 15 min, dried at 70°C under vacuum and autoradiographed.

2.2.4 Target specificity of methyltransferase

2.2.4.1 *In vitro* methylation and methylated target sequence analysis

The unmethylated strand of the hemimethylated 41 and/or 42mer was end labelled at the 5' terminus by using γ-[^{32}P] ATP and T4 polynucleotide kinase. The reaction was terminated by incubating the mix at 65°C for 15 min and annealed to the complementary methylated strand. The annealed products were purified from unincorporated ATP using a Chromaspin 10 spun column (Clonetech,USA). The

annealed products were further separated on a 20% acrylamide gel, the oligo band was cut from the gel and eluted by a crush and soak method (Maniatis *et al* 1989). The final pellet was dissolved in distilled water.

Typically, 5-10 µl (10 to 15 ng) of the radiolabelled oligo duplex was methylated with 15 units of the CNG enzyme or a vast excess of the CG enzyme in the presence of cold AdoMet (New England Biolabs, USA) at 30°C for 8 hrs. The methylated products were phenol-chloroform extracted twice and ethanol precipitated. The terminally labeled oligonucleotides were dissolved in sterile distilled water at 20,000 to 50,000 cpm /µl as determined by Cerenkov counting.

2.2.4.2 Maxam-Gilbert sequencing of the methylated products

Chemical cleavage and sequencing reactions were performed as described by Maxam and Gilbert (1980). The four base-specific cleavage reactions were carried out as follows:

G reaction:

5 µl radiolabeled DNA were mixed with 4 µl sonicated salmon sperm DNA (sssDNA, 1 µg/µl) and 190 µl DMS buffer (Section 2.1.6). The mixture was chilled to 0°C and 5 µl of 10% DMS (dimethyl sulphate, Aldrich) were added, mixed and incubated for 5 min at 20°C. To stop the reaction 50 µl of DMS stop solution (Section 2.1.6) were added, followed by 750 µl ethanol and incubation on dry ice.

A + G reaction:

10 µl radiolabelled DNA were mixed with 4 µl sssDNA (1µg/µl) and 10 µl dH_2O and chilled to 0°C. 25 µl formic acid (98%) were added and incubated for 4 min at 20°C.

The reaction was stopped by addition of 200 µl hydrazine stop solution (Section 2.1.6), followed by 750 µl ethanol and incubation on dry ice.

C + T reaction:

10 µl radiolabelled DNA were mixed with 4 µl sssDNA (1µg/µl) and 10 µl dH_2O and chilled to 0°C. 30 µl hydrazine (99%, Eastman Kodak, Hemel Hempstead, Herts.) were added and incubated 7 min at 20°C. The reaction was stopped by addition of 200 µl hydrazine stop solution, followed by 750 µl ethanol and incubation on dry ice.

C reaction:

5 µl radiolabelled DNA were mixed with 4 µl sssDNA (1µg/µl) and 15 µl NaCl (5 M) and chilled to 0°C. 30 µl hydrazine were added and incubated 5 min at 20°C. The reaction was stopped by addition of 200 µl hydrazine stop solution, followed by 750 µl ethanol and incubation on dry ice.

All four reactions were then processed in the same way.

The DNA was pelleted by centrifugation at 12,000 rpm for 10 min at 4°C, the supernatant was removed to the appropriate waste bottle and the DNA dissolved in 300 µl 0.3 M sodium acetate (pH 5.4). The solution was ethanol precipitated and the pellet subsequently washed with 98% ethanol and 70% ethanol. After drying the DNA in a rotary vacuum lyophilizer (Jouan RC10.10.), it was dissolved in 100 µl piperidine (1M). Cleavage of the sugar-phosphate chain of the DNA was carried out for 30 min at 90°C. The samples were then lyophilized to dryness, dissolved in 20 µl H_2O, transferred to a fresh tube and once more lyophilized in a rotary vacuum lyophilizer. This step was repeated until all piperidine was removed (smell test!!). The DNA was then dissolved in 10 µl sequencing gel-loading buffer (Section 2.1.6) of which 3-5 µl were used for separation on a 20% sequencing gel.

2.2.4.3 Sequencing gel

Denaturing polyacrylamide gels (0.2 mm thick) were used for the separation of the Maxam-Gilbert sequencing reaction. 50 ml of a 20% gel was made by mixing 9.5 g of acrylamide, 0.5 g of N,N'-methylenebisacrylamide, 5 ml of 10x TBE buffer and 24 g of urea. The solution was filtered through a nitrocellulose filter to remove any undissolved particles and polymerisation was initiated by the addition of 200 µl 10% ammonium persulfate and 50 µl TEMED. After complete polymerization (1 hour) the gel was assembled onto the electrophoresis apparatus and pre-run at 30 mA for 30 min in TBE buffer. The heat denatured samples were loaded and electrophoresis was carried out at 50 watt for the desired time. Upon completion of the run (as monitored by the migration of the two dyes in the loading buffer) the gel plates were separated and the wet gel was covered by Saranwrap (Dow Chemical Co., USA) and exposed to Hyperfilm (Amersham, UK) at –70°C.

2.2.5 Nucleic acids extraction and purification from plant tissue

2.2.5.1 Extraction and purification of plant DNA

Young plant shoot apices were harvested and quick frozen in liquid nitrogen. The tissue was grounded with liquid nitrogen to powder in a chilled mortar and pestle. Equal volume of hot 2 X CTAB was added to the tissue and mixed. Equal volume of chloroform/isoamylalcohol (24:1) was added to the above and mixed thoroughly and gently to form an emulsion. The emulsion was centrifuged and aqueous phase collected in a centrifuge tube. 1/5th volume of 5% CTAB solution was added and mixed gently.

Another chloroform/isoamylalcohol extraction was done and the aqueous phase was collected. Equal volume of the CTAB precipitate (Section 2.1.6) was added and mixed gently. At this stage the nucleic acids come out as a precipitate. The precipitated DNA was dissolved in high salt TE (Section 2.1.6) and ethanol precipitated. CsCl purification of the DNA was done as described in Section 2.2.1.12. The DNA-CsCl solution was dialysed against TE (Section 2.1.6) and stored at 4°C. The purity of the DNA was estimated by measuring A260/A280 and electrophoresis on a 0.5% agarose-TBE gel.

2.2.5.2 Extraction of RNA from plant tissue

Young plant shoot apices or tissues were harvested and quick frozen in liquid nitrogen. The tissue was ground to powder with liquid nitrogen in a chilled mortar and pestle. One ml of Solution D (Section 2.1.6) was added for each gram of tissue and was shaken. The following chemicals were added in a sequential manner with shaking (1). 2 M sodium acetate 0.2 ml/gm (2). Water saturated phenol 1.0 ml/gm and (3). chloroform:isoamyl alcohol 0.4 ml / gm of starting materials. After a vigorous shaking the material was kept on ice for 15 min. The mixture was centrifuged at 10000 x **g** for 20 min at 4°C. The pellet was collected and given several (at least 3) washes with 2.5 M of sodium acetate at pH 5.2. The final pellet was washed with 70 % ethanol and air dried. The RNA pellet was dissolved in sterile DEPC treated water and was kept frozen at -80 °C.

2.2.5.3 Poly A+ RNA fraction

A gram of oligo dT was resuspended in 0.1 N NaOH. The matrix was poured into a sterile disposable plastic syringe whose bottom was plugged with sterile glass

wool. The matrix was washed twice with three column volume of water followed by column loading buffer (Section 2.1.6) till the elutant's pH was about 8.0. The total RNA was heated to 65°C and cooled to RT before addition of equal column vol. of 2X loading buffer. The RNA solution was applied to the column and the elutant was collected, heated to 65°C, cooled to RT and was reapplied again. The matrix was washed with 10 column volume of 1 X Column loading buffer to get rid of the unbound RNA. Three column volume of elution buffer was applied to the column and fractions of 0.5 ml was collected. The fraction containing RNA was pooled and the NaCl conc. was adjusted to 500 mM and the above process was repeated once again. The final Poly A+ RNA fractions were pooled and ethanol precipitated. The final pellet was washed with 70% ethanol and air dried. The RNA pellet was dissolved in sterile DEPC treated water and was kept frozen at -80°C. The purity of RNA was determined by formaldehyde-agarose-MOPS gel electrophoresis (Section 2.2.5.4).

2.2.5.4 Electrophoresis of RNA in formaldehyde gel

For each 100 ml of 1% formaldehyde agarose gel, a gram of agarose was taken with 10 ml of MOPS 10X (Section 2.1.6) and 73 ml of water. After the agarose is dissolved in microwave, it was allowed to cool to 55°C and 17 ml of formaldehyde was added to it. The gel was cast in a fume hood and allowed to settle for 30 min. RNA samples were prepared (3-5 µg of Poly A+ RNA, 12.5µl of deionised formamide, 10X MOPS 2.5 µl, formaldehyde 4 µl, water up to 25 µl), denatured at 65°C for 5 min. and kept on ice. The electrophoresis buffer was MOPS (3-[N-Morphalino]-propanesulphonic acid). The samples were mixed with loading dye (Section 2.1.6) and loaded onto the gel. The electrophoresis was carried out at 80 to 100 volts till the bromophenol blue dye was about a cm from the bottom of the gel. The gel was directly

processed for northern blotting or was stained with ethidium bromide (0.5 µg/ ml in 10 mM ammonium acetate) for 30 min followed by destaining in sterile water.

2.2.6 Construction of cDNA libraries

2.2.6.1 Denaturation of RNA with methyl mercury hydroxide

5 µg of undegraded poly A+ RNA was taken in an microcentrifuge tube. The final volume was adjusted to 20 µl and was denatured at 65°C for 5 min. After cooling the RNA to room temperature (RT), 2 µl of 100 mM methyl mercury hydroxide (MeHg) was added and incubated at RT for another 1 min. To the RNA MeHg, 4 µl of 700 mM ß mercaptoethanol was added. The tube was left at RT for another 5 min. This RNA was now ready for 1st strand cDNA synthesis. Due to the toxicity of MeHg only the required amount was used at any time and the used waste was treated with DTT before disposal. All the manipulations were carried out in a fume hood.

2.2.6.2 First strand cDNA synthesis

To construct the cDNA library a commercial cDNA synthesis and cloning kit (cat no : 200400, 237211) from Stratagene (USA) was used. The steps followed are summarised below. To the above denatured RNA mix, 21.5 µl of 2 x reaction mix (Section 2.1.6) was added and mixed. After 5 min., 2.5 µl of Superscript RNAse H^- was added and mixed thoroughly. From the above mix 5 µl was aliquoted to another microcentrifuge tube and 0.5 µl of γ-[^{32}P] ATP (800 Ci/mmol) was added. Both the microcentrifuge tubes were incubated at 37°C for an hour. The radioactive sample was

stored at -20°C for analysis and the non-radioactive sample was processed for 2nd strand cDNA synthesis.

2.2.6.3 Second strand cDNA synthesis

All the steps for 2nd strand synthesis were carried out on ice. To the non -radioactive tube the following components are added in order: 40 µl of 10 X second strand synthesis buffer, 6 µl of second strand nucleotide mix, 288.1 µl of sterile water, 1 µl of γ-[^{32}P] ATP (800 Ci/mMol, Amersham cat no : PB10384), 2 µl of RNase H (1.1 u/µl), 5.7 µl, 13.2 µl of DNA pol I (7.6 u/µl). The total reaction volume was 400 µl. After mixing, the microcentrifuge tube was incubated at 16°C for 2.5 h. The mix was transferred on ice and the cDNAs were phenol/chloroform extracted and ethanol precipitated overnight at -20°C. The cDNA was pelleted at 14000 rpm at 4°C for an hour. The pellet was washed in 70 % ethanol, air-dried and resuspended in 45 µl of water. 3 µl of the 2nd strand synthesis product was kept for analysis. At this point the first strand and 2nd strand synthesis products were checked by alkaline agarose electrophoresis (Section 2.1.6) along with the radioactive λHE marker.

2.2.6.4 Blunting the cDNA termini

To the 42 µl of the second strand synthesis product the following were added and mixed in the following order. 5.0 µl of buffer 3 (Section 2.1.6), 2.5 µl of blunting dNTPs (2.1.6), 1.0 µl (1.0 u) of Klenow DNA polymerase. The mix was incubated at 37°C for exactly 30 min. After 30 min, the microcentrifuge tube was transferred to ice and the cDNAs were phenol/chloroform extracted and ethanol precipitated overnight at -20°C.

The cDNA was pelleted at 14000 rpm at 4°C for an hour. The pellet was washed in 70 % ethanol and air-dried.

2.2.6.5 Ligating *Eco*RI adapters

The pellet was resuspended in 7 µl of *Eco*RI adapter (checked with a Geiger counter). To the microcentrifuge the following components were added in the following order: 1.0 µl of 10 X buffer 3, 1.0 µl of 10 mM rATP, 1.0 µl (4 Wess unit) of T^4 DNA ligase. The components were spun down and incubated at 8°C overnight. The reaction was stopped with heat inactivation of the ligase at 70°C for 30 min.

2.2.6.6 Kinasing the *Eco*RI ends

After cooling down the microcentrifuge to RT these components were added in the following order. 1.0 µl of buffer 3, 2.0 µl of 10 mM rapt, 6.0 µl of sterile water, 1.0 µl (10 u) of polynucleotide kinase. The microcentrifuge was incubated for 30 min at 37°C. The reaction was stopped with heat inactivation at 70°C for 30 min.

2.2.6.7 *Xho*I digestion

The microcentrifuge was briefly centrifuged and these components were added in the following order. 10 X buffer XI 5 µl, water 22 µl, *Xho* I 3 µl (40 u/µl). The incubation was done at 37°C for 1.5 hours.

2.2.6.8 Size fractionation of cDNA

5 µl of 10 X SET (Section 2.1.6) was added to the digestion. A Sephacyl 400 spun column was used at 400 X g for 2 min. to size fractionate the products. This process was repeated two more times. One tenth of the eluant was run on a 5% non-denaturing polyacrylamide gel with radiolabeled marker as standard. The eluants were phenol/chloroform extracted and ethanol precipitated overnight at -20°C. The cDNA was pelleted at 14000 rpm at 4°C for an hour. The pellet was washed in 70 % ethanol, air dried. The pellet was dissolved in 3-10 µl of sterile water. A quantitative dot assay (Section 2.2.1.8) was done on all the fractions to estimate the cDNA concentration.

2.2.6.9 Ligating cDNA into Uni-ZAP XR vector arm

To the microcentrifuge tube these components were added in the following order: 0.5 µl of 10 X buffer 3, 0.5 µl of 10 mM rATP, 1.0 µl (1.0 µg) of vector arm, 100 ng of cDNA , water up to 4.5 µl and finally 0.5 µl (4 Weiss unit) of T4 DNA ligase. The ligation mix was mixed gently, centrifuged briefly and incubated for 16 hours at 12°C.

2.2.6.10 Packaging

The freeze thaw and sonic extracts were kept on dry ice. A µl of ligation mix was added to the melted freeze thaw extract. Gentle mixing was done and 15 µl of sonic extract was added to the tube. The contents were gently pipeted and mixed well. A quick spin was given and the tube was left at 22°C for exactly 2 hours. 500 µl of SM

(Section 2.1.6) and a drop of chloroform was added to the tube and the content was mixed and kept at 4°C for titering.

2.2.6.11 Plating and titering

A freshly grown XL1Blue MRF´ colony was inoculated in 50 ml of LB supplemented with 10 mM $MgSO_4$. The culture was grown at 37°C at 225 rpm in a shaker for 6 hours. The bacterial culture was pelleted at 2000 rpm for 10 min and the cells were suspended and diluted to a concentration of A600 to 0.5 in 10 mM sterile $MgSO_4$. These cells are stored at 4°C.

A serial dilution of packaged library was made with SM (Section 2.1.6) and 10 µl was taken and added to 200 µl of the bacteria. The tube was mixed and incubated at 37°C for 20 min for the phage to infect the host. The content was added to 3 ml of top agar at 45°C, mixed and plated on 2 day old bottom agar (Section 2.1.6) plates (90 mm). The inverted plate was incubated at 37°C for 8 hours and the efficiency of the library was calculated as follows:

$$\frac{\text{Number of plaques x dilution factor x total packaging volume (1000 µl)}}{\text{Total number of micrograms packaged x number of µl plated}}$$

2.2.7 Screening of the libraries and purification of recombinant phage

2.2.7.1 Plaque lifts

Plates were prepared as outlined in Section 2.2.6.11 Plates with a diameter of 130 mm were used in primary screens of a library and 90/130 mm diameter plates were

used for all subsequent rounds of screening. In the primary screens 5 x 10^4 plaque forming units (pfu)/plate were used. A much lower number, approximately 20-2000 pfu, were used for subsequent rounds.

Plates were chilled at 4°C for 30-60 minutes to allow the top layer of agar to harden before plaque lifts were attempted. Meanwhile, the appropriate number of nylon membranes (Hybond-N+, Amersham International plc., Amersham, UK) were labelled individually and marked asymmetrically in three locations with a permanent marker.

A filter was laid carefully onto each plate from the centre outward, so that it made direct contact with the plaques. The position of the reference points on the filter was recorded onto the bottom of the plate and after 60 seconds the first filter was peeled off the plate with a pair of blunt-ended forceps. When duplicate lifts were made, the second filter was left on the plate for a time duration of two minutes before peeling.

After its removal from the plate, each filter was placed with DNA side up, for four min onto two sheets of Whatman 3MM paper soaked in denaturation solution (0·5 N NaOH, 1·5 M NaCl). Then, the filters were carefully drained of excess solution and transferred to two sheets of 3MM paper soaked in neutralisation solution (0·5 M Tris-HCl; pH 7·5, 1·5 M NaCl) for a further four minutes. Finally, the filters were rinsed for five minutes in 2 x SSC and left, DNA side up, on Whatmann 3MM paper to dry.

All the nylon filters were alkali fixed with 0.05 N NaOH for 20 min, rinsed with 5 X SSPE and were air-dried.

Filters were hybridised in a solution which contained the appropriate [^{32}P]-labelled probe as described in Section 2.2.8.1.

2.2.7.2 Selection of bacteriophage λ plaques

When heterologous probes were used, filters were washed in 2 X SSPE, 0·1% SDS for five minutes at 50°C. At this stage, the filters were monitored, if more than 50 cpm were recorded on a Geiger counter the filters were washed for a further five minutes with a 2°C increase in temperature. Once, the cpm had fallen to about 20, filters were wrapped in clingfilm and autoradiographed at -70°C. In cases where the ratio of the background to specific signal was high, filters were washed again in 1 X SSPE, 0·1% SDS at 50°C for five to fifteen minutes. For homologous screening, the protocol from Sambrook *et al.* (1989) was followed.

For oligonucleotide screening of the library a quaternary alkylammonium (tetramethylammonium chloride) based hybridisation solution (Section 2.1.6) was used. The hybridisation was performed at 38°C for 16 h in a waterbath. the blots were washed with 2 X SSPE, 0·1% SDS for 5 min at RT as described above. Care was taken not to wash off the probe from the target by slow increase in temperature.

The pattern of dots produced on the film, by hybridisation of plaque DNA with a probe, were aligned with their corresponding position on the plate with the help of the reference marks on the filter and plate. Plaques of interest were removed from initial rounds of screening with the wide end of a pasteur pipette or a cut 1 ml disposable tip. In cases where individual plaques were well spaced apart, it was sufficient to use a sterile toothpick. The agar plug or toothpick was transferred to a 1·5 ml microfuge tube

which contained 1 ml of phage buffer (2.1.6) and 50 µl of chloroform, vortexed briefly and left at room temperature for one hour to allow the phage to diffuse out into the solution. The phage suspension was titered as described in Section 2-2.6.11. The process of screening was repeated with a lower plaque density per plate until plaque purity was attained. This typically took three to four rounds of screening.

2.2.7.3 Preparation of plate lysate stocks

An innoculum of 10^5 pfu was used to produce confluent lysis of a bacterial lawn grown on an 90 mm diameter plate. Phage were eluted from the agar by the addition of 5 ml of phage buffer to each plate, these were left on a rocker at 4°C overnight. The next day the bacteriophage suspension was recovered into a sterile 10 ml polypropylene tube and each plate was rinsed with a further 1 ml of phage buffer. A volume of 0.3% (v/v) chloroform was added to each suspension, vortexed and incubated at room temperature for 15 minutes. Bacterial cell debris was removed from the suspension by centrifugation at 3000 rpm in JA 20 (Beckman) for ten minutes. Then the supernatant was transferred to a fresh tube, chloroform was added to 0·3% (v/v) to prevent any bacterial growth and the lysate was titered. Phage lysates were stored at 4°C.

2.2.7.4 Purification of recombinant bacteriophage DNA

Magic lambda DNA purification kit from Promega (USA) was routinely used for recombinant phage DNA purification. The steps followed are briefly described here. Typically three 130 mm plates, each with 70,000 plaques are used for λ DNA preparation. Each plate was covered with 10 ml of SM and was shaken overnight in the

cold room. Next day the diffused phage were collected in a 50 ml tube and was spun at 5-8000 rpm for 10 min at 4°C in a JA 20 (Beckman, USA) rotor. The supernatant was collected in another fresh tube. 0.3% (v/v) chloroform was added to it and mixed gently. This lysate is now ready for either storage at 4°C or is used for DNA extraction. To get rid of bacterial nucleic acids 40 µl of nuclease mix (DNAse 10mg/ml and RNAse 10mg/ml) was added. The tube was incubated at 37°C for 30 min. For each 10 ml of nuclease-treated lysate, 4 ml of phage precipitant (Section 2.1.6) was added and mixed properly. The tube was incubated on ice for 30 min and spun at 10,000 rpm for 15 min at 4°C. The supernatant was discarded. The phage pellet was resuspended in 1 ml of phage buffer (Section 2.1.6).The microcentrifuge tube was spun for 1 min at 14000 rpm. The supernatant was collected in two microcentrifuge tubes. To each tube a ml of thoroughly resuspended purification resin was added. The mix was loaded to a 5 ml disposable syringe fitted to a Magic mini column (Promega, USA).The resin was gently pushed in to the mini column using a strile syringe and plunger. The syringe was detached and the plunger was removed. The empty syringe was again fixed to the mini column. The resin was washed with 3-5 ml of 80% isopropanol. A brief spin was given to the mini column to get rid of the residual isopropanol. The mini column was transferred to a new micro centrifuge tube and 100 µl of sterile hot water (80°C) was added to the resin. The assembly was immediately centrifuged at 14000 rpm for 1 min. The DNA was collected from the microfuge tube.

2.2.8 Transfer and detection of nucleic acids

2.2.8.1 Radiolabelling of DNA fragment

The Megaprime DNA labelling system from Amersham (RPN 1606) was used for the radioactive labelling of probes for hybridisation. 5 µl of primer solution

(containing random nonomer primers) and water to a final reaction volume of 50 µl were added to 50 ng of purified fragment DNA. The DNA was denatured for 5 min in a boiling water bath. Subsequently 10 µl of Megaprime reaction buffer (containing dATP, dGTP, dTTP, $MgCl_2$, 2-mercaptoethanol and Tris-HCl buffer [pH 7.5]), 5 µl [α-^{32}P]dCTP (3,000 Ci/mmole, 10 mCi/ml) and 2 µl Klenow enzyme (1 unit/µl) were added and the reaction incubated for 30 min at 37°C.

2.2.8.2 Removal of unincorporated nucleotides from radiolabelled DNA

Chromaspin-10 columns (Clontech Laboratories, Palo Alto, CA) were used to remove free radio label from radioactive labelled DNA as described in Section 2.2.1.3. The radioactively labelled sample was collected in an Eppendorf tube at the bottom of the column whereas the unincorporated nucleotides remained in the column. The specific activity of the radiolabelled sample was calculated by comparison of the radioactivity before and after removal of the free nucleotides.

2.2.8.3 Southern blot and hybridization

A combined vacuum/alkaline blot method was used to transfer DNA from agarose gels to nylon membranes. A vacuum blotting apparatus from Hybaid (Teddington, Middx.) was used to perform the alkaline Southern blot. A piece of Whatman 3MM filter paper and a piece of nylon membrane (Hybond N+, Amersham) were cut slightly bigger than the gel to be blotted and pre-wetted in 0.4 M NaOH. The paper was placed on top of the porous plate of the blotting device followed by the nylon membrane and followed by the rubber gasket, which was cut with an aperture

slightly smaller than the gel to be transferred. The agarose gel was briefly soaked in 0.4 M NaOH and then placed on the nylon membrane, avoiding trapping any air bubbles between gel and membrane. The lid was then placed onto the unit and tightened by turning the four levers into their locking positions. The vacuum was then applied and the transfer buffer (0.4 M NaOH) poured in the chamber so that the gel was completely immersed. The vacuum was reduced by a valve to 80 cm of water and the transfer time was 30 min. After the transfer was completed the nylon membrane was neutralised by washing in distilled water and 2 X SSPE buffer.

Southern blots were prehybridized in hybridisation buffer (Section 2.1.6) containing 100 µg/ml heat denatured sonicated salmon sperm DNA for 1 -2 hours at 65°C. The heat denatured probe (1 - 5 pg/ml hybridisation buffer, 10^8-10^9 cpm/µg) was added and incubated for 12 - 16 hours at 65°C in a shaking water bath. The filter was washed twice in 2 X SSPE, 0.1% SDS for 15 min at room temperature and subsequently twice in 0.1 X SSPE, 0.1% SDS at 65°C for 15 min The membrane was then autoradiographed for the desired time.

If the blot was subjected to a hybridisation with a second probe the membrane was not allowed to dry during or after hybridisation and washing. The blot was stripped of the first probe by pouring a boiling solution of 0.5% SDS on the membrane and allowing it to cool to RT.

2.2.8.4 Northern blotting and hybridisation

RNA from the formaldehyde gel was transferred by the capillary blotting method . The buffer used was 20 X SSPE. An overnight transfer to Hybond N was

routinely done. The blots were gently washed with 2 X SSPE and air-dried. RNA was fixed to the membrane by baking at 80°C for two hours. The hybridisation and washing protocols and conditions were identical to the Southern blot.

2.2.9 Generation of clones for sequencing by directed deletions

Erase-a-Base system from Promega was routinely used for generating nested deletion from one end of the cloned DNA. The basic steps are as follows.

2.2.9.1 Restriction digestion of plasmid DNA

Typically 5-7 μg of the plasmid was taken and restricted with two restriction enzymes from one end (either 5′ or 3′) of the insert. One of the enzyme produces a 5′ overhang and the other either a blunt end or a 3′ overhang. The 5′ overhang enzyme must be towards the vector thus protecting the primer binding site and the other enzyme must be close to the insert which it is desired to delete. A reaction volume of 50 μl was always preferred with two enzymes with compatible reaction buffer and incubation temperature. If the enzymes have different salt requirements then the lower salt buffer was added first and after the first digest the salt concentration was adjusted for the 2nd enzyme to work. After 2 to 3 hours incubation the digestion of the plasmid was checked on a gel. The rest of the digest was phenol chloroform extracted and ethanol precipitated. After a 70% ethanol wash the plasmid pellet was air-dried.

2.2.9.2 Exo III digestion, ligation and transformation

The plasmid pellet was thoroughly resuspended in Exo III 1 X buffer (Section 2.1.6). In the meantime, 7.5 µl of S1 nuclease mix (Section 2.1.6) was added to the appropriate numbers of tubes which were kept on ice. The DNA was warmed up to 37°C in a water bath and 500-600 unit of Exo III was mixed rapidly. After a 20 s lag period, 2.5 µl of the sample was pipetted out at 30 s intervals and mixed into the S1 nuclease mix tube. After all the samples were taken out, the tubes were left at RT for 30 min. 1 µl of S1 stop buffer was added to each tube. The S1 was heat inactivated at 70°C for 10 min. At this point the Exo III digestion was checked by electrophoresis of 2 µl of the above samples on 1% agarose TBE gel. To each sample 1 µl of Klenow mix (Section 2.1.6) was added. The samples were incubated for 5 min at 37°C. Then 1 µl of dNTP mix was added to each tube and incubation was continued for another 10 min. 40 µl of ligase mix was added to each of the tube and mixed. Ligation was continued at 16°C overnight. Next day, *E. coli* cells were transformed as described in the Section 2.2.1.17.

2.2.10 Cloning and characterisation of CWG islands

2.2.10.1 Isolation and cloning of PTFs

Routinely, 300-500 µg of CsCl purified pea nuclear DNA was digested to completion with *Pst*I restriction enzyme. The ethanol precipitated DNA was dissolved in TE and electrophoresed on a 1% agarose gel along with 100 bp ladder (Life Technologies Inc). A slit was made in the lane at the position of DNA of 90-100 bp and

a piece of DE81 (Whatman) paper (0.5 x 0.5 cm) was inserted. DNA fragments up to 350 bp were electrophoresed directly onto DE81 paper (as detected visually from the 100 bp ladder lane) which was transferred to a 1·5 ml microcentrifuge tube and the bound DNA fragments were eluted as described in 2.2.1.5a. The purified DNA was resuspended in 30 µl of TE and passed through a Chromaspin 30 (Clontech) column to remove any small particles of paper and remaining salt. The purified PTF fragments were phenol chloroform extracted and ethanol precipitated and cloned into *Pst*I digested, dephosphorylated pBlueScript KSII (+) vector. These plasmids were maintained in *E. coli* host XL1 Blue MRF′ (Stratagene, USA)

2.2.10.2 Purification and analysis of HLFs and PLFs

Typically 20-25 µg of CsCl purified DNA was digested to completion either with *Pst*I or *Hpa*II restriction enzyme and end labelled as described above. The ethanol precipited DNA was dissolved in TE buffer and separated on a 1% low melting point (LMP) agarose gel in TAE buffer (40 mM Tris-acetate, 1 mM EDTA, pH 8.0). The high molecular weight DNA band was cut from the gel and washed twice with β agarase buffer (10 mM Bis-Tris-HCl, 1 mM EDTA, pH 6.5). The slice was melted at 65° and incubated with 5-10 unit of β agarase (New England Biolabs) at 42°C for 10 to 16 hours. The DNA was isolated by phenol/chloroform extraction and ethanol precipitation. The *Hpa*II large fragments (HLFs) and *Pst*I large fragments (PLFs) were further digested with restriction enzymes and, in some cases, this second digestion was followed by relabelling of the fragments to produce, so-called, double labelled fragments. In order to check complete digestion an internal or parallel control of plasmid DNA was used. After electrophoresis, the quantification of labelled DNA was done using a Fujix BAS 1000 phosphoimager.

2.2.10.3 *In vitro* methylation of reporter gene constructs and transient gene expression assays

Prokaryotic DNA methyltransferases were used according to the manufacturers' instructions. The pea CWG methyltransferase was purified as previously described (Pradhan and Adams, 1995) and used to methylate plasmid DNA by overnight incubation at 30°C in the presence of 0.5 mM AdoMet. Methylation of plasmids were checked with corresponding restriction enzymes. The DNA was repurified and the amounts quantified using a microfluorometer (Hoefer Scientific Instruments) prior to use in transfection.

2.2.10.3 Plant growth and protoplast isolation

Phaseolus vulgaris L cv. tendergreen plants were grown for 10-15 days in continuous white light provided by warm white fluorescent tubes at a fluence rate of 10 μmol m^{-2} s^{-1}. Plants were transferred to dark for 2 days (dark adapted). All further steps, prior to culture of protoplasts were performed under a green safe light. The upper surfaces of primary leaves were rubbed with alumina which was subsequently removed by washing thoroughly in distilled water. Leaves were floated, (upper side down) on 0.8% cellulase R10, 0.25% macerozyme R10, (both from Yalkult Honsha, Nishinomiya, Japan) in culture buffer, (0.4 M mannitol with CPW salts – Power & Chapman, 1985) and 50 μg ml^{-1} ampicillin, pH 5.7) for 20 h at 22°C in the dark. Protoplasts were released from leaves by brief, gentle agitation and cells were filtered though one layer of muslin and a steel sieve, (70 μm pore size). Protoplasts were

allowed to pellet under gravity on ice for 1–1.5 hours after which they were washed once or twice in cold 0.4 M mannitol, CPW salts pH 5.7.

2.2.10.4 Electroporation and culture of protoplasts

Protoplasts were resuspended in electroporation buffer (0.4 M mannitol, 4 mM MES pH 5.7, 20 mM KCl) at 1-3 x 10^6 cells/ml. Aliquots of 0.5 ml were dispensed into cuvettes on ice and 25 µg of plasmid DNA were added in less than 30 µl. Cells were electroporated with a single pulse at a capacitance of 100 µF and a field strength of 325 V/cm. Cells were returned to ice for 5 min. Aliquots of 200 µl were added to 800 µl of culture buffer (0.4 M mannitol, CPW salts, 50 µg ml^{-1} ampicillin, pH 5.7). Cells were cultured in 24 well culture plates for 20 h in white light as above at a fluence rate of 80 µmol m^{-2} s^{-1}. Protoplasts were pelleted at 100 g for 5 min and were resuspended in 50 µl of 20 mM Tris-HCl, pH 8.0, 2 mM $MgCl_2$. After freezing at –80° and heating at 65°C for 15 min, debris was removed by centrifugation at 14,000 *g* for 5 min. CAT activity was assayed by the method of Seed and Sheen (1988) using 40 µl of extract for all constructs apart from 35SCAT where 40 µl of a 1/100 dilution of the extract was used. The electroporation was carried out by Dr. Nigel Urwin.

2.2.10.4 Chimaeric promoter CAT constructs

pCAMVCN (Pharmacia) is a 35S-CAT construct containing a 400 bp cauliflower mosaic virus *35S* promoter fused to the coding region of the chloramphenicol acetyltransferase gene (CAT) of Tn9 and the termination sequence of the nopaline synthase gene (NOS).

An *rbcS-CAT* construct containing a promoter fragment (-1433 bp to +26 bp) generated from an *Eco*RI–*Kpn*I digest of the *rbcS2* promoter was cloned into pUC19. The promoter was re-isolated as an *Eco*RI–*Bam*HI fragment, which was then cloned into pBluescript SKII-. The *Xba*I–*Bam*HI (CAT-NOS) fragment of pCAMVCN was then cloned into this to give pRBCSCAT2.

Both the constructs were made and supplied by Drs. Nigel Urwin and Garath. I. Jenkins, IBLS, University of Glasgow.

2.2.11 Construction of genomic library and isolation of m5C MTase gene promoter

P. sativum genomic high molecular genomic DNA was prepared by the CTAB method (2.2.5.1) and a genomic library of partial *Sau*3AI (0.01 u/µg of DNA at 37°C for 0.5 h) fragments (partially filled with dATP and dGTP) were ligated to *Xho* I digested partially filled λFIX (Stratagene) vector arms. The ligation, packaging and titering was done as described in Section 2.2.6.9-11.

The 5´ end of the pea MTase cDNA (the 825 bp *Eco*RI-*Eco*RI fragment of clone RP23-1) was used to screen a library of 3.5 x 10^6 primary plaques by hybridisation according to Sambrook *et al* (1989). Clone F 3.1 was digested with *Kpn*I-*Xba*I to give a 3.5 kbp fragment which was cloned in to *Kpn*I-*Xba*I digested pGEM 7Zf (+).

2.2.12 N terminal sequencing of proteins

2.2.12.1 Preparation of the gel and protein separation

Proteins for sequencing were separated using SDS polyacrylamide gel electrophoresis as described in Applied Biosystem user bulletin no. 25 (1986). The lower tris (4X) was made by dissolving 36.34 g of tris and 0.8 g of SDS in 150 ml of water. The solution was titrated to pH 8.8 with 6N HCl and the final volume was adjusted to 200 ml. The upper Tris (4X) was made by dissolving 12.11 g of Tris and 0.8 g of SDS in 150 ml of water. The pH was titrated to 6.8 with 6N HCl and the final volume was adjusted to 200 ml. Easy gel, acrylamide gel solution (29:1) was purchased from Scotlab, UK. A 7.0% resolving and 3.6% stacking gel was prepared using lower Tris 4X as described in Section 2 2.1.10. The gel was allowed to age overnight at 4°C. The upper electrode buffer was upper Tris 4X with 0.05 mM glutathione. A pre-run was done at 15-45 milliamp constant current with only the dye till the dye front was about to touch the resolving gel. At this point the upper chamber buffer was replaced with Tris-gly-SDS buffer (Section 2.1.6) with 7 mM sodium thio glycolate. The samples were boiled with equal volume of SDS sample loading buffer (Section 2.1.6) for 3 min and loaded onto the gel. Rainbow protein molecular wt markers (2.1.6) were used along side. The electrophoresis was resumed till the dye front moved to the bottom of the gel.

2.2.12.2 Blotting, staining and sequencing of the protein

Following a modification of the method of Matsuda (1987) it is possible to electroblot proteins/peptides onto PVDF membrane. The blotting was carried out in

CAPS buffer (Section 2.1.6). The PVDF membrane was pre-treated in 100% methanol for few seconds followed by a few seconds in 50% methanol/water and finally in electroblotting buffer. The gel was soaked for 15 min in blotting buffer with two changes of 250 ml of buffer. The blotting was carried out at RT for 2 hrs at 50V. After the electroblotting the proteins could be stained with Coommassie Blue R-250 (0.1% Coommassie Blue R-250 in 50% methanol) for 2 min with several changes of destain (50% v/v of methanol, 10% v/v of acetic acid) till the background is clear. The blot was rinsed with several changes of distilled water and air dried. The protein bands were cut and sequenced at the SERC protein sequencing facility, Aberdeen using a ABI 477A protein sequencer.

2.2.13 Automated diedeoxy sequencing of the plasmid DNA

2.2.13.1 Sequencing reaction

Plasmids are grown in XL1Blue MRF' host with ampicillin and tetracycline. Plasmid DNA was purified either by CsCl-EtBr gradient centrifugation or using a Qiagen tip. The Taq dyedeoxy terminator sequencing kit used was from Applied Biosystem Inc (USA, cat no 401113). Accurately 1-1.2 µg of plasmid DNA was used for each sequencing reaction. A typical sequencing reaction consists of 1.0 µg of the plasmid DNA, 5 picomole of primer (universal or reverse), 9.5 µl of dyedeoxy reaction pre mix (Section 2.1.6) and sterile water up to a total volume of 20 µl. Each reaction mix was covered with 100 µl of mineral oil. The cycle sequencing reaction was carried out in a Perkin-Elmer Cetus thermal cycler model 480 with the following cycling parameters:

Programme 1: Rapid thermal ramp to 96°C for 2 minutes

Programme 2: * 96 °C for 30 seconds

* Rapid thermal ramp to 50°C

* 50°C for 15 seconds

* Rapid thermal ramp to 60°C

* 60°C for four minutes

25 cycles total

Programme 3: Soak at 4°C

After the cycling reaction was over the products were purified by a modified phenol-chloroform extraction and ethanol precipitation.

2.2.13.2 Purification of the extention products

To the 20 µl of the extention products 80 µl of sterile water was added. The products were removed to a new tube without the mineral oil. 100 µl of phenol chloroform mix (ABI, USA) was added to each tube and vertexed. After a centrifugation at 14000 rpm the lower organic phage was discarded. Another 100 µl of the phenol/chloroform mix was added and the above process repeated. The clear aqueous phase was taken to a new tube and DNA precipitation was done with 1/10 vol of sodium acetate pH 5.2 and 2.5 vol of ethanol. After a spin at 14000 rpm for 10 min the pellet was washed in 70% ethanol and air-dried. The air dried pellet was kept at -70 °C until use.

2.2.13.3 Sequencing of extention products in ABI 373A automated sequencer

Sequagel-6 from National Diagnostics, USA (cat no: EC - 836) was routinely used to cast the gel for DNA sequencing. The ratio of buffer reagent to monomer was 2:8. The gel was casted just after the addition of 0.8 ml of 10% ammonium persulphate per each 100 ml of the gel. After two hours the gel was pre-run for an hour with 1 X TBE (2.1.6). The samples were mixed with 3 µl of formamide and denatured at 90°C for two minutes. The samples were quick chilled on ice. First the odd number wells were loaded and after a 5 minutes run the even number wells were loaded. The sequencing data was collected and stored in the work station.

2.2.14 Sequence analysis and assembly

2.2.14.1 Use of GCG package

Genetic Computer Group (GCG), Wisconsin sequence analysis package version 8.1 was used routinely for analysis of the sequence information. The generated sequences from both sense and anti-sense strand were fed by using Seqed. The contigs were established by GelAssemble. The programmes used for protein translation were Pepdata and Translate. Profile search was done by Motifs. Comparison of the sequences were done by Compare, DotPlot, BestFit and Gap. WordSearch was used for data base search. CG and CNG frequency was calculated by composition programme.

CHAPTER THREE

Purification and characterisation of cytosine DNA methyltransferases from pea

3.1 Introduction

The DNA of higher eukaryotes is methylated at carbon 5 of some cytosine residues. In vertebrates, 3 to 8% of cytosine residues are methylated (Shapiro, 1975), whereas in plants as many as 30% of the total cytosines are methylated (Adams and Burdon, 1985). The higher content of methylated cytosine in some plants could be partly attributed to the large genome which contains many repetitive DNA sequences. However, in the vertebrate genome 5-methylcytosine (5mC) is largely confined to CG dinucleotides, whereas in higher plants both CG dinucleotides and CNG trinucleotides are methylated (Gruenbaum *et al.*, 1981). In non-vascular plants, methylation appears to occur only at CNG trinucleotides (Belanger and Hepburn, 1990). Methylation of DNA in plants, as in vertebrates, is implicated in the regulation of gene expression (Finnegan *et al.*, 1993; Antequera and Bird, 1988); an effect that may be direct, through DNA:transcription factor interaction, or indirect via an alteration in chromatin structure (Razin and Cedar,1991; Lewis and Bird, 1991; Adams, 1990).

Plant methylases have been partially purified from pea (Yesufu *et al,* 1991), wheat (Theiss *et al,* 1987) and rice (Giordano *et al,* 1991). The wheat and rice enzymes have an

apparent molecular weight of 55 kDa, with a possible breakdown product of 35 kDa for wheat methylase. The predicted translated sequence of the *Arabidopsis* enzyme encodes 1534 amino acids which would correspond to a protein of at least 150 kDa.

Sequence specificity of methylation in plants has been investigated *in vitro* for wheat, rice and pea by using partially purified enzymes with various DNA substrates. In wheat, 80% of C-G sequences, 60% of C-A and C-T sequences and 25 % of C-C sequences are found to be methylated (Theiss *et al*, 1987). Similar results were obtained for pea methyltransferase (Yesufu *et al*, 1991). Nearest-neighbour analysis of wheat germ DNA gives an estimation of 82% methylation for C-G and >80% methylation of CAG and CTG sequences confirming the authenticity of CG and CNG methylation in plants (Gruenbaum *et al.*,1981). However, information about methylation of both di and trinucleotide sequences is incomplete.

Using synthetic oligonucleotide substrates, Houlston *et al* (1993) have indicated the possible existence of two different DNA methylases in pea and the affinity of the enzymes varied dramatically for the different substrates. Application of total crude methyltransferase to Q-Sepharose led to partial fractionation of activities as determined using substrates containing di- or trinucleotide target sites. Moreover, the relative activities with these substrates changed during purification, plant growth and on preincubating the enzyme at 35°C. Bezdek *et al* (1992) have presented evidence for the differential drug sensitivity of CG and CCG methylation in repetitive DNA sequences of *Nicotiana* which might imply the presence of more than one DNA methyltransferase in plants.

Refinements of previous purification methods (Yesufu *et al*, 1991, Houlston *et al*, 1993) have made possible the complete separation of two DNA methylases from pea shoot

tips. The separated enzymes have been used to methylate synthetic DNA substrates *in vitro* and their target sequences were identified by Maxam-Gilbert chemical cleavage methods. One activity is specific for the dinucleotide 5'-CG-3' and the other methylates the trinucleotides 5'-CAG-3' and 5'-CTG-3'. Both the CG and CAG/CTG enzyme can perform both maintenance and *de novo* methylation *in vitro*, though the latter shows a strong preference for hemimethylated DNA. We have found no evidence for an enzyme methylating the 5´ cytosine of the 5´-CCG-3´ sequence but the CG enzyme does interact weakly with such sites.

3.2 Separation of two different methyltransferase activities from extracts of pea nuclei

For routine assays we have used either (poly dI-dC).(poly dI-dC) or a double stranded oligonucleotide substrate formed by annealing (CAG)7 with (MTG)7. These two substrates allowed us to separately study CG and CNG methylation. Previous studies using crude, nuclear extracts (Houlston *et al*, 1993) have shown the relative activities of *de novo* versus maintenance methylation with such di and trinucleotide target sequences and this aspect of the work is extended later in this chapter. The solubilised methyltransferase activity binds to heparin-Sepharose and elutes in a single activity peak at higher salt concentrations (Houlston *et al*, 1993). Similar results are seen using pea DNA affinity columns. However, fractionation of this activity on a MonoQ column results in the separation of two different activities: one using (poly dI-dC).(poly dI-dC) and the other using the CNG substrate (Fig. 3.1). The NaCl concentrations for elution of the CG and CNG enzymes were 360 and 400 mM respectively. Figure 3.2 shows the SDS/PAGE of the active fractions from the MonoQ gradient. Although several protein bands are obvious, the doublet at 110 and 100 kDa varies directly with the CNG enzymic activity (Fig. 3.2b).

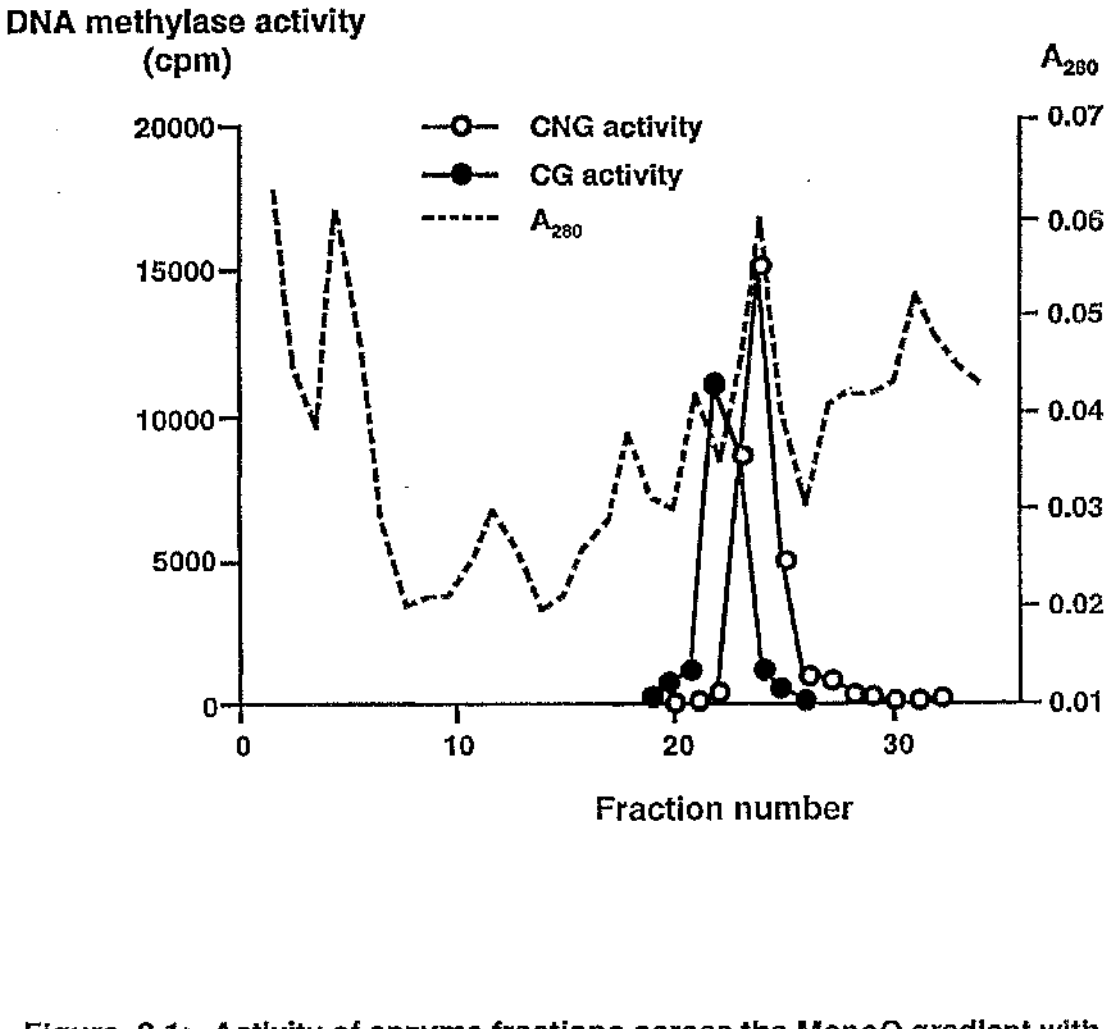


Figure. 3.1: Activity of enzyme fractions across the MonoQ gradient with different substrates

Partially purified enzyme from the pea DNA affinity fractionation was loaded onto an FPLC MonoQ column and the dialysed fractions assayed with either (poly dI-dC).(poly dI-dC) (●) or the hemimethylated CNG substrate (o). The dotted line indicate the protein absorbance at A280.

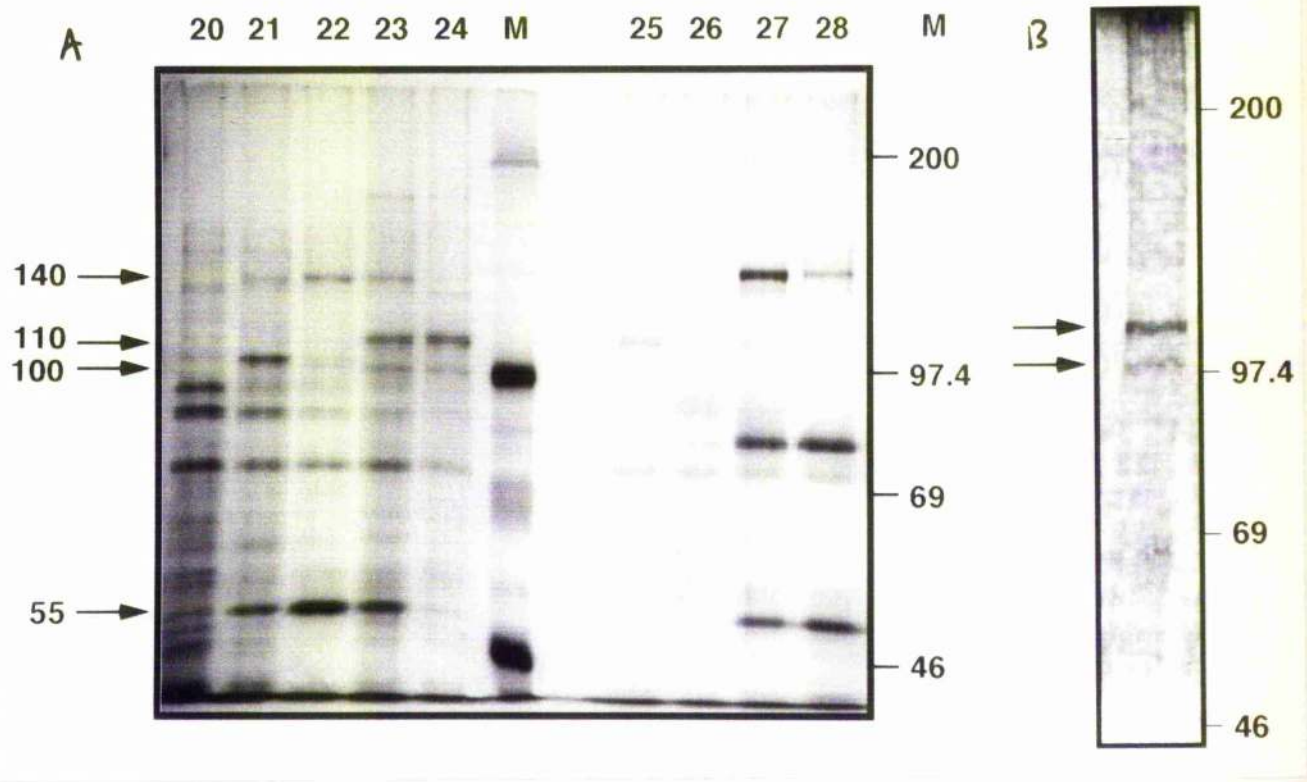


Figure. 3.2: Silver stained gel of fractions of different DNA methyltransferases from the MonoQ column

(A) Silver stained SDS/7%-PAGE of fractions from the MonoQ column (fraction numbers across the top). M is the set of Rainbow markers (Amersham) and the size (in kDa) is indicated down the right hand side. The size of relevant bands is indicated (in kDa) down the left hand side. Figure (B) shows a similar gel of the final, purified fraction from the CNG affinity column (see Table 1).

We attribute these proteins to this enzyme, and propose that the smaller arises from partial proteolysis of the former. The CG peak fraction corresponds to proteins of 140, 70 and 55 kDa similar to those observed by Yesufu *et al* (1991) and Houlston *et al* (1993).

The CNG enzyme bound successfully to a CNG-Sepharose affinity column and was eluted at 0.8 M NaCl. This resulted in a 142 fold purification of the enzyme (Table-1). The pure protein was a doublet with molecular mass 110 and 100 kDa (Fig.3.2b).

Table1 Purification of CNG methyltransferase from pea

Fraction	Total activity (units)	Protein (mg)	Specific activity (units/mg)	Purification (fold)
Nuclear extract	4970	47	105.7	1
Heparin Sepharose	4703	14.3	329	3.1
Q Sepharose (fractions 2&3)	3168	2.8	1131	10.7
Pea DNA affinity column	1291	0.78	1655	15.7
MonoQ (fractions 22&23)	600	0.12	5000	47.3
CNG affinity column	300	0.02	15000	142

Gel filtration of the purified enzymes from the MonoQ fractionation shows that the CG enzyme elutes with a native size of 150 kDa whereas the CNG enzyme elutes with a native size of 120 kDa (Fig. 3.3). With the CG enzyme, fractions eluted from the gel filtration column again show the presence of bands of 140, 70 and 55 kDa (Fig. 3.4).

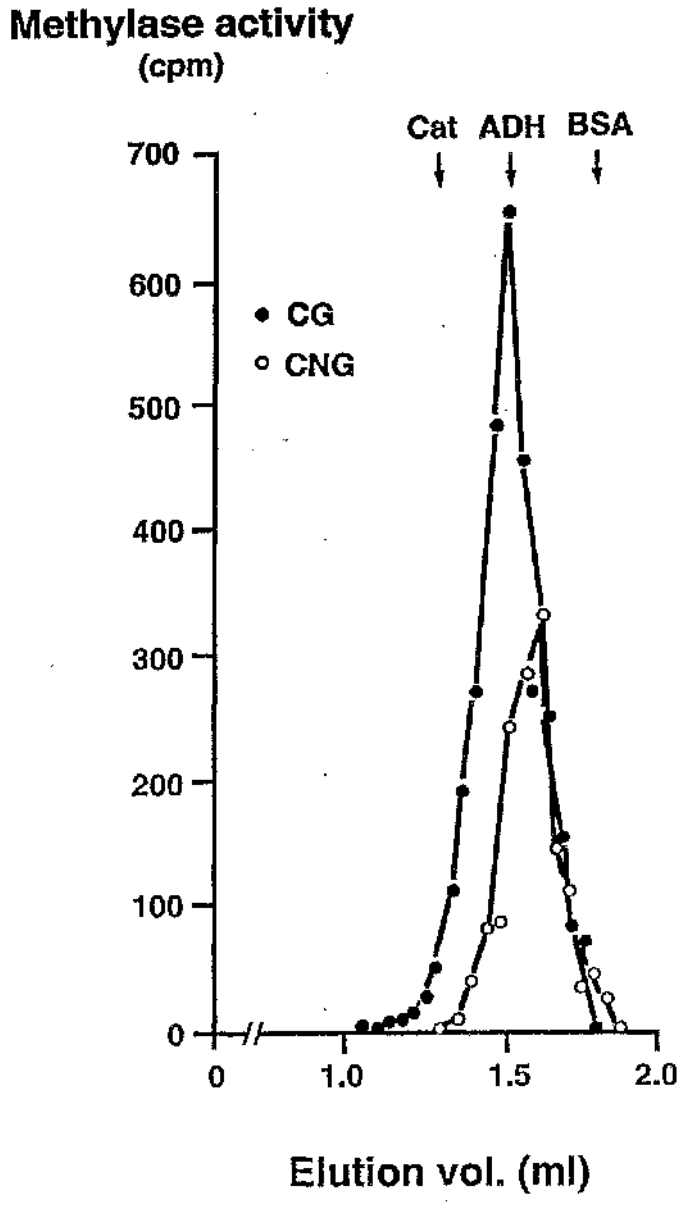


Figure. 3.3: Gel filtration profile of MonoQ fractionated enzyme

Samples (50 µl) from the two peak fractions from the MonoQ column were loaded separately onto a small gel filtration column. Single drop fractions were collected and assayed for enzymic activity with either (poly dI-dC).(poly dI-dC) (•) or the hemimethylated CNG substrate (o).

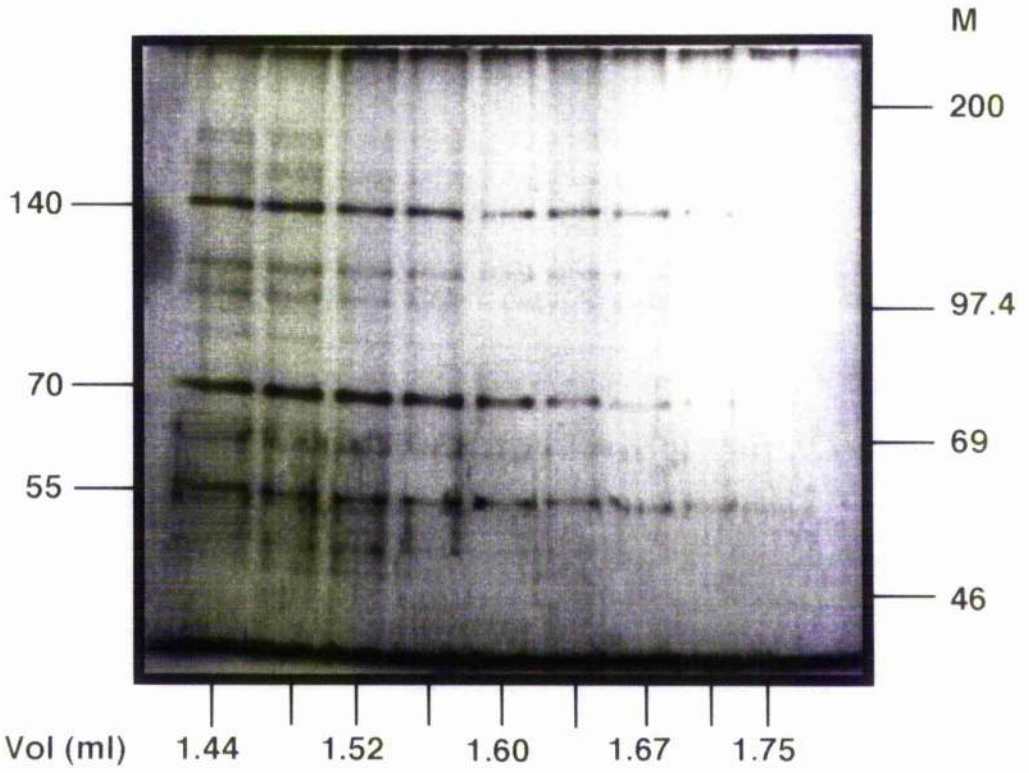


Figure. 3.4: Silver stained gel of the CG enzyme fractions

Silver stained SDS/7%-PAGE of fractions from the gel filtration of the CG enzyme (fraction numbers across the top). M is the set of Rainbow markers (Amersham) and the size (in kDa) is indicated down the right hand side. The size of relevant bands is indicated (in kDa) down the left hand side.

It is possible that the smaller proteins are derived from some larger protein either by proteolysis or as a result of dissociation of the subunits of a polymeric protein. However, from these data we cannot definitively conclude which bands correspond to the CG enzyme.

3.3 Cross linking of the enzymes to their substrates

The peak fractions from the MonoQ separation were uv-crosslinked to the CNG substrate that had been end-labelled with ^{32}P on the unmethylated strand. These fractions were also fractionated by SDS/PAGE and showed the presence of radiolabelled bands of

110 and 100 kDa in the fractions corresponding to the peak CNG enzyme activity (Fig. 3.5A).

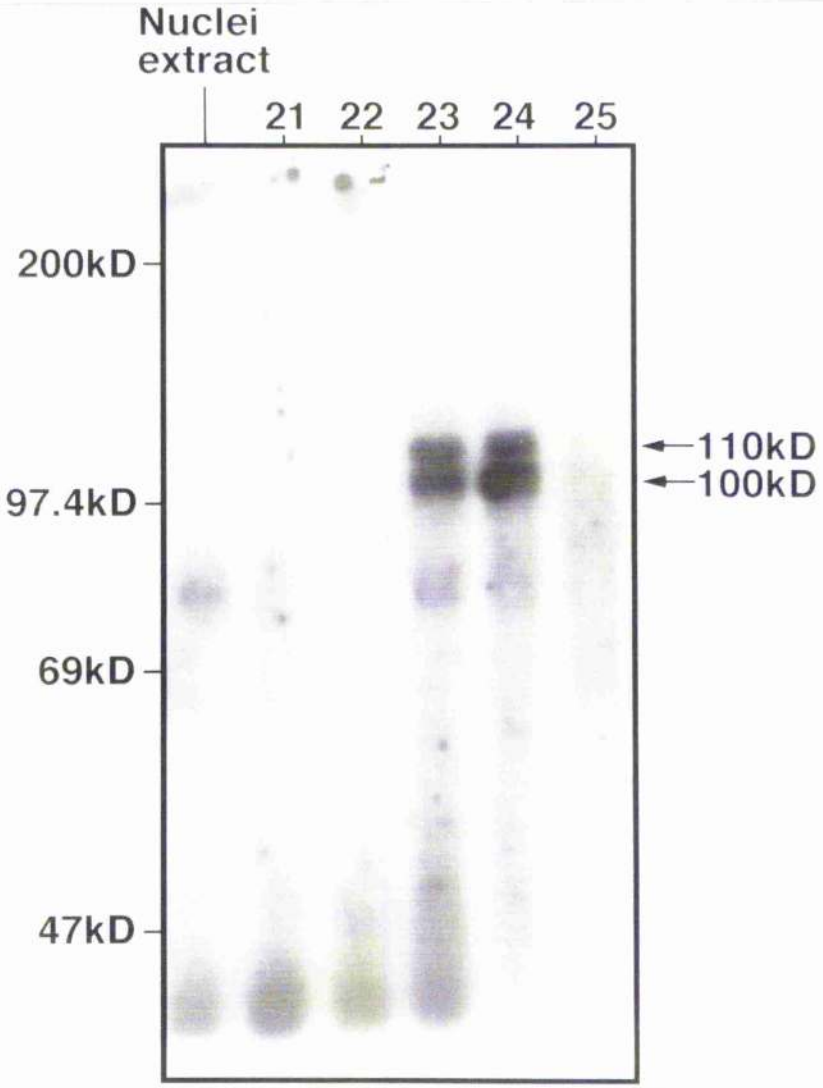


Figure 3.5 A: UV crosslinking to hemimethylated CNG substrate

Fractions (30 µl), collected from across the FPLC MonoQ column (numbers along the top) were cross-linked to ^{32}P-end labelled $(CAG)_7(MTG)_7$ as described in the experimental Section. After separation of unbound oligonucleotides the fractions were separated by SDS/PAGE. The size (in kDa) of marker proteins is indicated down the left hand side and the size of relevant bands is indicated (in kDa) down the right hand side.

Gel filtration of this crosslinked fraction shows that the ^{32}P radioactivity coelutes with the CNG activity at a molecular mass of about 120 kDa (results not shown).

These results support the contention that the CNG enzyme is a monomer of 110 kDa.

No crosslinked material is seen in fractions 21 to 23 corresponding to the bands of 140, 70 or 55 kDa indicating that the CG enzyme does not crosslink to the CNG substrate.(Fig. 3.5A). Crosslinking to a CG substrate was not successful probably because the target site was over 10 bp from the labelled end of the molecule and the latter would be removed by the DNase treatment (see methods Section 2.2.3).

Therefore an attempt to covalently bind the CG enzyme to substrates containing 2-pyrimidinone-1-β-D-2′-deoxyribofuranoside ($d^{4H}C$) in place of deoxycytidine was undertaken (Taylor *et al*, 1993; Ford *et al*, 1993). This analogue reacts with the enzyme to form a covalent adduct and so can be used as an affinity detection reagent. Substrates were used that contained the sequence CCCG in which the first (oligo1), second (oligo 2) or third (oligo 3) cytosine had been replaced with $d^{4H}C$. Using the peak fraction containing the CG enzyme, bands are seen with oligos 2 and 3 but not with oligo 1 (Fig. 3.5B) indicating that the enzyme interacts with $d^{4H}CG$ and, less strongly, with $d^{4H}CCG$ target sequences but not with the sequence $d^{4H}CCCG$. Whether the cofactor present is AdoMet or AdoHcy appears to make little difference.

The major band observed is at about 70 kDa with lesser bands of higher molecular weight. As the bound oligonucleotide (which, in this case was not trimmed by DNase treatment) will contribute significantly to this value, the major band almost certainly arises

from the protein of 55 kDa. However, in view of the known instability of the enzyme and the presence of more slowly migrating bands, we consider that the bands observed represent binding to proteolytic fragments and that the 140 kDa band represents the intact enzyme.

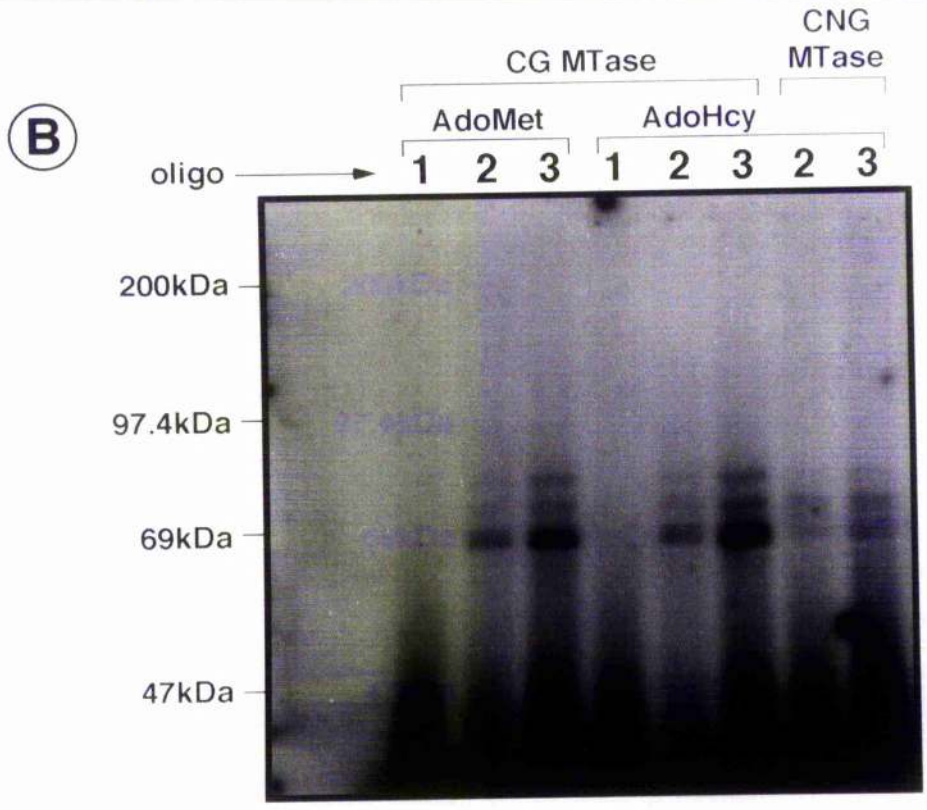


Figure 3.5 B: Covalent crosslinking to affinity substrate

Samples from the CG enzyme peak (8 units) or from the CNG peak (160 units) taken from the MonoQ column were crosslinked to oligonucleotides containing 2-pyrimidinone-1-b-D-2′-deoxyribofuranoside in place of deoxycytidine as described in the text and in the methods Section. The substituent ($d^{4H}C$) was in position 1, 2 or 3 of the CCCG target site as indicated and the incubation was for 2 h in the presence of AdoHcy (300 mM) or AdoMet (500 mM) as indicated. Separation was on an SDS:7.5% polyacrylamide gel and the size (in kDa) of marker proteins is indicated down the left hand side.

Attempts are underway to obtain enzyme preparations in which little or no breakdown has occurred and to study any changes in binding and specificity that might accompany degradation. The 70 kDa material is an impurity (possibly keratin) present in all samples whether they possess enzymic activity or not.

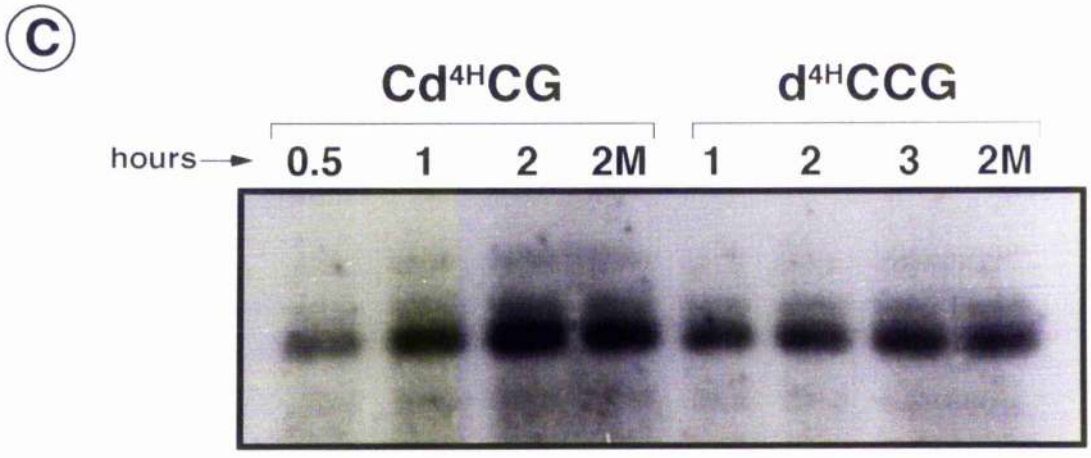


Figure 3.5 C: Time course of crosslinking

Samples as in (B) above were incubated with oligonucleotide containing $d^{4H}C$ at position 2 or 3 for various time from 30 min to 3 h and then separated on SDS PAGE. The incubation was in the presence of AdoHcy with the exception of lanes labelled M where AdoMet was used. The figure only shows the band at about 70 kDa.

Although I have not been able to detect any enzymic methylation of the 5´C in CCG target sequences, the enzyme clearly does react with such target sites, albeit weakly. However, in contrast to the binding to $d^{4H}CG$ targets which increases steadily with incubation time, binding to $d^{4H}CCG$ does not change with time of incubation. This indicates that binding, although rapid, is readily reversible and hence the complex formed is of low stability (Fig 3.5C). This may explain the failure of the enzyme to transfer a methyl group to the 5´C in CCG sequences, but whether this is a consequence of partial degradation or is also a property of the undegraded enzyme remains to be resolved.

Similar crosslinking studies were done with the CNG enzyme (Fig. 3.5B). Using a 20-fold greater amount of enzyme than was used in the experiments with the CG enzyme, no bands were obtained of 110 kDa or greater. The only bands observed were produced by a small amount of contaminating CG enzyme. This shows that the CNG enzyme fails to interact with CG and CCG target sequences to form a stable complex, although it can be uv crosslinked to CAG/CTG target sequences (Fig. 3.5C).

3.4 Sequence specificity of DNA methyltransferase

To confirm the target site specificity of the CG and CNG methyltransferases we designed two substrates with all possible combinations of hemimethylated and unmethylated di- and trinucleotide target sites. The unmethylated strands were ^{32}P labelled at the 5' end prior to annealing with the complementary methylated strand to give the partially hemimethylated substrate (see Figs 3.6A and 3.6B). Methylation was for an 8 h period and Fig 3.7 shows a trial time course of the reaction for the two enzymes with the unlabelled substrate. With the CNG enzyme, the rate drops after 2 h at a point commensurate with the total methylation of all the hemimethylated 5'-CAG-3' and 5'-CTG-

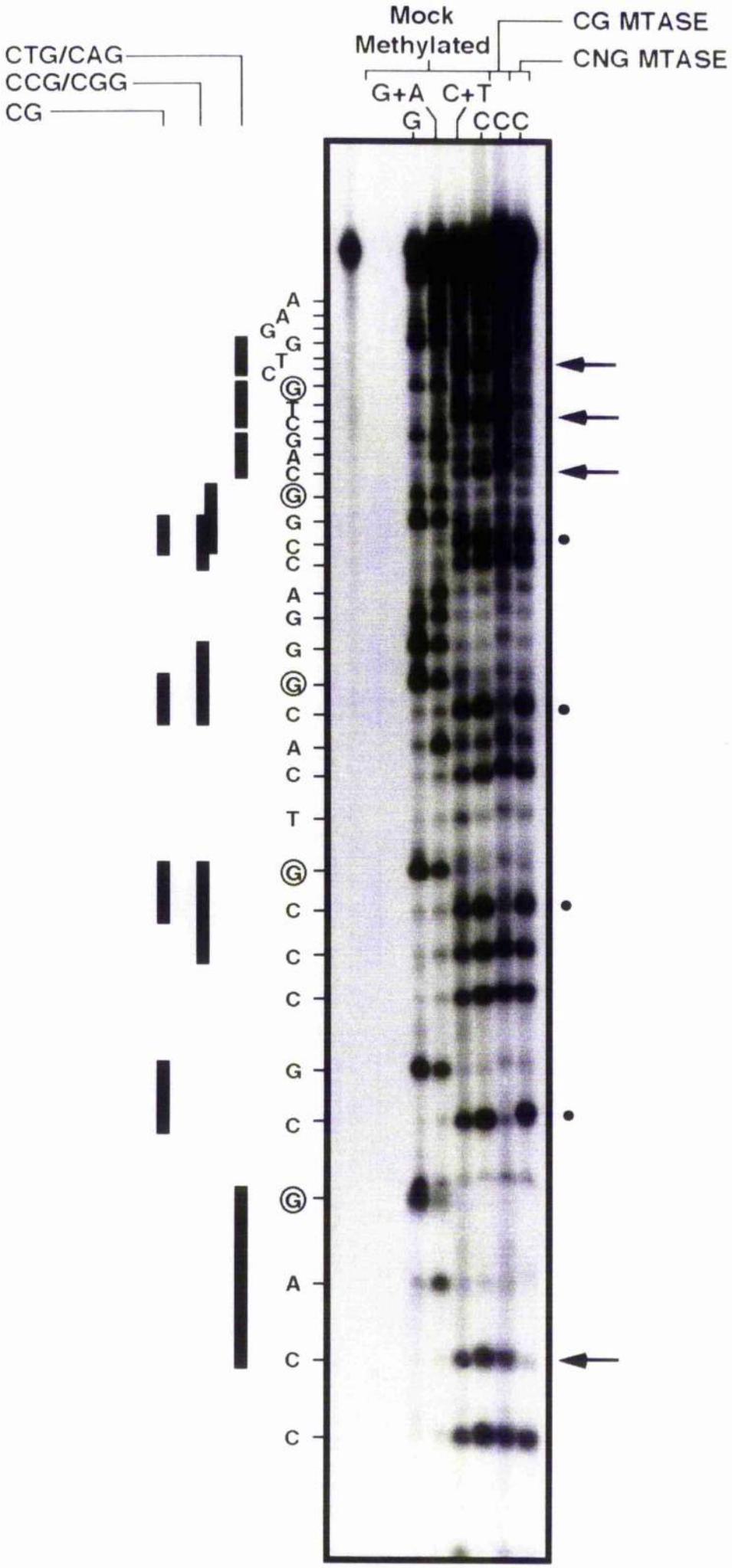
3' sites. Thereafter, methylation continues at a slower rate. With the CG enzyme, methylation was linear over the 8 h period but incorporation reached only a third of the maximum possible. For this reason more CG enzyme was used in the experiment with the ^{32}P-labelled substrate.

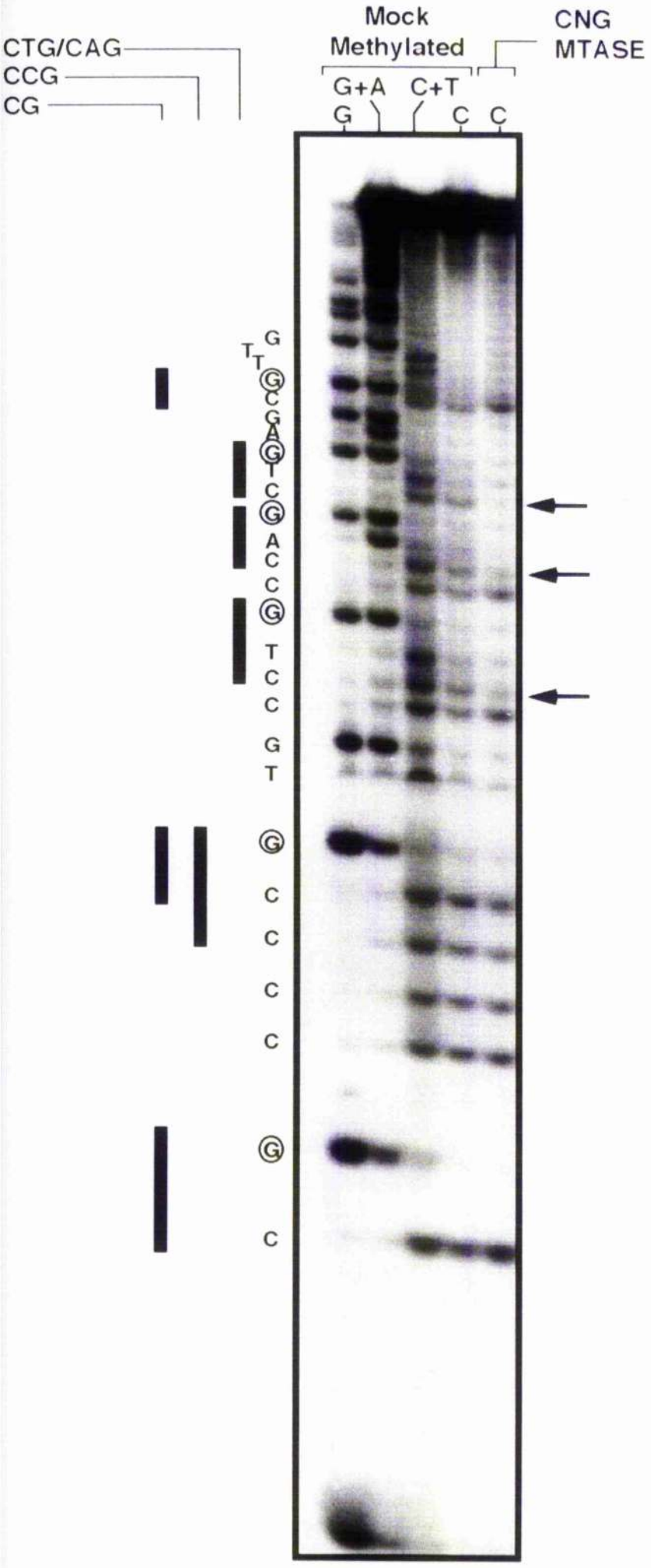
Figure 3.6: Maxam-Gilbert chemical sequencing of *in vitro* methylated oligo 42mer (Fig 3.6a) or 41mer (Fig 3.6b) duplexes

The sequences and sites of methylation are given in the experimental Section and down the left hand sides of the figures. G in a circle indicates that it is paired with a methylcytosine. The bars down the left hand side indicate potential target sequences. The lanes in figure 3.6a show unreacted substrate; mock methylated substrate after the G, G+A, C+T and C reaction and the C reaction on the oligonucleotide after methylation with the CG or the CNG methylases from the peak fractions from the MonoQ gradient. Arrows indicate sites of CNG methylation and dots the sites of CG methylation. Figure 3.6b shows results from a similar experiment where only the CNG enzyme was used.

Please see page no 98 for Figure 3.6a

Please see page no 99 for Figure 3.6b


Mock
Methylated
CG MTASE
CNG MTASE
G+A
C+T
G
CCC
CTG/CAG
CCG/CGG
CG


CTG/CAG
CCG
CG
Mock
Methylated
CNG
MTASE
G+A
C+T
G
C
C

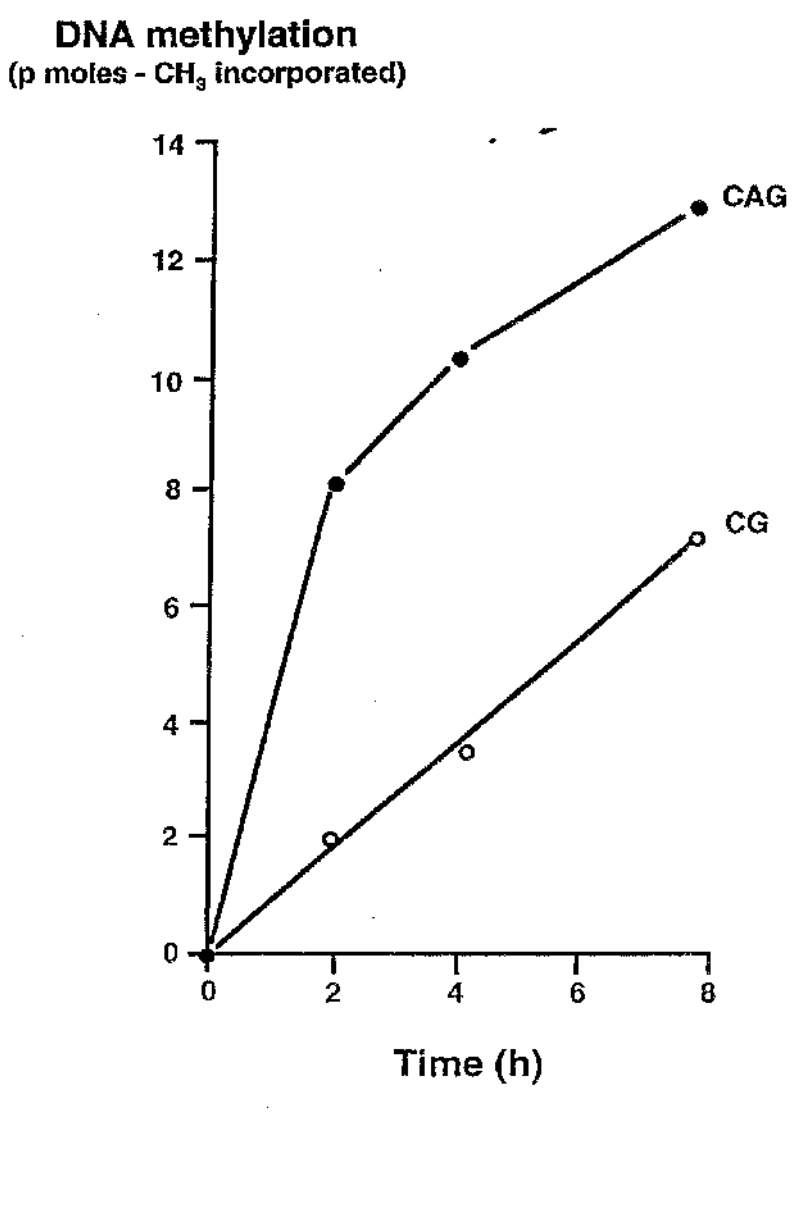


Figure 3.7: Time course of *in vitro* methylation of the 42mer oligonucleotide duplex

Oligonucleotide (100 ng) with enzyme (10 µl) from the peak fractions (CG or CNG enzyme) from the MonoQ separation.

After methylation with the two separate enzymes, chemical cleavage using the technique of Maxam and Gilbert (1980) allowed us to recognize the sequences that became

methylated *in vitro*. The missing bands in the C reaction track represent the sites of methylation. The CG enzyme was able to methylate all the 5´-CG-3´ sites whether hemimethylated or unmethylated (Fig. 3.6a). However, the CNG enzyme methylated only 5'-CAG-3' and 5'-CTG-3' and methylation of unmethylated sites was less efficient than for hemimethylated sites (Figs 3.6a and 3.6b). The strong preference of the CNG enzyme for hemimethylated substrates has already been recorded (Houlston *et al*, 1993) and is further confirmed by these results and the time course shown in Fig. 3.7. No methylation by the CNG enzyme of 5'-CCG-3' or 5'-CGG-3' sites was seen, whether these were unmethylated or hemimethylated and no methylation of 5'-CCG-3' sites could be detected with the CG enzyme.

3.5 Lack of evidence for 5´-CCG-3´ methylation in pea DNA

In order to explain our failure to find an enzyme that can methylate the 5´-cytosine in the sequence 5´-CCG-3´ we reassessed the evidence for the existence of methylcytosine in this position. The evidence for methylation of the 5´ cytosine in the CCG sequence is primarily based on the resistance of genomic DNA to cleavage by the *Msp*I restriction enzyme. *Msp*I digestion is inhibited by methylation of the 5´cytosine in the sequence 5´-CCGG-3´.

Figure 3.8 shows the digestion of pea genomic DNA with a variety of restriction enzymes. *Msp*I digestion gives fragments in the size range 200 to 800 bp, similar to that expected for DNA in which the *Msp*I sites are unmethylated (see Discussion). Hence, there is evidence for little or no methylation of the 5´-cytosine in the CCG sequence in pea DNA.

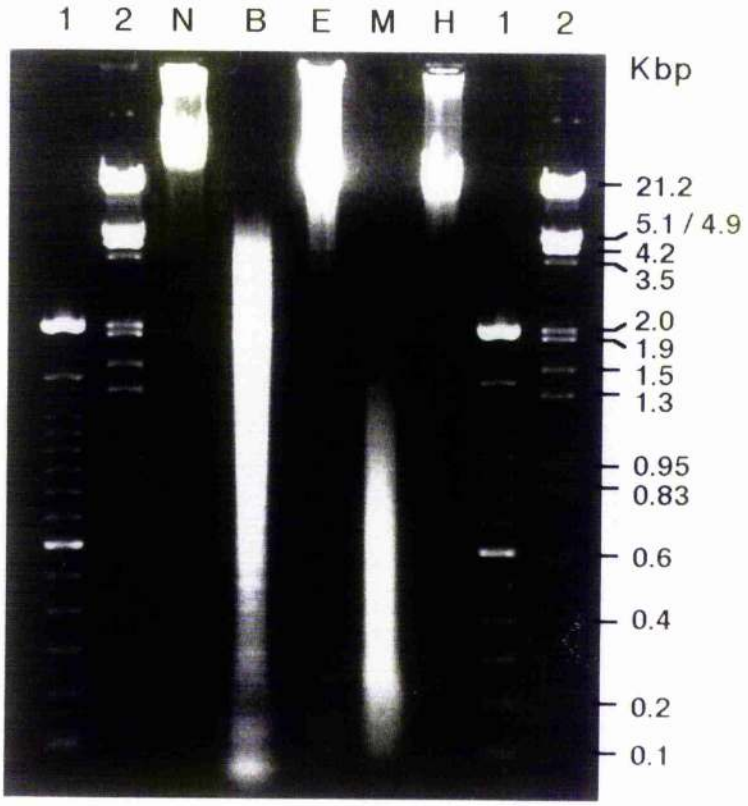


Figure 3.8: Extent of methylation of pea nuclear DNA

Pea DNA was digested with *Bst*NI (B), *Eco*RII (E), *Msp*I (M) or *Hpa*II (H) as described in the experimental Section. N is an undigested control and 1 and 2 represent a 100 bp ladder and I *Hin*dIII/*Eco*RI digest respectively. Sizes (in kb) are given down the right hand side.

The DNA is largely resistant to *Hpa*II and *Eco*RII, indicating extensive methylation of the internal cytosines in the sequences 5′-CCGG-3′, 5′-CCAGG-3′ and 5′-CCTGG-3′ sequences. The *Bst*NI digest acts as the control for *Eco*RII – both cleave CC(A/T)GG but the former is unaffected by methylation. Parallel digestions of plasmid DNA were performed for all enzymes and confirmed that the reaction had gone to completion (results not shown).

3.6 Discussion

Methylation of DNA provides a mechanism for the stable alteration of the local nucleotide sequence of a gene and may thereby play an important role in control of gene expression (Adams and Burdon, 1985; Hepburn *et al.*, 1987). In plants, clusters of non-methylated CpGs are found associated with the promoter regions of some genes (Hepburn *et al.*, 1987; Antequera and Bird, 1988) and there is an inverse correlation between the level of methylation of a gene and its transcriptional activity (see Finnegan *et al.*, 1993 for a detailed review).

Most reports consider only CG methylation, but Langdale *et al.* (1991) have reported that the cell specific accumulation of maize phosphoenolpyruvate carboxylase was correlated with demethylation at a *Pvu*II site (5'-CAGCTG-3') approximately 3 kb upstream from the transcription start site. This is the only *in vivo* evidence that CAG/CTG methylation might be involved in the control of gene expression.

Using a construct containing the CaMV *35S* promoter and the GUS reporter gene methylated at every cytosine using methyldCTP, Weber *et al.* (1990) showed that both transient and long term gene expression in transformed tobacco cells was inhibited and that the integrated DNA from non-expressing plants was methylated at both CG and CNG sequences. However, Hershkovitz *et al* . (1990) showed that methylation with M.*Sss*I was sufficient to block gene expression from the same promoter and they suggested that CNG methylation may have a function other than in gene silencing. If this is so then the enzymology and control of CG and CNG methylation may be significantly different.

Multiple methyltransferases of different types (i.e. a type I, a type II or a Dam methyltransferase) are often found in *Bacillus subtilies*. These different types of methyltransferase probably serve different functions within the cell. Different type II MTases have the capacity to methylate cytosine or adenine to produce 5-methylcytosine or 6-methyladenine. The Dam methyltransferase is involved in mismatch repair and in regulation of chromosome replication but not in restricting the growth of invading organisms. Trautner's group (e.g. Trautner *et al.*, 1988) have shown more than one MTase activity in one *E. coli* , which can arise from a single enzyme having more than one target recognising domain (TRD). These domains can be assembled in a modular fashion, although they can also act independently. Mutations can activate or eliminate TRDs (Trautner *et al.*, 1990) and so domains may be present in an enzyme that have the potential to be activated by a simple amino acid substitution. The TRDs are linked with a single active site for methyltransfer and hybrid methylases have been constructed containing domains from more than one enzyme (Trautner *et al.*, 1988; Klimasauskas *et al.*, 1991).

Although the major target site for vertebrate DNA methyltransferases is the CG dinucleotide, they also have a reduced activity on CA and CT target dinucleotides (Hubrich-Kühner *et al.*, 1989; Adams *et al.* 1993, Clark *et al,*. 1995) but no *in vitro* methylation of CC dinucleotides has been reported. This CA and CT methylation may be a result of the enzyme showing some activity with CAG and CTG and the mouse enzyme shows significantly greater activity with hemimethylated CAG/MTG sequences compared with CAG/CTG sequences (Adams *et al.*, 1993). Clark *et al.* (1995) demonstrated that the methylation machinery in mammalian cells is capable of maintenance as well as de novo methylation of CpNpG sites. Although the *in vivo* importance of this is unclear, it might indicate how two plant enzymes could arise from a common ancestral protein that had

properties similar to those of the vertebrate enzyme, by gene duplication and divergence of activity.

Plant DNA methyltransferases have been purified from wheat and rice (Theiss *et al.*, 1987; Giordano *et al.*, (1991) and these are of relatively low molecular weight (the wheat enzyme contains proteins of 55 and 35 kDa) and show the ability to methylate unmethylated and hemimethylated CG target sequences. However, there is no clear data available regarding their target recognition sequences. Using pea DNA as a substrate Yesufu *et al.* (1991) were unable to detect two different methylases in pea nuclear extracts but, using synthetic oligonucleotide substrates, Houlston *et al.* (1993) found evidence that there may indeed be separate CG and CNG methyltransferases. By using MonoQ chromatography I have now succeeded in separating these two activities (see Fig. 3.1). The elution of the two activities correlates with bands on SDS gel electrophoresis of about 140 kDa (with a presumptive breakdown product of 55 kDa) and 110 kDa respectively.

In order to unequivocally determine the substrate specificity of the two activities Maxam Gilbert sequencing of the oligonucleotides after *in vitro* methylation was undertaken. Methylated deoxycytidine residues fail to react with hydrazine thus leaving a gap in the sequence. All the cytosines in CG dinucleotides were either partially or fully methylated by the CG enzyme, indicating both maintenance as well as *de novo* methylation. However, the CNG enzyme was only able to methylate cytosines in 5′-CAG-3′ and 5′-CTG-3′ trinucleotides and failed to use any hemimethylated 5′-CCG-3′ or 5′-CGG-3′ sequences as a substrate. Different reaction conditions were used either by increasing the concentration of AdoMet and by supplementing ATP or divalent ions and scanning otherwise inactive fractions but no other enzymic activity was observed arguing against the existence of a third enzyme.

The question arises as to whether or not methylation of the 5′ cytosine in the CCG sequence occurs to a significant extent in peas. Such methylation has not been detected by genomic sequencing of the 1400 bp of 5′-flanking sequence of the maize alcohol dehydrogenase gene present in non-expressing leaf DNA (Nick *et al.*, 1986). However, looking at transgene methylation by genomic sequencing, Meyer *et al.* (1994) found a single mCCG in the inactive 35S promoter. They also found methylation at many non-symmetrical sequences especially in the inactive promoter. Similar methylation of non-symmetrical sequences has been seen in animal cells and the mouse methyltransferase shows a low level of activity on such target sequences *in vitro* (Hubrich-Kühner *et al.*, 1989; Adams *et al.*, 1993). Methylation of such sequences will not be maintained by a maintenance methyltransferase and is presumably introduced *de novo* following each round of DNA replication. Sites selected for such methylation activity might be distinguished by abnormal chromatin structure or as a result of their presence in multiple copies. The extent to which mCCG methylation is observed may, then, depend on the species, tissue and gene under study.

Gruenbaum *et al.* (1981) digested genomic wheat DNA with *Msp*I and found an average fragment size of 980 bp which was longer than their predicted size and they concluded that this was a result of some of the CCGG target sequences being resistant to digestion as a result of methylation of the 5′ cytosine. Using the same approach with genomic pea DNA fragments ranging in size from about 200 to 800 bp were obtained (Fig. 3.8). Assuming a base composition for the bulk of pea DNA of 37.4% G+C and a CpG suppression of 0.9 (Montero *et al.*, 1992), then the expected size of *Msp*I fragments is 740 bp which is close to the value obtained by Gruenbaum *et al.*, (1981) and ourselves (Fig. 3.8). In addition, about 1% of the DNA will consist of CpG islands (Antequera and Bird,

1988) with a G+C content of about 60% and no CpG suppression and these islands will be cut to give fragments of 120 bp (Gardiner-Garden and Frommer, 1992) but these fragments will not be visible on the stained gel. These results, therefore, provide no evidence for CCG methylation in pea DNA. Vongs *et al.* (1993) have produced mutant *Arabidopsis* lines that show reduced methylation at *Taq* sites, i.e. in CG dinucleotides. These mutations must, therefore, be in the CG methyltransferase or affect its action. The latter interpretation was favoured as the mutants were also affected in the methylation at other, non CG, sites in the genome.

Bezdek *et al.* (1990) demonstrated that in *Nicotiana* methylation of CG and CCG motifs in repeated sequences is differentially sensitive to the inhibitor ethionine. The fragments that they observed on the stained gels for the *Msp*I digests are very large, implying a very high level of methylation of this sequence or incomplete cleavage by the nuclease for some other reason. Controls were always included (sometimes internal) to confirm that the digestion had gone to completion. Several reports propose that apparent methylation of the 5′cytosine in the sequence CCGG may simply reflect the lower rate of cleavage of *Msp*I when the internal cytosine is methylated (Goldsbrough *et al.*, 1982) but it is also quite possible that CCG methylation is significant in other species.

Although there is no evidence of any enzymic methylation or *in vivo* evidence for CCG methylation in peas, the CG enzyme does interact with such target sites. It is possible that there is a low level of methylation of such targets that I have been unable to detect using the techniques employed. Based on the pyrimidinone studies (Fig. 3.5b) it could be speculated that there could be a latent activity of the CG enzyme on CCG sites. The overlap in target sites between a CCG:CGG enzyme and a CG enzyme could lead to complications where specific target site methylation is required. However, in light of the *in*

vivo evidence, I do not favour the possibility that the undegraded pea CG methylase might also methylate CCG sequences.

The restriction enzyme *Hpa*II also cuts the sequence CCGG but its action is blocked by methylation of the *internal* cytosine. Treatment of pea DNA with this enzyme yields fragments of about 20 kb implying that about 95% of such sites (and, by inference, 95% of all CG sites) are methylated. The enzyme *Bst*NI recognises the sequence CC(A/T)GG whether or not it is methylated. In a 2 h digestion genomic DNA is cut to fragments ranging in size from 500 to 4000 bp, consistent with the average predicted size of 1770 bp. The same sequence is cut by *Eco*RII, but this enzyme is sensitive to methylation at the internal cytosine. Although some cleavage does occur, the fragment size is greater than 20 kb implying that most (about 90%) of these sites are methylated *in vivo*.

CHAPTER FOUR

Isolation and characterisation of CWG islands in the pea genome and studies on the effect of CG and CWG methylation on gene expression

4.1 Introduction

The methylation of CG dinucleotides is well known in the vertebrate genome, but in addition the genomes of higher plants contain methylcytosine in the trinucleotide sequence mCNG, where N is reportedly any of the four common DNA bases (Gruenbaum, *et al.*, 1981). As there is little depletion of the CG dinucleotides and even less depletion of CNG trinucleotides (Gardiner-Garden *et al.*, 1992), this results in methylcytosine representing as much as a third of the cytosine bases in plant DNA.

Little is known about the localisation or function of methylcytosine in plant DNA. While Antequera and Bird (1988) have shown that, as with vertebrates, angiosperm genomes contain clusters of unmethylated CG dinucleotides it is not known whether the bulk of CG dinucleotides are found in such clusters. Indeed, this seems unlikely in view of the lack of CG suppression which indicates that about one in five cytosines will be followed by a guanine (i.e. the majority of CG dinucleotides are not in CG islands). The resistance of plant DNA to cleavage by restriction enzymes such as *Hpa*II indicates that the majority of the CG dinucleotides are methylated and are spread throughout the genome (Pradhan and Adams, 1995).

In a similar manner, restriction enzyme studies have shown that most cytosines in CWG sequences (where W is A or T) are methylated, although there is controversy about methylation of the 5′ cytosine in CCG sequences (Pradhan and Adams, 1995).

In chapter 3, I have reported the purification of separate CG and CWG DNA methyltrtansferases from *Pisum sativum* (Pradhan and Adams, 1995). The finding of different methyltransferase activities in the same cell raises the possibility that the action and control of these enzymes may be different, even to the extent that methylation of CG and CWG sequences may serve completely different functions. It is therefore interesting to have further information on the biological significance of CG and CWG methylation on reporter gene expression and on the distribution of the target sequences for these methyltransferases.

This chapter reports an attempt to dissect the genome of pea. My findings are summarised as follows: *Pisum sativum* contains unmethylated CWG islands which are distinct from the CG islands; rather than being depleted in the methyltransferase target sequence, the CWG islands are enriched in CWG trinucleotides; in addition, the pea genome contains a repeated DNA sequence containing widely dispersed unmethylated *Hpa*II sites and unmethylated *Pst*I sites; transcription of some promoters that are rich in CWG sequences but deficient in CG sequences is inhibited by CWG methylation.

4.2 CG and CWG islands

Cleavage of pea genomic DNA with *Pst*I (CTGCAG), *Pvu*II (CAGCTG) or *Eco*RII (CCWGG) - all of which are inhibited by cytosine methylation - followed by end-labelling, produces DNA fragments in two or three size ranges (Fig 4.1a).

The largest fragments (referred to as *Pst* large fragments or PLFs) are greater than 20 kb and represent sequences lacking unmethylated sites for these restriction enzymes. The smallest fragments are less than 100 bp in size and an intermediate group with an average size of about 500 bp arise from sequences containing frequent sites for these enzymes. Similar results have been reported by Messeguer *et al.*, (1991).

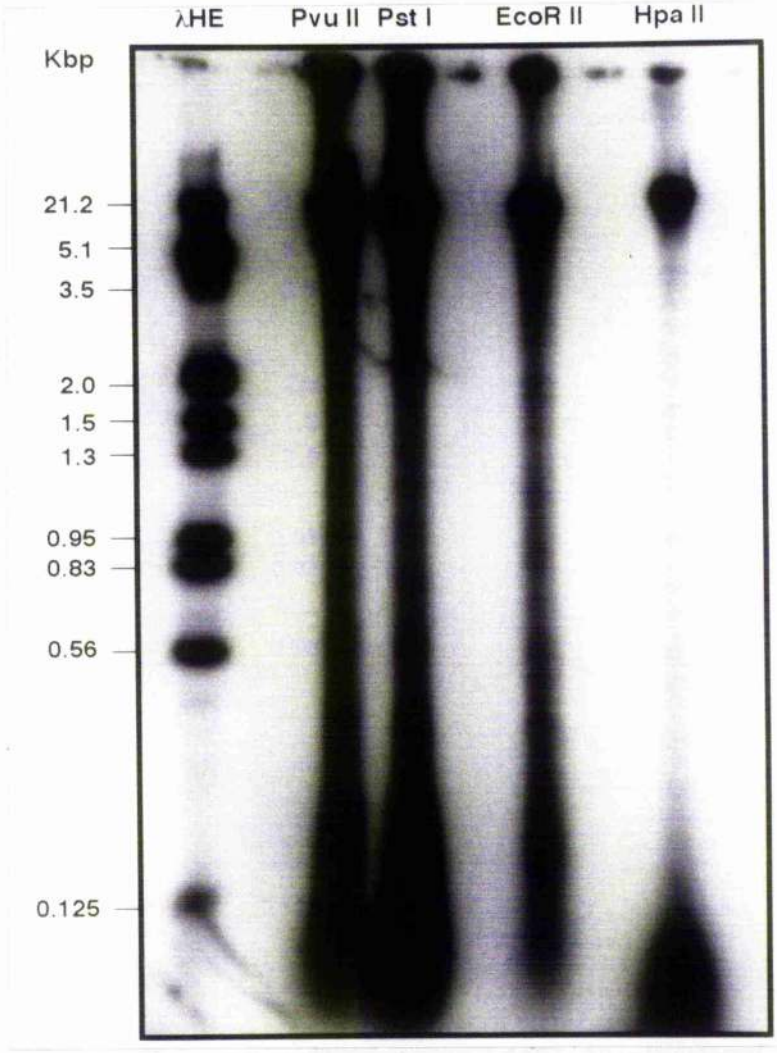


Figure 4.1a CG and CWG islands in pea nuclear DNA.

Pea nuclear DNA was digested with restriction enzyme *Pvu*II, *Pst*I, *EcoR*II and *Hpa*II and end labelled fragments were separated on a 1% agarose gel. The photograph shows an autoradiograph of the dried gel. The sizes of the marker λHE bands are indicated down the left hand side.

The size of these *Pst*I fragments is much smaller than predicted from a consideration of a random distribution of bases. Were *Pst*I sites distributed evenly, the smallest mean size of *Pst*I tiny fragments (PTFs) would be obtained from DNA in which two thirds of the bases were G+C (Fig. 4.1b) and these fragments would have a size of about 2900 bp. It was considered that some or all of these small fragments might have evolved through cleavage of an array of satellite DNA but there is no obvious banding pattern typically seen when satellite DNA is cut by a restriction enzyme. Sequence analysis of these CWG rich clones rules out this possibility for the majority of the fragments. Other possible reasons for the small size of the fragments obtained will be discussed below.

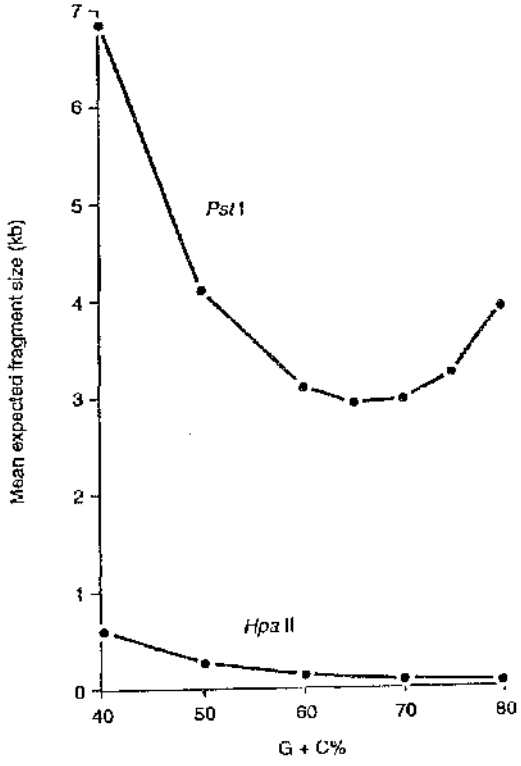


Figure 4.1b Size prediction of fragments

Predicted size of fragments obtained from DNA of the indicated composition (% G+C in the X axis and mean distribution on the Y axis) and containing a random distribution of bases after cleavage with *Pst*I or *Hpa*II.

*Hpa*II digestion gives a similar distribution of fragments as previously reported by Antequera and Bird (1988). In the genome of mouse a large-sized component (*Hpa* large fragments or HLFs) lacking unmethylated CCGG sequences and a component containing fragments of about 80 bp (HTFs) derived from a region of DNA containing frequent, unmethylated CCGG sequences (CG islands) is present. It was this type of evidence that led to the notion that CG islands exist in both animal and plant DNA. Similarly, the results of the digestions with *Pst*I, *Pvu*II and *Eco*RII are evidence for the presence of CWG islands in plant DNA.

It is possible that the CG and CWG islands are one and the same and that islands exist that are enriched in both unmethylated CG and CWG sites. The alternative explanation is that two distinct types of island exist in plant nuclear DNA. The following experiments were designed to distinguish between these possibilities.

When the end-labelled HLFs are isolated and treated with *Pst*I a very small fraction of the HLFs are digested to small fragments, but the majority of the HLF DNA remains unaffected (Fig. 4.2a). This indicates that a very small proportion of CG oceans have an unmethylated CWG close to their boundary but that long regions of DNA exist with no unmethylated CG or CWG sequences. Digestion with *Bgl*II (recognisation seq: 5´-AGATCT-3´) leads to a reduction in size of most HLFs. To further clarify this point, end labelled HLFs were isolated, cleaved with *Pst*I and end-labelled again before electrophoresis (Fig 4.2a) showing that CWG islands do not occur in regions of the genome lacking unmethylated CCGG sequences : i.e. they occur in the regions of DNA that also contain unmethylated *Hpa*II sites

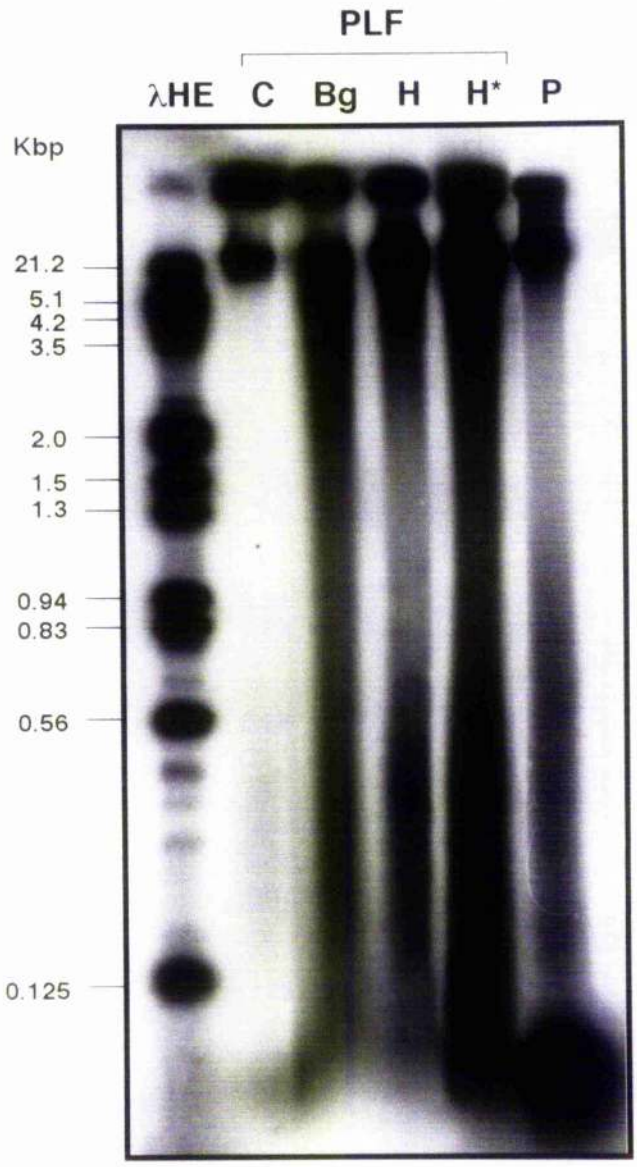


Figure 4.2a Analysis of PLF DNA for CpG islands

The end-labelled *Pst*I large fragments (PLFs) were excised from a gel and subjected to further digestion with *Pst*I, *Bgl*I of *Hpa*II. The asterisk indicates that, in this case, the fragments were end labelled again after the second digestion. The figure shows an autoradiograph of the fragments separated on a 1% agarose gel. C are control PLFs not subject to a second digestion.

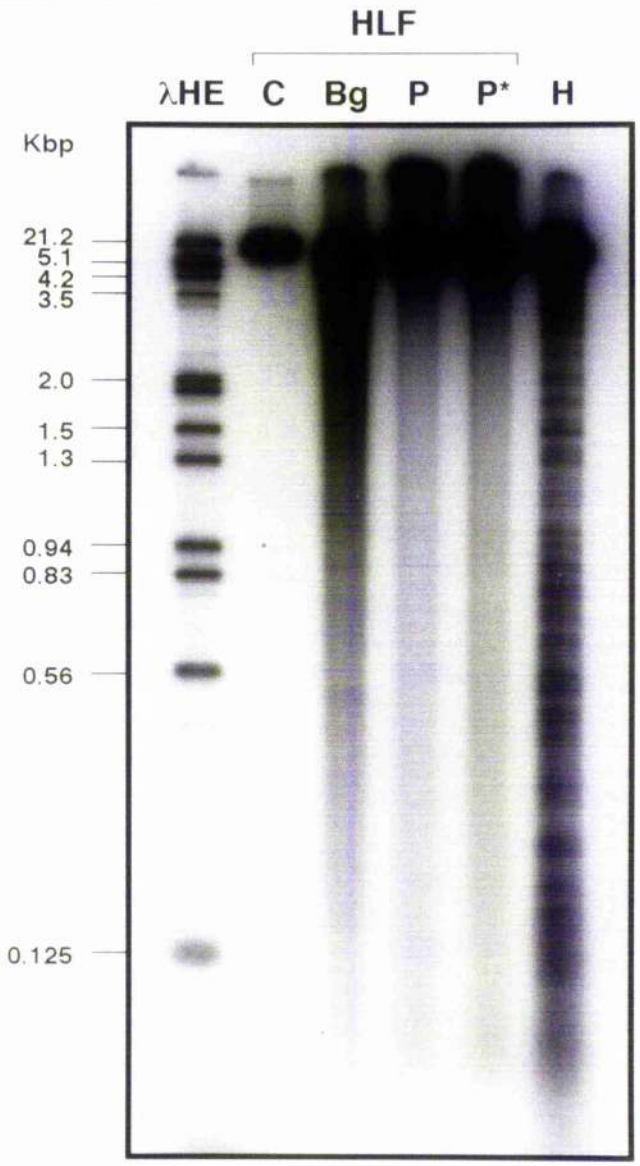


Figure 4.2b Analysis of HLF for CWG islands

The end-labelled *Hpa*II large fragments (HLFs) were excised from a gel and subjected to further digestion with *Pst*I, *Bgl*I of *Hpa*II. The asterisk indicates that, in this case, the fragments were end labelled again after the second digestion. The figure shows an autoradiograph of the fragments separated on a 1% agarose gel. C are control HLFs not subject to a second digestion.

In the reverse experiment (Fig. 4.2b), when end-labelled PLFs were treated with *Hpa*II, most labelled ends remained in large fragments but a significant proportion were converted to small fragments (mean size about 300 bp) indicating that some unmethylated CCGG sequences are present, at least, near the shores of some CWG oceans (i.e. at least some PLFs are bordered by unmethylated *Hpa*II sites). This result would be obtained if unmethylated *Pst*I and *Hpa*II sites were present in a common region where the frequency of *Hpa*II sites (a four base cutter) would be expected to be far greater than that of *Pst*I sites (a six base cutter). Such islands, if they exist, are far more likely to have a *Hpa*II site at their boundaries than a *Pst*I site. This result is consistent with at least some CG islands containing unmethylated CWG sequences.

In this experiment the fragments had been labelled only following the *Pst*I treatment and hence cleavage at multiple *Hpa*II sites located centrally in the PLFs would still generate only relatively large *labelled* fragments (any HTFs generated would be unlabelled and hence not observed). This could mean that even if PLFs contain CG islands they are unlikely to be detected. To clarify this point, end-labelled PLFs were isolated, cleaved with *Hpa*II and labelled again before electrophoresis. It is clear from Fig. 4.2b that there has been a considerable increase in the labelling, particularly of small fragments which indicates that many PLFs do contain CG islands and generate HTFs. This would be the case if a proportion of CG islands lack (unmethylated) *Pst*I sites.

These results indicate that separate CG and CWG islands occur in pea nuclear DNA. Some CG islands clearly contain unmethylated *Pst*I sites but others lacks them. Despite the lack of *Pst*I sites these islands may contain unmethylated CWG sites. Few CWG islands are present in CG oceans; the exception being that some of the very short *Pst*I fragments may arise from CG oceans.

4.3 Cloning of PTFs

PTFs were cut from the gel and cloned into pBluescript and several hundred clones were isolated. Fifty of these have been sequenced and the sequence data is analysed in Table 2. The G+C content of these clones ranges from 33 to 57% with an average of 43% which is typical of the genome as a whole and which is very atypical of CG islands. Moreover, these sequences are deficient in the CG dinucleotide and the CCG trinucleotide that occur at only 50% the expected frequency. However, the CWG trinucleotides occur more frequently than expected with the observed frequency being 170% of that expected from a random distribution of bases within each clone (i.e. o/e = 1.7).

Of the fifty PTFs sequenced, four (102, 104, 127, 130) are identical to each other, although in different orientations. PTF145 differs from 130 in one position and PTF157 differs in seven positions. Two other clones (119, 139) are identical to each other but are in opposite orientations. It is possible that the latter pair arose from the same clone. The first six clearly arise from a multi-copy sequence sequence. As most repetitive sequences are highly methylated (Deumling, 1981) it would be surprising to find one with unmethylated *Pst*I sites every 98 bp. Although DNA was purified from isolated nuclei, it was considered that this sequence might represent chloroplast DNA. However, it shows no homology to chloroplast DNA sequences present in the Gen/EMBL data base and a more likely explanation is that the sequence is derived from ribosomal DNA. Indeed clone PTF130 shares considerable sequence homology with ribosomal DNA sequences in the database.

Table2 **Analysis of PTF clones**

49 clones were sequenced and they are arranged here in size order. Based on their G+C content, the expected frequency of occurrence of CG, CCG and CWG sequences was calculated. The o/e values are the observed frequency divided by the expected frequency for the indicated sequence. The asterisks indicate clones with identical or nearly identical sequence (the numbers in brackets indicate the number of differences from PTF130 or PTF102) and ! indicates clones in the reverse orientation. Clones containing *Hpa*II sites are indicated by †.These sequences will be deposited in EMBL database.

Name	length: bp	%G+C	CG: o/e	CCG: o/e	CWG: o/e
PTF155	66	53	0.7	0	1.4
PTF150	80	55	0.5	0.5	3.7
PTF129	88	43	0.5	0	2.9
PTF110	88	48	0.7	0.7	1.8
PTF31	92	41	0	0	1.4
PTF156	98	45	1	1	2.6
PTF130 *	98	51	0.2	0	1.6
PTF127 *	98	51	0.2	0	1.6
PTF102 *	98	51	0.2	0	1.6
PTF104 !	98	51	0.2	0.6	1.6
PTF145 !	98 (1)	54	0.2	0	1.6
PTF157 *	98 (7)	49	0.2	1	1.4
PTF113	101	37	0.6	1.9	2.4
PTF147	102	48	0.3	0.7	1.7
PTF140	106	40	0	0	1.6
PTF6	113	39	0.5	0	3.8
PTF153	114	57	0.3	0.5	1.6
PTF158	114	39	0	0	0.5
PTF105	117	49	1.2	1	2.3
Average	**98.3**	**47.4**	**0.4**	**0.4**	**2.0**

PTF174	126	45	0.6	0	1.4
PTF141 †	135	40	0.6	1.3	1
PTF17	139	44	0.6	1.7	2
PTF125	140	39	0	0	1.7
PTF159	141	41	0.2	0	1.3
PTF167	141	43	0.2	0	1.2
PTF119	141	43	0.2	0.9	1.1
PTF139	141	43	0.2	0.6	1.1
PTF121	142	46	0.5	0.6	1.7
PTF106	142	38	0.5	0	1.2
PTF118	142	42	0.2	0	1.1
PTF170	144	42	0.2	0	1.6
PTF21	147	43	0.3	1	1.5
PTF138	148	37	0.4	0	1.9
PTF133	153	42	0.3	0	2.6
PTF115	160	43	0.4	0.8	2.3
PTF166	161	45	1	1.8	1.3
PTF160	164	45	0.4	0	1.8
PTF151	166	54	0.9	0.8	1.5
PTF134 †	166	45	0.5	1.4	1.3
PTF165	170	45	0.5	1.8	1.3
PTF131	175	39	0.5	1	2.1
PTF136	177	45	0.5	0	2.1
PTF126	180	43	0.5	0	2.5
PTF173	196	42	0.5	0.6	1.2
PTF137	206	33	0.2	1.4	2.2
PTF169	233	39	0.3	0	1.9
PTF111 †	257	41	0.3	0.8	1.1
PTF152 †	299	42	0.2	0.5	1.4
PTF154 †	307	45	0.7	0.5	0.8
Average	**171.3**	**42.5**	**0.4**	**0.6**	**1.6**
Overall	143.9	44.4	0.4	0.5	1.7

From these sequence data it is clear that PTFs do not arise from typical CG island DNA (i.e. DNA with a G+C content greater than 60% and no deficiency in the CG dinucleotide). Not all organisms contain typical CG islands and those in fish DNA are reported to be poor in G+C. Nevertheless, fish islands still contain the expected frequency of CG dinucleotides (Cross *et al.*, 1991). Thus the CWG islands in plants are distinct from both the typical CG islands and from those found in fish.

4.3 Southern analysis using PTF probes

Nearly all of the PTF clones contain CG dinucleotides but only 10% contain a *Hpa*II site (i.e. 5 *Hpa*II sites occur in about 7 kb of DNA). If *Hpa*II sites (and CG dinucleotides in general) are unmethylated in CWG island DNA, then these regions will be sensitive to *Hpa*II digestion and CWG island DNA would not be present in HLFs. This was what was found in the experiments reported above. Using a cloned PTF as a probe, *Hpa*II treatment of genomic DNA would be expected to produce fragments from CG-unmethylated CWG island regions of an average size of about 1000 bp. However, if the CG dinucleotides are methylated in CWG islands, then CWG island DNA would be present in the HLFs. The former result was found by Southern blot analysis of nuclear DNA cleaved with *Pst*I or *Hpa*II. The blots (Fig. 4.3) were probed with the labelled inserts from clones PTF152, PTF154 or PTF111 and also PTF130. With the first two probes the *Pst*I digest shows a single band of the expected size indicating that this is a single copy sequence and that the bounding *Pst*I sites are unmethylated in all cells. Probing with PTF111 highlights a doublet indicating two very similar copies (perhaps alleles) of this sequence: once again all bounding *Pst*I sites are unmethylated. As these three clones contain an internal *Hpa*II site, the *Hpa*II digest would be expected to show double the number of bands if all *Hpa*II sites are

unmethylated; and this is what is found (Fig. 4.3). (With probe PTF154 the *Hpa*II site is very near the 5´ end and the second band is not visible on the autoradiograph.) There are alternative explanations for these findings but the simplest is that PTFs occur in totally unmethylated regions of the genome and that the CGs that occur in or around CWG islands are unmethylated. Thus CWG islands do not occur either in CG islands (where *Hpa*II sites are frequent and HTFs are about 80 bp long) nor embedded in long stretches of DNA lacking unmethylated CGs. From the size of the *Hpa*II fragments identified with the PTF probes an estimate of the minimum extent of the CWG island can be made. Thus the two *Hpa*II fragments lit up by PTF152 are of 650 and 480 bp indicating that the unmethylated island stretches for at least 1100 bp and contains 3 unmethylated CCGG sequences. The islands that contain the 2 copies of PTF111 could be between 3 and 4 kb long; rather larger than what would be expected for DNA of 42% G+C and no deficiency in CG dinucleotides.

When the multi-copy PTF130 was used as a probe in Southern transfers, a single 98 bp fragment is observed after *Pst*I digestion of pea DNA (Fig. 4.3). No ladder of bands is seen: a result consistent with this not being satellite DNA. This probe picks out a single, 410 bp band in a *Hpa*II digest indicating that the *Pst*I fragment is embedded in a longer region of unmethylated DNA. The most likely explanation for the presence of a multi-copy sequence being present as a single band is if that DNA is part of a conserved multi-copy sequence and, as explained above, we believe this could be rDNA.

It is, perhaps, surprising that of the 50 clones sequenced none has the characteristic high G+C content of CG island DNA and only 5 contain a *Hpa*II site. This is totally consistent, however, with CG and CWG islands being separate entities

with different characteristics and independent patterns of methylation. It is clear then that the target sites of the two pea DNA methyltransferases are found in distinct regions of the genome.

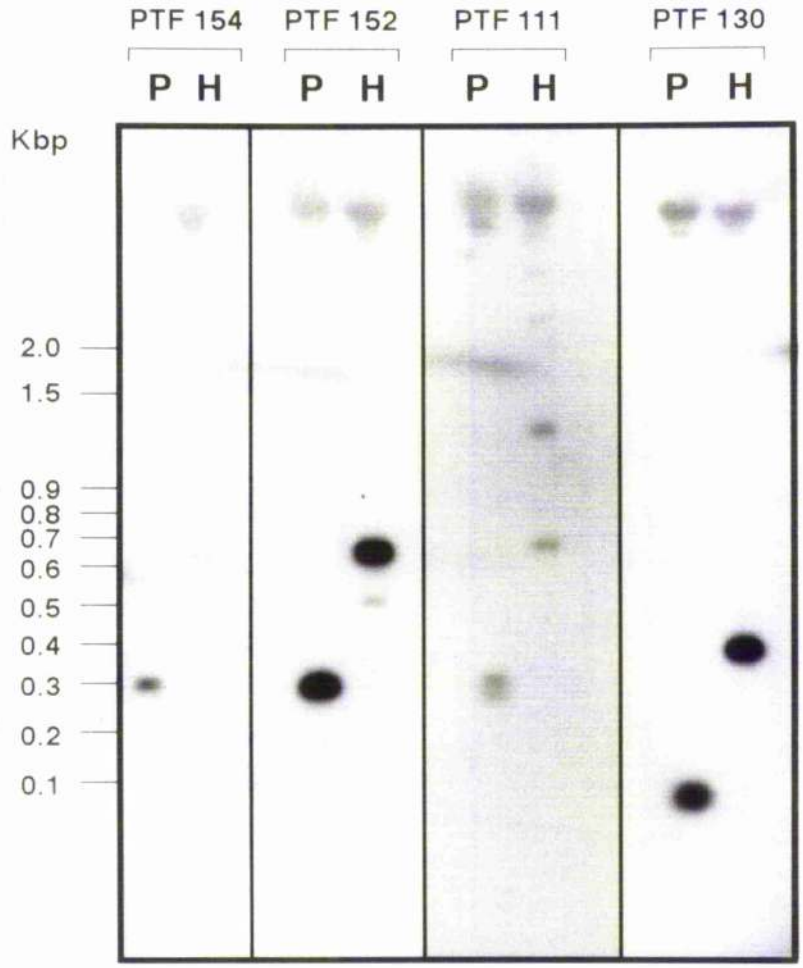


Figure 4.3 Southern analysis of PTF cDNA clones

Genomic DNA was cleaved either with *Hpa*II (H) or *Pst*I (P) and separated on a 1% agarose gel prior to transfer to Hybond N+ membrane for hybridisation. Random primed probes were used. THe clone identity numbers are indicated on top and the size of marker (100bp ladder) bands is indicated down the left hand side

4.4 Frequency of occurrence of islands

In order to assess the number of CG and CWG islands in the pea nuclear DNA, *Eco*RI cleaved DNA was end labelled and run on a gel, when the number of fragments expected from a haploid genome of 4.8×10^9 bp with 38% G+C content would be 1.6×10^6 ; of mean size, 3000 bp. On further cleavage with *Hpa*II or *Pst*I the proportion of labelled ends present in molecules shorter than 250 bp increased by 7.5% and 0.4% respectively (Fig. 4.4). These new, small fragments arise from cleavage at island sequences within an original *Eco*RI fragment. As islands are 1–2 kb in length, the effect of cleavage of an island (with *Hpa*II or *Pst*I) would be to shatter the island to small pieces and so the proportion of ends in small molecules is a reflection of the proportion of *Eco*RI fragments containing islands. We can, therefore, estimate that there are about $1.6 \times 10^6 \times 7.5/100 = 120000$ CG islands and $1.6 \times 10^6 \times 0.4/100$ CG = 6400 CWG islands in the haploid pea genome. The generation of these numbers involves several assumptions, but gives an approximation of the frequency of islands in pea genomic DNA. These figures are tentative estimations, the major assumption being that the distribution of *Eco*RI sites is random in the pea genome and that the sizes of the CG and CWG islands are similar. The value for the number of CG islands is about four times higher than the number estimated to be in the mammalian genome (Bird, 1986: Gardiner-Garden *et al.,* 1992).

If the ribosomal DNA sequence contains a *Pst*I site within 250 bp of an *Eco*RI end then this end will now appear in low molecular weight DNA and this will affect our estimate of the number of CWG islands. However, the effect will be small and give rise to about a $6 \times 100/49 \times 2 = 6\%$ over-estimate of the number of islands. (Six out of 49 clones correspond to this sequence but only one end will be released as low molecular weight DNA). If the *Pst*I site is further away from an *Eco*RI site, the resulting labelled ends will remain greater than 250 bp and so will not affect the calculation.

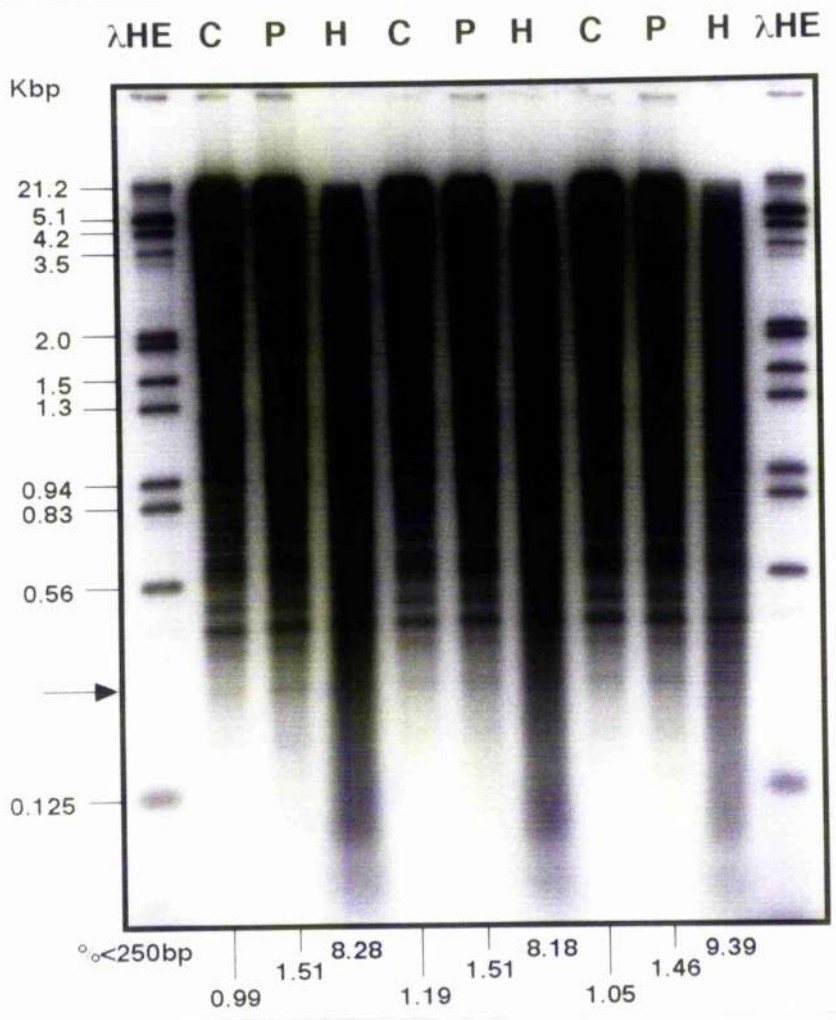


Figure 4.4 Island frequency in pea

Genomic DNA was cleaved with *Eco*RI and the fragments end labelled using Klenow DNA polymerase. Unincorporated triphosphate and the smallest fragments were removed on a Chromospin 10 column (Clontech). The DNA was then cleaved with either *Pst*I or *Hpa*II and the fragments separated on a 1% agarose gel. The proportion of ends in fragments smaller than 250 bp (indicated by an arrow in the left) was quantitated using the phosphoimager and these values appear below each lane. The size (in kb) of marker bands appears to the left of the gel. The gel shows triplicates of control DNA (C), DNA cleaved with *Pst*I (P) and DNA cleaved with *Hpa*II (H)

4.5 Data base analysis of plant promoter sequences

From the GenEMBL database 40 plant gene promoters were fetched and analysed using the GCG program. The *Phaseolus vulgaris rbcS2* promoter (Urwin and Jenkins, unpublished) was also analysed. Fig 4.5 shows the relationship between the frequency of occurrence of CG and CWG sequences. The proportion of cytosines in CWG trinucleotides varies between 2 and 17% and, in the majority of cases, is independent of the frequency of occurrence of CG dinucleotides. In most cases this is fairly constant with 6.5% to 12% of cytosines being in CG sequences. The *rbcS2* promoter is unusual in having a very low CG content and five promoters, including that of the cauliflower mosaic virus *35S* gene, have a high CG content. There is no parallel or reciprocal relationship between the CWG and CG contents of these promoters. The *rbcS2* promoter is clearly a candidate for a region containing a CWG island in that it has a deficiency in CG (the CG/GC ratio of the pea *rbcS2* promoter is less than 0.1) and a higher than average CWG content, even though, on closer inspection, it was found to contain no *Pst*I site. In contrast, the 12 promoters with 15% or more of their cytosines in CG dinucleotides show no deficiency in CG (average CG/GC = 1.1) and a range of values for the proportion of Cs in CWGs that have an average value (9.6%) very close to that for all the promoters (9.3%).

4.6 Effect of methylation on expression from different promoters

With the collaboration of Dr. Nigel Urwin a few preliminary investigation on the effect of *in vitro* methylation on the transient expression of genes introduced into bean protoplasts were carried out to distinguish the biological effect of differential methylation. The expression of the *CAT* reporter gene under the control of either the *CaMV 35S* promoter or the bean *rbcS2* promoter, was investigated.

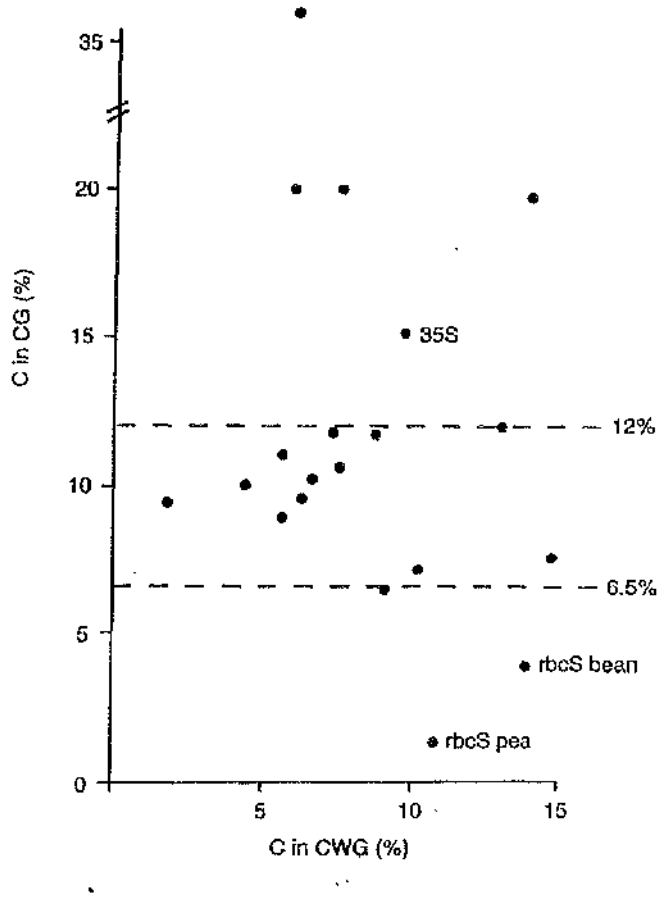


Figure 4.5 Analysis of plant promoter

1 kb upstream sequences from the translation start site of 41 plant promoters were analysed. The figure shows 20 of them. The graph shows the percentage of cytosine in CG dinucleotides plotted against the percentage in CWG trinucleotides. The dotted lines enclose the majority of sequences. The promoter sequences and their accession numbers are: phenylammonia lyase (*PAL*, M83314), tomato anionic peroxidase *tap -1* (S63739), pea *UBC4* (L39921), Pea rbcs3A (M21356), pea plastocyanin PETE (S66544), *Arabidopsis* cdc2a (U19862), *Arabidopsis* plastocyanin (S67901), *Arabidopsis* PAL (S45847), Maize Zein (K00543), *Nicotiana* promoter (X55365), *Nicotiana*seed coat protein (U08931), *Phaseolus* hydroxyproline rich glycoprotein (U18991), *Phaseolus* glutamine synthetase (S44882), Rice actin (S44221), Petunia *adh2*, (U25536), Petunia chalcone synthase A (S52984), Soyabean leghaemoglobin Iba (X06438), soyabean auxin responsive promoter (D11429).

So as not to be confused with any pre-existing methylation, the plasmids were grown in an *E. coli* strain (GM2163 – New England Biolabs) that was dam⁻ and dcm⁻ (i.e. lacking methylation of GATC and CCWGG sequences). Complete plasmids were methylated *in vitro* with either the purified pea CWG methyltransferase (Pradhan and Adams, 1995) or one of a variety of prokaryotic methyltransferases that would lead to the introduction of methyl groups into the following sequences : C^{m}CGG (M.*Hpa*II), ^{m}CCGG (M.*Msp*I) or ^{m}CG (M.*Sss*I). Compared with the *rbcS2* promoter, the *35S* promoter leads to a 3000-fold greater level of CAT expression with the control, unmethylated plasmids (i.e. the *rbcS2* is a weak promoter and the *CaMV35S* a strong promoter).

Methylation of CCGG sequences has little effect on gene expression from either promoter (Fig. 4.6). This is, perhaps, not surprising as neither promoter contains CCGG target sites and what sites there are in the CAT gene and the vector are well spread out. In contrast, methylation of all CGs with M.*Sss*I leads to a complete inhibition of expression from both promoters. Whether this is a direct effect of promoter methylation or an indirect effect mediated partly by vector methylation (Kass *et al.*, 1993) we are unable to say and this needs further investigation.

Methylation of CWG sequences leads to a dramatic inhibition of expression from the *rbcS2* promoter but has only a limited effect on expression from the *35S* promoter (Fig. 4.6). As the plasmids used are very similar, differing substantially only in the promoter region, this difference must be attributable to promoter methylation.

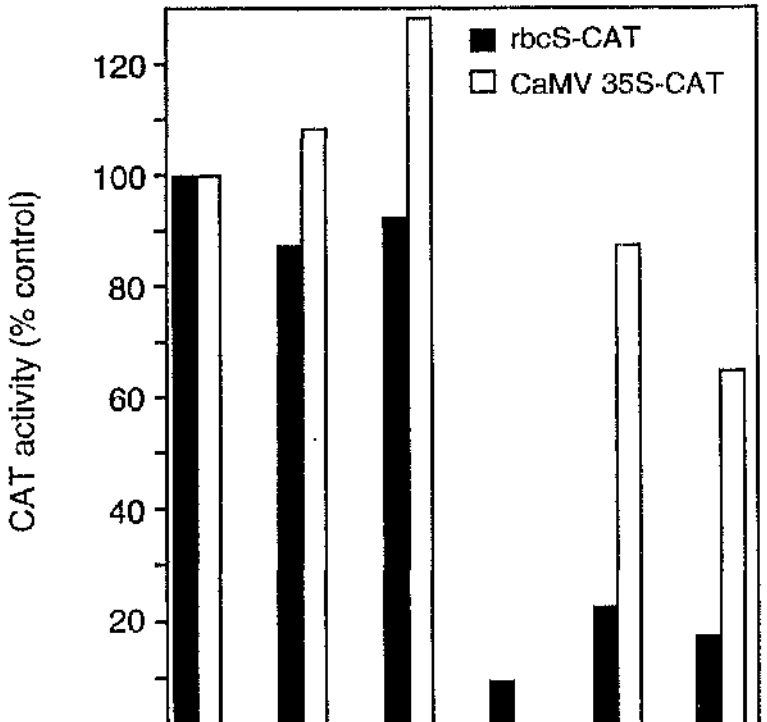


Figure 4.6 Transient expression of methylated plasmids

The plasmids were grown in dam⁻, dcm⁻ bacteria and methylated *in vitro* as described in the methods Section. They contained the CAT reporter gene linked to either the CaMV 35S promoter (CaMV *35S-CAT*) or to the bean *rbcS2* promoter (*rbcS-CAT*). Results, which are the average of two experiments each carried out in duplicate, are expressed as a percentage of expression with a mock-methylated control plasmid. Methylation with the pea CWG methyltransferase was for 4 h (25% saturation) or 24 h (80% saturation).

4.7 DISCUSSION

Selection of PTF clones may not reflect the actual distribution of PTFs. Thus the average insert size of the sequenced clones is 144 bp whereas the autoradiograph (Fig.4.1a) suggests clones might fall into two size groups; one with inserts of less than 100 bp and the second with inserts of about 500 bp. Clearly there is some bias; the cloning efficiency appears to be greater for the smaller fragments. Furthermore, when the clones are arranged in size order (Table 2) a not-very-convincing trend is apparent in which the G+C content and the o/e for CWG fall as the size of the cloned fragments increases.

Among the fifty clones, six (PTF145, PTF157, PTF130, PTF127, PTF102, PTF104) contain inserts of nearly identical (though sometimes inverted) sequence. There is no evidence that these arise from chloroplast DNA that might contaminate the nuclear DNA preparation. The total chloroplast genome of several plants such as rice (acc no:X15901), tobacco (Z00044) and liverwort, Marcantia polymorphia (acc no: X04465) are available in the database, but none of the clones matched with them. Thus we conclude that these sequences may represent clones of a ribosomal DNA sequence that is unmethylated at both CG and CWG sequences. Wordsearch using GCG programme indicates two best fit from *Toxoplasma gondii* ribosomal RNA large and small subunit genes (acc no: U18086) and *Euglena gracilis* extra chromosomal DNA for large subunit (LSU) rRNA (acc no:X53361).

Can the clustering of CWG trinucleotides in CWG islands explain the small size of the PTFs? The higher the concentration of CWG sites, the greater the chance of two sites occurring together to give a *Pst*I site. From Table 2 it is apparent that the

smallest clones have the greatest concentration of CWG sites and vice versa. If we consider clone PTF6 that has the greatest concentration of CWG sites, these are still separated, on average, by 20 bp: to have two pairs of contiguous CWG trinucleotides (i.e. *Pst*I sites) within 113 bp would be very unlikely from a statistical point of view. This argument is even more forceful for PTF111. For the average PTF with o/e of 1.7 for CWG trinucleotides, a *Pst*I site would be expected only every 2068 bp compared with the actual frequency of 144 bp. It thus appears that, not only does CWG occur more frequently than expected but, over and above this, the same is true for CWGCWG sequences. The conclusion must be that there has been a positive selection for these sequences and a similar positive selection might explain the high G+C content of CG island DNA.

In 1986, Tautz *et al.*, analysed the current EMBL nucleotide sequence data base for the presence of simple repeats and found that in nearly all cases these occurred with higher frequency than expected. Repeats of the trinucleotide CAG/CTG were particularly common occurring over 80 times as often as found in a random DNA sequence. Not only were simple repeats found to occur more commonly than expected, but cryptic repeats (i.e. repeats with a variable number of other bases separating the target sequence) were also found to be much more common than predicted. A number of mechanisms have been proposed to explain how directly repeated sequences might be generated. A recent study by Strand *et al.*, (1993) indicates that slippage at replication is the most likely mechanism. Whether this can account for the increased frequency of cryptic repeats and inverted repeats (as in the *Pst*I site) is not known.

From these results it could be proposed that CG and CWG islands share the common property of being a cluster of unmethylated CG or CWG sites respectively.

These islands are present in different but largely unmethylated regions of the genome. Although CWG islands contain unmethylated CGs and unmethylated *Hpa*II sites, these occur at a frequency lower than expected from a random distribution and markedly lower from that seen in the CG islands. From the spacing of sensitive *Hpa*II sites associated with particular PTFs, we can conclude that the unmethylated CWG islands extend for more than a thousand bp and resemble CG islands in this respect. However, only about 5% of genes seem to be associated with CWG-rich promoters compared with most genes being associated with CG islands.

The bean *rbcS2* promoter contains a CG in the sequence CACGTG, known as a G-box, that binds the factor, GBF, implicated in regulation by a variety of environmental signals (Gilmartin *et al.*, 1990; Williams *et al.*, 1992) and that is essential for expression in particular contexts (Donald and Cashmore, 1990; Urwin and Jenkins, unpublished). However, Southern analysis by Nigel Urwin, using a probe stretching from –1433 to –4 bp, relative to the transcription start site of the bean *rbcS2* gene indicates an absence of CG methylation of the G-box sequence in genomic DNA from dark grown, dark adapted or light grown leaves, roots or seeds. This analysis was possible as *Eco*72I (Promega) is sensitive to methylation of the internal cytosine in the sequence CACGTG (results not shown). Furthermore, the G-box in the pea *rbcS3A* promoter lacks the CG dinucleotide (Gilmartin *et al.*, 1990). It is possible that the very strong inhibitory effect of CG methylation in transient expression assays is mediated directly by action on GBF binding but it may equally well be mediated through the formation of an inactive chromatin structure (Kass *et al.*, 1993). This could also be true for the *35S* promoter. In contrast, the selective effect of CWG methylation on the weak *rbcS2* promoter implies a direct effect on the promoter but whether this involves inhibition of transcription factor binding or whether a mCWG-binding protein is

involved awaits further investigation. As we have said, the bean *rbcS2* promoter contains very few CGs (CG/GC = 0.2) but the two promoters have a very similar level of CWG sequences. Southern analysis indicates that there is no more than a very low level of methylation of the *Eco*RII site at –850 in the *rbcS2* promoter in DNA from all bean plant tissues tested. A similar complete lack of methylation was observed at the *Bal*I site (TGGCCA*G*) at – 138 that overlaps a CAG trinucleotide. Thus, although we have shown that, in transient expression assays, both promoters are sensitive to CG methylation and the *rbcS2* promoter sensitive to CWG methylation, we have not found a situation where the endogenous bean rbcS2 promoter is methylated despite looking in tissues showing a wide range of expression. This may be because all tissues may have at least a low level of expression and because complete inactivation by methylation would be strongly selected against.

Although in this analysis only one gene that is preceded by a CWG island has been identified , analysis indicates that about 5% of plant genes fall into this category. CWG methylation could play a part in the regulation of such genes although, if the CG analogy applies to CWG islands, they will be unmethylated in most situations as is true for the bean *rbcS2* gene. Island methylation may precede the complete inactivation of genes by allowing inactive chromatin structures to spread from surrounding regions, while an unmethylated island protects a gene from this insidious invasion. If CG and CWG islands perform similar functions, what advantage is gained by having two sorts of island? If it is considered that, the islands are essential and their lack of methylation is important then it could be that a small fraction of genes (those with CWG islands) remain active when the majority are inactivated by blanket CG methylation.

CHAPTER FIVE

Isolation and characterisation of the cDNA encoding cytosine DNA methyltransferase from pea

5.1 Introduction

Methylated DNA is found in organisms ranging from bacteria to plants and mammals. In eukaryotes, the most commonly found (if not the only) modified base in DNA is 5-methylcytosine. The level of methylation varies from 3-8% of the cytosines in vertebrates (Shapiro, 1975) and as high as 30% in plants (Adams and Burdon, 1985). This large proportion of methylated cytosines in plants could be attributed to the large fraction of repetitive DNA which is usually highly methylated. The second contributing factor is the presence of methylcytosine in the mCNG trinucleotide in plant DNA (Gruenbaum *et al.*, 1981). Thirdly there is little or no depletion of CG and CNG sequences. Though the exact function of DNA methylation is not understood, methylation is implicated in the regulation of a number of cellular processes including transcription (Busslinger *et al.*, 1983; Cedar, 1988; Boyes and Bird, 1991), developmental regulation (Antequera *et al.*, 1987), genomic imprinting (Reik *et al.*, 1987), mutagenesis (Cooper and Youssoufian, 1988), transposon mobilisation (Fedoroff, 1989), transgene inactivation (Weber *et al.*, 1990) and chromatin organisation (Bird, 1986; Lewis and Bird, 1991; Kass *et al.*, 1993).

Cytosine 5 methyltransferases (m5C-MTases) catalyse the transfer of a methyl group from the cofactor, S-adenosyl-L-methionine (AdoMet), to C5 of cytosine (Wu and Santi, 1987). All the m5C-MTases share a set of well conserved amino acids motifs; designated I to X (Kumar *et al.*, 1994). It is believed that the strong evolutionary conservation is essential for common functions, namely; binding AdoMet; recognising, flipping out and forming a covalent linkage to C6 of the target cytosine; transfer of the methyl group from AdoMet to the C5 of the target cytosine; release of AdoHcy and cleavage of the covalent link to the C6 of the target cytosine. From the co-crystal structure of the DNA-M.*Hha*I-AdoHcy complex, the significance of the conserved motifs became clear (Klimasauskas *et al.*, 1994). As speculated, motif I (F-G-G) is intimately involved in the binding of AdoMet along with motif X. The conserved motif IV (PC) was positioned close to AdoMet in a groove that had appropriate dimensions to accommodate the target cytosine. In the presence of a 5-fluoro-substituted substrate and AdoHcy the extended loop region that contains motif IV (including the catalytic cysteine residue) undergoes a shift of 180° that results in movement of the cysteine towards the DNA (Klimasauskas *et al.*, 1994). Motif VI (ENV) plays an essential structural role in the positioning of the PC dipeptide of motif IV. The motif VII (Q-R-R) probably makes non-specific contacts neutralising the charges of the phosphodiester bond.

In contrast to the prokaryotic MTases the eukaryotic enzymes have an additional N-terminal domain that makes up the first two thirds of the protein (Bestor *et al.*, 1988). The murine MTase N-terminal domain has a nuclear localisation signal, a cysteine-rich Zn binding site and a region responsible for targeting the enzyme to DNA replication foci. Some of the amino acids in this region also help the enzyme to discriminate between unmethylated and hemimethylated target sequences in DNA. In

the mammalian enzyme, the N and C terminal domains are joined by repeats of the dipeptide, GK (Bestor *et al.*, 1988). Similar structural domains are also observed in the MTase of *Arabidopsis* (Finnegan and Dennis, 1993). This plant enzyme has only eight conserved motifs and shows a high degree of homology with the C-terminal of the mammalian protein. It has been proposed that a family of MTase genes are present in *Arabidopsis*.

The high level of methylation is not a characteristic of all plants. Between *Arabidopsis* and pea there is a six-fold difference in the proportion of DNA cytosines methylated (Leutwilder *et al.*, 1984). In peas (as we have recently reported; Pradhan and Adams, 1995) there are two distinct MTases; one methylating CG and the other CWG target sequences . The CWG m5C-MTase is capable of methylating CAG and CTG sites in an *in vitro* assay system and we found no evidence that either enzyme is capable of methylation the 5′ cytosine in CCG sequences. To help better understand the complexity and evolution of the plant enzymes, I have cloned the cDNA for m5C-MTase from *Pisum sativum*. The sequence homology with other eukaryotic enzymes and the tissue specific expression pattern is also reported in this chapter.

5.2 Identification and cloning of the cDNA encoding pea 5mC-MTase

CWG DNA MTase from 5 days old pea shoot apices was purified as described by Pradhan and Adams, (1995). The purified proteins were blotted on to the PVDF membrane. Protein bands at 110 kDa and 100 kDa were submitted for sequencing to the SERC protein sequencing facility in Aberdeen, UK. The N terminal amino acid sequence of the 110 kDa protein was used to design a degenerate oligonucleotide probe (Fig. 5.1).

a: (110 kDa)	Amino acid sequence:	D A R T G D K
	Degenerate probe:	GAIGCIMGIACNGGNGAYAA
	Deduced amino acid sequences:	K A R T G L K
b: (100 kDa)	Amino acid sequence:	D N A K E
	Deduced amino acid sequences:	D K A K E

Figure 5.1 Amino acid and nucleotide sequences used in identification and cloning of clone p3 and identification of (a) 110 and (b) 100 kDa peptides

The amino acids sequences are those obtained by sequencing the purified protein. The deduced peptide sequence is that obtained from the sequence of cloned cDNA. Underlined amino acids are where the two sequences differ. The degenerate probe used is also indicated. The abbreviation for codon usage are as follows: I (inosine), N (A, T, G, C), Y (A, T), M (A,C).

The 5´ end labelled probe was used to screen a λgt 11 library containing cDNA inserts made from mRNA of 5 days old etiolated pea seedlings (Clontech, USA). The screening conditions are described in Section 2.2.7.2 using tetramethylammonium chloride based hybridisation solution (Section 2.1.6). An initial screening of 1.5 x 10^6 PFU led to the isolation of clone p3 (Fig. 5.2). This clone also strongly hybridised with a 1.8 kb fragment corresponding to the total conserved domain from *Arabidopsis* (the*Pst*I-*Bam*HI fragment of YC8 clone kindly provided by Jean Finnegan).

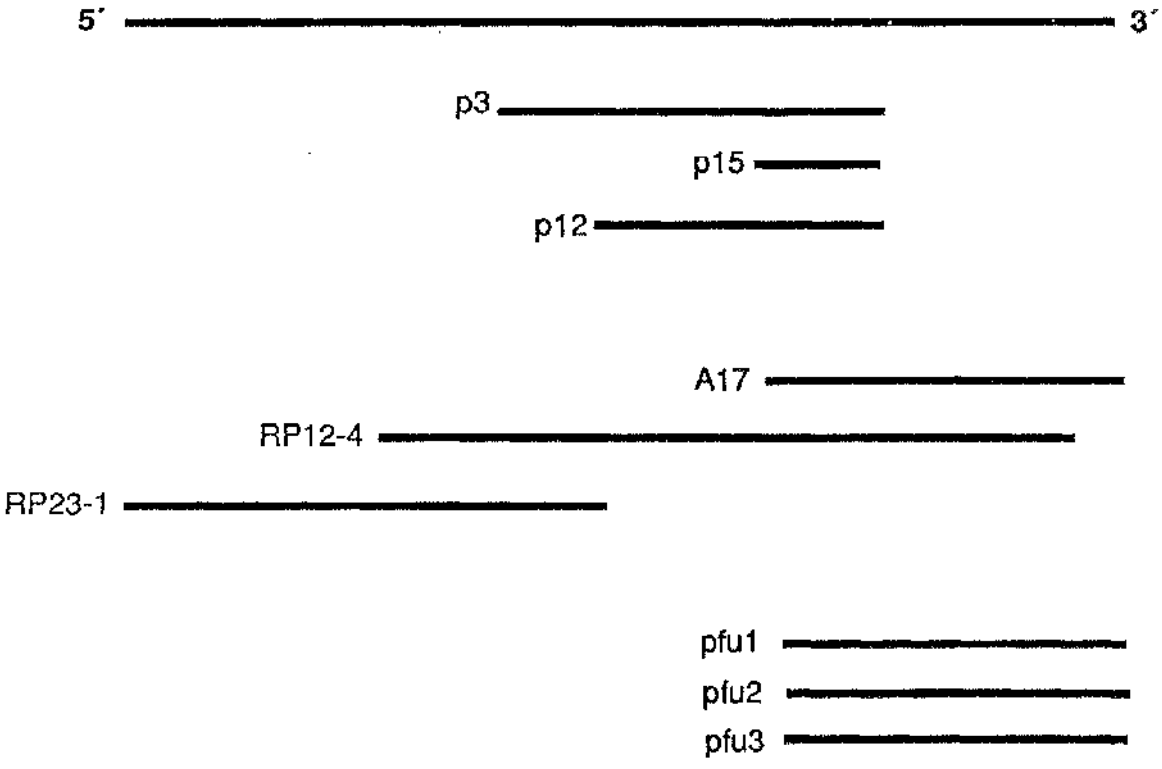


Figure 5.2 Diagram of isolation of overlapping cDNA clones for pea cytosine methyltransferase

The top line represents a line drawing of methyltransferase cDNA. Clones p3, p15 and p12 were isolated from a λ gt11 library. Clone A17 was from a oligodT primed library. Two clones RP23-1 and RP12-4 were from a random primed library. Pfu clones were obtained by M. Cummings by 3´ RACE.

The 1.8 kb cDNA insert from clone p3 was excised using *Eco*RI and ligated into the *Eco*RI cloning site of pBluescript (pBKS II). Sequencing showed that clone p3 did contain a 1.8 kb Section homologous to *Arabidopsis* MTase clone YC8. A second round of screening was done using fragments from the 5′ end (a 225 bp *Eco*RI-*Sca*I fragment), or the 3′ end (a 540 bp *Eco*RV-*Eco*RI fragment) of the 1.8 kb insert from clone p3. From 1.5 x 10^6 PFU, two more positive clones (p12 and p15) were obtained (Fig. 5.2). Sequencing of these clones revealed that neither of them have sequences extending 5′ to the region corresponding to conserved region IV. These results led us to believe that full length MTase cDNA clones could well be under-represented in this library. This is probably because the library had been amplified (Clontech catalogue no: FL1101a). In order to obtain the total cDNA sequence, random primed and oligo dT primed primary λZapXR libraries from 4-day old green shoot apices were constructed and screened simultaneously using the 5′ and 3′ specific probes from clone p3 . A series of overlapping clones were obtained (Fig. 5.2). After 5′ and 3′ end sequencing and restriction mapping of each clone, three clones RP23-1, RP12-4 and A17 were identified as overlapping clones and were selected for sequencing of both the strands (Appendix-III). The sequence of the 5′end of RP23-1 is also found in the genomic clone (data not presented).

The assembled nucleotide sequences from clones RP23-1, RP12-4 and A17 represent a message of about 5.0 kb (Appendix-II) and include 73 nucleotides of 5′untranslated region and a poly A tail of 21 nucleotides. Northern analysis of poly A+ message probed using the 1.8 kb cDNA insert from p3 indicates the size of the MTase transcript to be 4.9 kb, which agrees with the size of the assembled sequence. The deduced amino acid sequence has an open reading frame of 1560 amino acids (Fig. 5.3; for detail please see Appendix-II).

```
   1  MGSASLLNPS DSSLPGGKDS TSKEEPVSNT EGEVMAGGKQ KKRSLSESSE

  51  QPAPTRKVAK RSASAASKNL EEEVFFHILI SLVLLKLRRI RLQKENCSQS

 101  AMTAGQEDDR PNRRLTDFIL HDESGAAQAL EMLEIKDLFI TGLILSQKEM

 151  LTRKKSQGVR CHGFGRIESW DISGYEDGSP ADMDFLAEIA DYDCQKPAGT

 201  YKKYYDLFFE KARACLEVYK KLAKSSGGDP DISLDELLGW HGTGNEWSKY

 251  FSGTVSLKEF IISQGDFIYK QLIGLDTMLK ANDKGFEDIP ALIALRDESK
                poty virus NIa-NLS     SV40-NLS / P
 301  KQAHFAKHTS AGHQMRLYGL VRELSMKRKR HQMDSVDEED EDEDAKLARL
                TGA-1B-NLS
 351  LLDEEYWKSN RQRKNSRIII FILIKFYIKI NEDEIANDYP LPAYYKTSLQ

 401  ETDEFIVFDN DCDIYDTERS SRSMLHNWAL YNSDSRLISL ELLPMKPCSE

 451  MDVTIFGSGT MTSDDGSGFN LDTEAGQSSV ASGAQDTDGI PIYLSAIKEW

 501  MIEFGSSMVF ISIRTDLAGI GLGKPSKQYT PWYDTVLKTA RIAISIITLL

 551  KEQSRVSRLS FPDVIKKVSE YTQDNKSYIS SDPLAVERYI VVHGQIILQL
                                                  P
 601  FAEFPDDKIR KSPFVTGLMN KMEERHHTKW LVKKKKLSPK SEPNLNPRAA
                                          Acidic motif
 651  MAPVVSKRKA MQATATKLIN RIWGEYYSNH LPEESKEGTA IEEKDDDEAE

 701  EQEENEDEDA EEETVLLEET LKPRIVSKQI KAFSDDGEVR WEGVPERKTS
                                    Zn binding motif
 751  SGLPLYKQAI IHGGSCFCGN ICVSRKLMNQ MSFLIYITLN ICLNPKNGEK

 801  MFHGRMMQHG CHTVPWQSRS EREVFLTNEC RDLGLQDVKQ INVASIRKTP
                      BP-NLS       P          pea110
 851  WGHQHRKASN AAGKIDRERA DERKKKGLPT EYYCKARTGL KRGAFFSLPF
                             pea100
 901  DTLGLGSGVC HSCNIQEADK AKEIFKVNSS KSSFVLEGTE YSLNDYVYVS

 951  PFEFEEKIEQ GTHKSGRNVG LKAFVVCQVL EIIAKKETKQ AEIKSTELKV

1001  RRFFRPEDVS SEKAYCSDVQ EVYFSDETYT ISVQSVEGKC EVRKKIDIPE
                                                  SV40-NLS / P
1051  GSAPGAFHNV FFCELLYDPA TGSLKKLPSH IKVKYSSGPT ADNAARKKKG
                                        I
1101  KCKEGDSISV PDLKSKTSNE NCLATLDIFA GCGALSEGLH KSGASSTKWA
                II
1151  IEYEEPAGNA FKANHPEALV FINNCNVILR AIMEKCGDID ECISTAEAAE
                                    IV
1201  LASKLDDKDL NSLPLPGQVD FINGGPPCQG FSGMNRFNTS TWSKVQCEMI
                          VI                          VII
1251  LAFLSFADYF RPRYFLLENV RNFVSFNKGQ TFRLTLASLL EMGYQVREGI
                VIII
1301  LEAGAFGVSQ SRKRAFIWAA SPEDVLPEWP EPMHVFSAPE LKITLAENVQ

1351  YAAVCSTANG APLRAITVRD TIGELPAVGN GASRTNMEYQ SDPISWFQKK
```

```
1401   IRGNMAVLTD HISKEMNELN LIRCQKIPKR PGCDWRDLPD EKIKLSTGQL

1451   VDLIPWCLPH TAKRHNQWKG LFGRLDWQGN FPTSITDPQP MGKVGMCFHP
                IX                                    X
1501   DQDRILTVLE CARSQGFPDH YQFSGNIIHK HRQIGNAVPP PLAFALGRKL

1551   KEALDSKSAN *
```

Figure 5.3 Translated amino acid sequence of pea cytosine DNA methyltransferase

Sequence domain and functional elements are underlined. The putative function is indicated on top. Putative proteolytic sites are in bold and indicated with a P. The Zn binding motif is highlighted by the putative C and H aminoacids. The following abbreviations are used: NLS, nuclear localisation signal; BI, bipartitate; SV40, SV40 like. The conserved motifs are underlined and the motif number is on top. Sequences similar with pea 110 and 100 kDa proteins are underlined. Please note that some of the putative functional domain overlap with each other. Clone p3 is between aminoacid sequence 647-1227.

There is a translation start site with a high degree of identity with the Kozak sequence; a consensus sequence required for initiation of translation in higher eukaryotes (Kozak, 1986). The putative Kozak sequence is ACTCAUGG compared to the consensus A/GCCAUGG with AUG as the initiating methionine. The first stop codon (UAA) in this frame is at position 4752 of the assembled cDNA sequence. Short open reading frames are present in the other frames and are not associated with a potential Kozak sequence. The first two-thirds of the enzyme represent the amino terminal domain and the last one-third the catalytic domain (Fig 5.4).

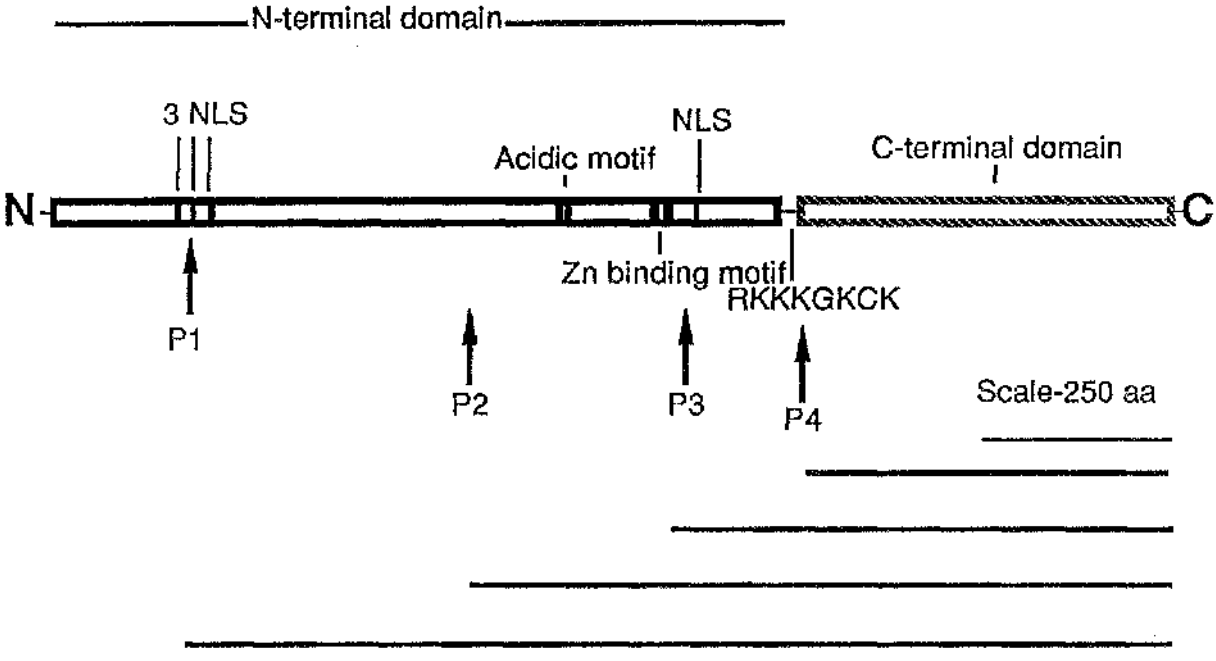


Figure 5.4 Diagram showing the functional domain and putative proteolysis sites on the pea cytosine DNA methyltransferase

The protein is divided into N-terminal and C-terminal (catalytic) domains. NLS represents nuclear localisation signals. P1-4 show the putative proteolysis sites on the protein. For detail see Fig. 5.3

5.3 Sequence comparison with other eukaryotic DNA MTases

Eukaryotic DNA MTase sequences from mouse (Bestor *et al.*, 1988), human (Yen *et al.*, 1992) and *Arabidopsis* (Finnegan and Dennis, 1993) have been cloned. The cDNAs are all about 5 kb in size, potentially encoding a protein of mass about 180 kDa. The size of the pea MTase c DNA is similar. The inferred amino acid sequence shows the presence of eight conserved motifs that are homologous and present in the same order in pea as in other eukaryotes and prokaryotes (Fig. 5.5).

	motif-I	motif-IV
Pea:	LATLDIFAGCGALSEGLHKSGA	LPGQVDFINGGPPCQGFSGMNRFN
Arabidopsis:	LATLDIFAGCGGLSHGLKKAGV	LPGQVDFINGGPPCQGFSGMNRFN
Mouse:	LRTLDVFSGCGGLSEGFHQAGI	QKGDVEMLCGGPPCQGFSGMNRFN
Human:	LRTLDVFSGCGGLSEGFHQAGI	QKGDVEMLCGGPPCQGFSGMNRFN
EcoRII:	FRFIDLFAGIGGIRKGFETIGG	HVPDHDVLLAGFPCQPFSLAGVSK
	F G G G	G PCQ FS

	motif-VI	motif-VIII
Pea:	RPRYFLLENVRNFVSFNK	ILEAGAFGVSQSRKRAFIWAA
Arabidopsis:	RPRYFLLENVRTFVSFNK	ILEAGAYGVSQSRKRAFIWAA
Mouse:	RPRFFLLENVRNFVSYRR	VLQAGQYGVAQTRRRAIILAA
Human:	RPRFFLLENVRNFVSFKR	VLQAGQYGVAQTRRRAIILAA
EcoRII:	KPAIFVLENVKNLKSHDK	KVIDGKHFLPQHRERIVLVGF
	F ENV S	Q R R

	motif-IX	motif-X
Pea:	DRILTVRECARSQGFPDHY	HKHRQIGNAVPPPLAFALGRKLKEA
Arabidopsis:	HRILTVRECARSQGFPDSY	HKHRQIGNAVPPPLAFALGRKLKEA
Mouse:	HRVVSVRECARSQGFPDSY	DRHRQVGNAVPPPLAKAIGLEIKLC
Human:	HRVVSVRECARSQGFPDTY	DKHRQVGNAVPPPLAKAIGLEIKLC
EcoRII:	PRRLTPRECARLMGFEKVD	QSYRQFGNSVVVPVFEAVAKLLEPY
	RECAR F	GN P A

Figure 5.5 Conserved motif sequence comparison of Methyltransferases

The accession numbers for methyltransferases from GenEMBLwere following *Arabidopsis* (L10692), mouse (X14805); human (X77485) and M.*Eco*RII (X05050). The consensus amino acids are written below each column

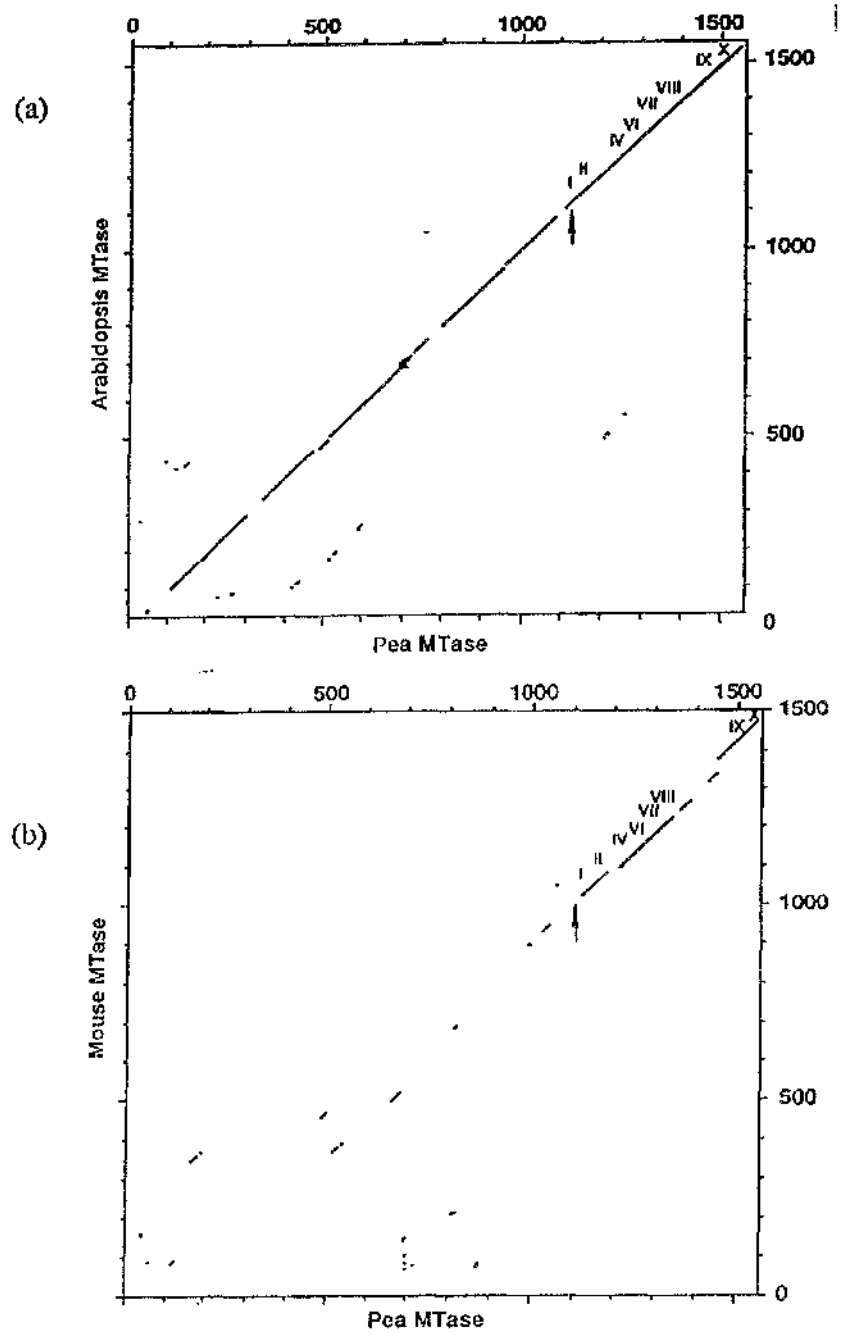


Figure 5.6 Dot plot comparisons of DNA methyltransferases

Comparisons of pea DNA methyltransferase with a: Arabidopsis MTase, b: mouse MTase. The stringency of search was 18 in a window size of 30. The fusion point of N- and C-termini is shown by an arrow. The conserved motifs are marked with roman numerals.

At the amino acid level, pea enzyme shares 67% overall identity with the *Arabidopsis* MTase. The N terminal two-thirds of the protein has very poor identity with the mammalian enzyme. However, the catalytic domain shares 50% identity with that of the mammalian enzymes and over 80% identity with *Arabidopsis* (Fig. 5.6a). Conserved motif I, the AdoMet binding domain, has the consensus fingerprint FxGxG, identical with other MTases (Kumar *et al.*, 1994). A proline-cysteine (PC) dipeptide is present in the conserved motif IV that has been conclusively identified as the functional catalytic motif of m5C MTase by crystallographic study of M.*Hha*I (Klimasauskas *et al.*, 1994) and by the use of the suicide substrate, FdC, for M.*Eco*RII (Friedman and Ansari, 1992). The variable region between conserved motifs VIII and IX determines the sequence specificity of methylation in prokaryotic m5C MTases and is often referred as the target recognition domain (TRD). Hybrid swap experiments in the mono-specific MTases established that not only did the variable region determine the sequence specificity, but also the choice of the specific base to be methylated. For a thorough review, see Noyer-Weider and Trautner (1993). Prokaryotic enzymes showing similar target specificity share strong homology in this region, whereas enzymes with different target specificity have little or no homology (Walter *et al.*, 1990; Szilak *et al.*, 1990). The TRD of the pea enzyme is 90% identical with that of *Arabidopsis* and has very little homology with either of the mammalian enzymes.

The amino terminal domain of the pea enzyme is 49% identical to the corresponding region of the *Arabidopsis* protein and lacks significant homology with the mammalian enzymes (Fig. 5.2b). There are several clusters of the basic amino acids lysine and arginine that could serve as nuclear localisation signals (NLS) for plant proteins (Raikhel, 1992). The amino acid cluster between 299-332 is similar to the NLS of potyvirus, NIa (Carrington *et al.*, 1991). It exhibits the characteristic bipartitate

structure of a spacer of 25 amino acids flanked by two regions of basic amino acids (KK-25aa-KRK). A similar bipartitate NLS is also present between amino acid 856-874. There are also short stretches of basic amino acids in regions 327-332 (KRKRH) and 1095-1101 (RKKKGK) on the pea MTase. These sequences have strong identity with basic amino acids KKKRK of the NLS of SV40 virus that can act as nuclear targeting sequences in tobacco (Van der Kroland and Chua, 1991). The amino acids between 345 and 364 have 41% identity with the NLS, TGA-1B, of tobacco (Van der Krol and Chua, 1991). In addition to these basic regions there is an acidic region of 37 amino acids, between 682 and 719, that contains 17 glutamic acid and 5 aspartic acid residues. This region is present in the *Arabidopsis* but not in the mammalian methyltransferase. In the mammalian enzymes a Zn binding domain has been identified (Bestor, 1988) but this domain is absent from the *Arabidopsis*.MTase. However, a putative Zn binding domain with amino acid configuration $CXCX_3CX_{19}CX_{10}HX_5H$ is present in the pea protein (for a detailed diagram see Fig. 5.3).

5.4 3′ RACE and PCR amplification of the conserved motifs

I have previously separated two DNA MTase activities from the nuclei of pea (Pradhan and Adams, 1995). The CG MTase is believed to be a 140 kDa protein but is unstable, readily breaking down to a 55 kDa form. The CNG (CWG) enzyme has a size of 110 kDa, with a possible breakdown product of 100 kDa. In *Arabidopsis*, Finnegan and Dennis (1993) reported a family of homologous genes for m5C MTase and so it is of immense importance to understand whether two genes and/or mRNAs are present or whether both the enzymes can be coded for by a single mRNA.

The obvious hypothesis is that the CG and CWG MTases have different TRDs and so a 3´RACE experiment was designed by M. Cummings to amplify the sequence between conserved motif II to X. At least two distinct bands were amplified of 1400 and 1320 bp. The amplified material was cloned and four different clones sequenced. All four clones are identical at the nucleotide level except that they differ in the length of the 3´ untranslated region (i.e. they differ only in their 3´ polyadenylation signals). A GU-rich far upstream element (FUE) with the consensus UUGUA is present betw4800-4890 (Morgen *et al.*, 1992) and there are at least two putative polyadenylation signals at position 4903, and 4938. The full length cDNA clone corresponds to the 3´ RACE product of intermediate length.

To study the catalytic motif in particular, a similar approach was employed using degenerate oligonucleotides for conserved motifs IV and VI to amplify this region from 1st strand cDNAs. PCR clones were sequenced and all had the same nucleotide sequence. A Southern blot using a 470 bp *Pst*I-*Pst*I probe corresponding to the conserved region between domains IV and VIII showed a single major band of 0.8 kb when pea genomic DNA was cut with *Pst*I. This indicates the presence of a single gene with intron(s) of about 300 bp. Single major bands were also seen in genomic DNA digested with*Pvu*II, *Eco*RI, *Sph*I (Fig. 5.7). These enzymes were chosen as none have a recognition site within the probe.

5.5 Developmental and tissue specific expression of the 5mC MTase

Northern blot analysis using, as probe, the 1.8 kb cDNA insert from clone p3 highlights a single 4.9 kb message in the poly A+ RNAs (Fig. 5.8).

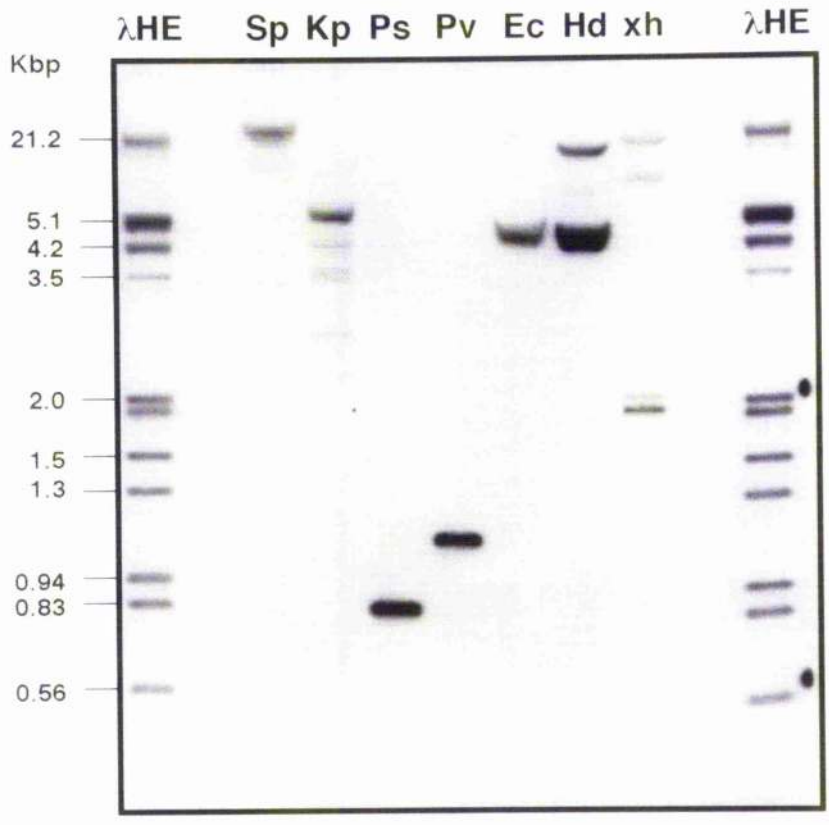


Figure 5.7 Southern analysis of pea DNA

DNA digested with restriction enzymes and probed with a 400 bp region corresponding to that between conserved motif IV-VIII. The abbreviation for restriction endonucleases are: Sp, *Sph*I; Kp, *Kpn*I; Ps, *Pst*I; Ec, *Eco*RI; Hd, *Hind*III; xh, *Xho*I. End labelled . End labelled λ DNA digested with *Hind*III and *Eco*RI (λHE) was used as marker.

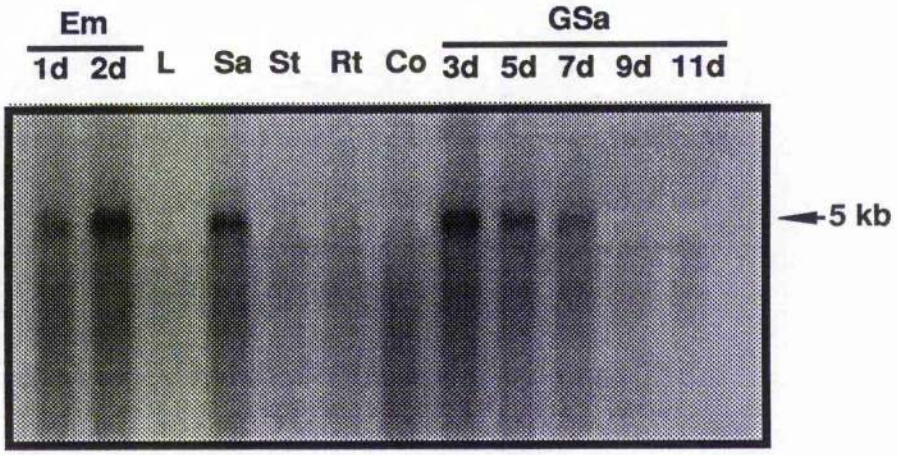


Figure 5.8 Tissue specific MTase gene expression

Poly A+ RNA (3 μg) from different tissues was blotted and probed with a 1.8 kb MTase cDNA clone. RNA samples from embryo (Em), leaf (L), shoot apices (Sa), stem (St), root (Rt), cotyledon (Co) and green shoot apices (GSa) were blotted. The age of the tissue is shown in days (d). Please note that the 5 kb MTase specific message varies with age and stage of development in the tissue

The MTase specific message is more abundant in the meristematic tissues, where rapid cell division and DNA synthesis takes place. Cells in leaf, hypocotyl and mature roots do not have any detectable mRNA Fig. 5.8) though root tips taken 2 and 3 days after imbibition do show the presence of the mRNA (not shown). The steady state level of the mRNA in the shoot apex falls with development and is barely detectable on the 11th day after germination. This time course parallels the changes observed for enzymic activity (Houlston *et al.*, 1993).

5.6 Discussion

Based on the homology with other eukaryotic DNA MTases it is clear that the assembled cDNA sequence and the corresponding translated protein described in this paper is a m5C-MTase . The amino acid sequence obtained from the N-terminal sequence analysis of 110 kDa and 100 kDa is also found in the translated protein. Like the other known eukaryotic DNA MTases, the pea enzyme has eight of the ten conserved motifs characteristic of m5C MTases. However, sequence comparison of TRDs of this and the *Arabidopsis* enzyme with known mammalian CG MTase (human and murine) shows poor homology, suggesting that the plant and mammalian enzymes may methylate different sequences. Yet there is also no obvious homology with the group of bacterial enzymes that have a CWG in their target recognition sequence.

In mammalian enzymes, the N terminal and the C terminal domains are separated by a linker of 13 alternating lysine-glycine residues as compared with a shorter sequence in *Arabidopsis* and pea where the domains are separated by only six or eightamino acids (RKKKGKCK). The most striking feature of the pea enzyme is the presence of clusters of the basic amino acids lysine and arginine at several places

within the N terminal domains. As well as possibly being signals for nuclear localisation, these sequences could be targets for processing by endogenous proteases. Thus, the translated protein could be proteolytically processed to smaller species keeping the conserved MTase domain intact (Fig. 5.4). A closer look at the peptide sequence suggests that, if the proteolysis occurs at the linker, the smallest possible peptide generated from this protein will be of 55 kDa; the size of one of the peptides present in the purified CG enzyme preparation from pea, wheat and rice (Pradhan and Adams, 1995; Theiss *et al.*, 1987; Giordano *et al.*, 1991). In the event of processing occurring at the sequence RKKK (aa 873-876), 684 amino acids in from the N terminus, a protein of predicted size of 80 kDa will be generated and within 6 amino acids of this site there occurs an amino acid sequence that shows 85% identity to the sequence, determined by microsequencing techniques, for the N-terminus of the CWG MTase. The discrepancy in the molecular weights between the predicted protein and that of the isolated enzyme (110 kDa) could be due to intracellular glycosylation or phosphorylation thus altering its apparent molecular weight. Alternatively , the protein sequence obtained could have come from a futrher proteolytic fragment derived following cleavage of a 110 kDa protein generated by cleavage at KKKK (aa 633-636). 230 amino acids in from the N-terminus there is the sequence KRKR (AA-327-330) and processing at this site could lead to the production of a 140 kDa CG MTase. All these proteins would retain the linker regions that could act as a nuclear localisation signal. However, this would be lacking from the 55kDa protein so that, if this is generated intracellularly, it would be unlikely to be transported to the nucleus. Proteolysis at the linker may represent a mechanism of disabling the intracellular functioning of the MTase mediated by its loss from the nucleus.

In the murine enzyme, both the protein domains are reported to fold independently and proteolytic cleavage leads to the isolation of an intact C-terminal on SDSPAGE (Adams *et al.*, 1983). It is not clear whether the native enzyme falls apart following cleavage of the linker but the enzyme retains some activity. The specificity is, however, changed as cleavage stimulates *de novo* methylation several fold, whereas maintenance methylation rate does not increase (Adams *et al.*, 1983; Bestor, 1992). This indicates a role for amino terminal sequences in the down regulation of *de novo* enzymatic activity (Bestor, 1987). Thus, as with the prokaryotic type I DNA MTases, the amino acid sequences recognising the methylation status of the complementary strand are distinct from those at the catalytic centre. Nonetheless, they must interact with the same short sequence of DNA and so must affect the conformation of the active site of the enzyme. The ability to bring about this change could equally well affect the target specificity of the enzyme such that, depending on the length of the N-terminal domain, the enzyme methylates either CG or CWG target sequences.

It has been suggested that the mammalian MTases have evolved by fusion of two ancestral genes, one with MTase activity and the other a sequence specific DNA binding protein (Bestor, 1990). The finding that the N terminal domain of the pea protein has poor homology with mammalian MTase and the acidic pocket of the plant enzymes is not present in mammalian enzyme suggests that both the genes could have evolved separately via different gene fusion events. The driving force for the fusion event would primarily be the addition of a NLS but the presence of a DNA interaction centre (the Zn binding domain) could have been instrumental in the evolution of a maintenance methyltransferase activity.

The expression pattern of DNA MTase is confined to the actively dividing tissues where DNA synthesis rate is high. The steady state level of mRNA in young pea seedling is similar to that for MTase activity during plant growth and development (Yesufu, *et al.*, 1991). The expression of this gene is also stage dependant indicating the basic regulation of this gene could be at the transcriptional level. However, protein processing may not be identical at all developmental stages leading to differing ratios of the two enzymic activities at different stages. This hypothesis has not been tested yet. Recently the murine enzyme has been expressed both in microbial as well as insect cell system (Fraser and Reich, 1995; Tollefsbol and Hutchison, 1995). In future it should be possible to dissect the various structural domains of the initial translation product generated in heterologous systems and study its processing by plant extracts.

CHAPTER SIX

General discussion

The most common modification in the plant genome is methylation of cytosine residues in carbon 5. Plant DNA methylation in plants is different from in mammals in two ways. The first difference is the percentage of methylated cytosine that may be as much as 6 fold when compared with vertebrate DNA. Secondly both CG and CNG nucleotides act as target sequences for methylation to occur. In plants over 80% of CG and CNG were found to be methylated (Gruenbaum *et al.*, 1981), in contrast to animals where methylation is mainly in the CG dinucleotides. Before I started my work there were a few reports available on partial purification of DNA methyltransferases from wheat (Theiss *et al.*, 1988), rice (Giordano *et al.*, 1991) and pea (Yesufu *et al.*, 1991). The reported size of these enzymes varies from 35 kDa (wheat) to 160 kDa (pea) which was confusing. Thus, it was of considerable interest to understand the enzyme and the molecular mechanism involved in CG and CNG methylation. Pea plant was chosen as a model system because the seeds are readily available and because of the ease of growth and relatively ease of obtaining large amounts of shoot apices that I used for my research work.

Fractionation of low salt extracted nuclear proteins has allowed the purification of distinct CG and CNG MTases from pea. The use of synthetic oligonucleotide substrates has allowed their separate assay in the mixture of the two enzymes. The target specificity for both the enzymes was confirmed by the Maxam-Gilbert chemical

sequencing method. These findings demonstrate two different enzymes in a single pea cell. The presence of two different MTases in the same cell is very similar to the situation in *E. coli* where multiple methyltransferases (i.e. a type I, a type II or a Dam methyltransferase) are found. Different type II MTases have the capacity to methylate cytosine or adenine to produce N-4 methylcytosine, C-5 methylcytosine or N-6-methyladenine. There is another interesting situation, where a single MTase recognises more than one specific sequence. These are called multispecific MTases. Several *Bacillus* bacteriophages encode such mutispecific MTases. These enzymes are single polypeptides as are the monospecific enzymes but have the versatile property of methylating more than one target sequence. One such example is M.ϕ3TI, which recognises and methylates GGCC tetranucleotide and GCNGC pentanucleotide sequences (Weidner *et al.*, 1983). Subsequently Trautner's group has shown that a single mutation in the variable region of M.ϕ3TI can abolish the function of the enzyme to recognise one of the specific sequences, while still allowing the other sequences to be recognised (Balganesh *et al.*, 1987; Wilke *et al.*, 1988). The variable region or TRDs are linked to a single active site for methyltransfer. Hybrid MTases have been constructed containing domains from more than one enzyme (Trautner *et al.*, 1988; Klimasauskas *et al.*, 1990; Mi and Roberts, 199). So what is responsible for the CG and CNG substrate specificity of pea MTase? Could there be two different genes for two enzymes as hypothesised by Finnegan and Dennis, (1983) in *Arabidopsis*. A series of 3'RACE experiments was done to amplify the region between conserved domain II and end of the cDNA (poly A+ tail of the message). None of the amplified products sequenced so far have any difference between conserved domains VIII-IX. Southern blotting of the genomic DNA with a probe between domain IV-VIII proved that pea has only one gene for methyltransferase. All the above observations suggested

that pea MTase is coded by one gene and possibly the target recognising function cannot lie completely in the region between motif VIII-IX.

At this point a close look at the methylating activity of the vertebrate DNA methyltransferases may be considered. Though the mammalian MTase is known as a CG MTase, it exhibits a low level of methylation on CA and CT dinucleotides (Hubrich-Kühner *et al.*, 1989; Adams *et al.* 1993, Clark *et al*,. 1995). Efficient *in vitro* methylation of CC dinucleotides by vertebrate MTase has not been reported. This CA and CT methylation may be as a result of the enzyme showing some activity with CAG and CTG sequences. The mouse enzyme shows significantly greater activity with hemimethylated CAG/MTG sequences compared with CAG/CTG sequences (Adams *et al.*, 1993). Clark *et al.*, (1995) demonstrated that the methylation machinery in mammalian cells is capable of maintenance as well as *de novo* methylation of CpNpG sites.

In the mouse a single MTase gene has been detected. This could suggest that the CpG methylation pattern is established and maintained by a single species of enzyme. However in murine erythroleukemia (MEL) cells at least two different closely related enzymes were detected in early experiments (Bestor and Ingram, 1985). Such differences could be as a result of alternative splicing of hnRNA, protein modification or proteolysis. Bestor *et al.*, (1988) have detected only one message. Purified pea CG and CNG enzymes exhibit a native molecular weight of 140 and 110 kDa. On SDS/PAGE the CG enzyme shows 140 and 55 kDa peptides whereas the CNG enzyme is 110 kDa, indicating proteolysis during purification or electrophoresis. The message for pea MTase is about 5 kb and the assembled cDNA code for a protein of 1560 amino acids. The estimated molecular weight of such a protein would be 180 kDa, similar to

the other eukaryotic MTase reported so far (Bestor *et al.*, 1988; Yun *et al.*, 1992; Finnegan and Dennis, 1993). The amino terminal of the translated protein has several nuclear localisation signals preceded by putative proteolysis sites as described in chapter five. In the event of proteolysis some of these peptides will have the ability to transport across to the nucleus with the intact catalytic domain. Thus there is a strong argument in favour of the proteolytic processing of MTase in pea. Indeed the N terminal sequence analysis of CNG MTase matches at position 875 of the predicted protein. However, non of these hypotheses has been tested so far. One way of finding out is to obtain massive amount of the intact pea protein through over expression. Such a task will not be impossible, since the mammalian enzyme has been successfully expressed in microbial as well as insect cell system (Fraser and Reich, 1995; Tollefsbol and Hutchison, 1995). Once the protein is over expressed it could be used for *in vitro* processing using plant extracts. The other experimental approach would be to obtain a series of deletion protein products and to test their nuclear transport as well as methylating activities on different substrates.

In an *invitro* assay system very little or no identifiable methylation in CC dinucleotides has been detected has been reports for both mammals and pea methyltransferase. This observation raises the question as to whether or not methylation of the 5′ cytosine in the CCG sequence occurs to a significant extent in peas. Such methylation has not been detected by genomic sequencing of the 1400 bp of 5′-flanking sequence of the maize alcohol dehydrogenase gene present in non-expressing leaf DNA (Nick *et al.*, 1986). Although there is no evidence of any enzymatic methylation or *in vivo* evidence for CCG methylation in peas, the CG enzyme does interact weakly with such target sites when the suicide substrate, pyrimidinone, was substituted in place of C at CG sequences (Pradhan and Adams,

1995). It is possible that there is a low level of methylation of such targets that I have not been able to detect using the techniques employed. If there is a very low level of CCG methylation then the bisulphite genomic sequencing technique could be used to examine the potential sites after *in vitro*. methylation.

The finding of different methyltransferase activities in the same cell raises the possibility that the action and control of these enzymes may be different even to the extent that methylation of CG and CWG sequences may serve completely different functions. It, therefore, becomes of interest to have further information on the biological significance of CG and CWG methylation on reporter gene expression and on the distribution of the target sequences for these methyltransferases. In preliminary observations using a CWG methylated *rbcs2*-CAT construct, I have observed inhibition of CAT expression, strengthening the conclusion that such methylation can alter gene expression. A small fraction of genes (those with CWG islands) could remain active when the majority are inactivated by blanket CG methylation. It has been estimated that 5% of the pea genes contain CWG islands (section 4.4, chapter four).

The expression pattern of DNA MTase is confined to the actively dividing tissues where the DNA synthesis rate is high. The steady state level of mRNA in young pea seedlings parallels to that for MTase activity during plant growth and development (Yesufu, *et al.*, 1991). The expression of the MTase gene is also stage dependant indicating the basic regulation of this gene could be at the transcriptional level. However, protein processing may not be identical at all developmental stages leading to differing ratios of the two enzymatic activities at different stages-a hypothesis yet to be tested.

To summarise the work reported in this thesis I have shown two different MTases in the cells of *Pisum sativum*. The preliminary results suggest that both the enzymes could have arisen from one translation product, allowing us to speculate that pea could have an entirely different mechanism from mammals to produce two different enzymatic activities from one precursor protein of a theoretical mass of 180 kDa. Selective proteolysis is one of the most favoured mechanism, which needs to be tested. Neither of the enzymes methylates the 5´ C in a CCG sequence and such methylation in pea nuclear DNA has not been detected. I have identified a fraction of the pea nuclear genome free from methylation at the trinucleotide CWG sequences that is separate from traditional CpG islands. This fraction is named as CWG islands. The CpWpG islands are characteristically similar to CpG islands in vertebrates. An estimated 5% of the pea genes have such islands at their 5´ region. Identification of a CNG methyltransferase and unmethylated CWG islands opens up new avenues to understand the molecular mechanism of methylation mediated plant gene expression. The processing of two different enzymes from one gene product asks more question about the control mechanism and evolution of such processes in plants that could be fundamentally different from those seen in other organisms.

References

Achwal, C. W., Ganguly, P. and Sharat Chandra, H. (1984). *The EMBO J.* **3**, 263-266

Adams, R. L. P. and Burdon, R. H. (1985). *Molecular Biology of DNA Methylation*, Springer Verlag, New York

Adams, R. L. P., Burdon, R. H., McKinnon, K. and Rinaldi, A. (1983) *FEBS Lett.* **163**, 194-198.

Adams, R. L. P., Davis, T., Fulton, J., Kirk, D., Qureshi, M. and Burdon, R. H. (1984). *Curr. Top. Microbiol. and Immunol.* **108**, 143-156.

Adams, R.L.P and Burdon, R.H (1985) *Molecular Biology of DNA methylation.* Springer Verlag. New York.

Adams, R.L.P. (1990) DNA methylation. The effect of minor bases on DNA–protein interactions. *Biochem. J.* **265**, 309–320.

Adams, R.L.P., Lindsay, H., Reale, A., Seivwright, C., Kass, S., Cummings, M. and Houlston, C. (1993) Regulation of *de novo* methylation in *DNA methylation: Molecular biology and biological significance*. (J. P. Jost and H. P. Saluz, eds), 120–144, Birkhäuser, Basel.

Antequera, F,. Macleod, D. and Bird, A.P. (1989). *Cell* **58**, 507-517.

Antequera, F. and Bird, A.P. (1988). *The EMBO J.* **7**, 2295–2299.

Baldwin, G. S., Kelly, S. M., Price, N. C., Wilson, G. W., Connolly, B. A., Artymiuk, P. J. and Hornby, D. P. (1994). *J. Mol. Biol.* **235**, 545-553.

Balganesh, T. S., Reiners, L., Lauster, R., noyer-Weidner, M., Wilke, K. and Trautner, T. A. (1987) *EMBO. J.* **6**, 3543-3549.

Becker, P. B., Ruppert, S. and Schütz, G. (1987). *Cell* **51**, 435-443.

Bednarik, D. P., Duckett, C., Kim, S. U., Perez, V. L., Griffis, K., Guenthner, P. C. and Folks, T. M. (1991). *New Biol.* **3**, 969-976.

Belanger, F.C. and Hepburn, A.G. (1990). *J. of Mol. Evol.* **30**, 26–35.

Ben Hattar, J., Beard, P. and Jiricny, J. (1989). *Nucleic Acids Res.* **17**, 10179-10190.

Bestor, T. H. and Ingram, V. M. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 5559-5563.

Bestor, T. H. and Ingram, V. M. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 2674-2678.

Bestor, T. H., Laudano, , A., Mattaliano, R. and Ingram, V. (1988) *J Mol. Biol.* **203**, 971-983.

Bestor, T.H. (1990). *Phil. Trans. R. Soc. Lond. B* **326**, 179–187.

Bestor, T.H. (1992). *The EMBO J.* **11**, 2611–2617.

Bezdek, M., Koukalova, B., Kuhrova, V., and Vyskot,B. (1992). *FEBS Letts.* **300**, 268–270.

Bird, A. P. (1986). *Nature* **321**, 209-213.

Boyes, J. and Bird, A. (1991). *Cell* **64**, 1123-1134.

Boyes, J. and Bird, A. P. (1992) *EMBO. J.* **11**, 327-333.

Bradford, M.M. (1976). *Anal. Biochem.* **72**, 248–254

Bryans, M., Kass, S., Seivwright, C. and Adams, R. L. P. (1992). *FEBS Lett.* **309**, 97-102.

Buschhausen, G., Wittig, B., Grässmann, M. and Grässmann, A. (1987). *Proc. Natl. Acad. Sci. USA* **84**, 1177-1181.

Busslinger, M., Hurst, J. and Flavell, R. A. (1983) *Cell* **34**, 197-206.

Carrington, J. C., Freed, D. D. and Leinicke, A. J. (1991) *Plant cell* **53**, 3-4.

Chen, L., MacMillan, A. M. and Verdine, G. L. (1993). *J. Am. Chem. Soc.* **115**, 5318-5319.

Chen, L., MacMillan, A. M., Chang, W., Ezaz-Nikpay, K., Lane, W. S. and Verdine, G. L. (1991) *Biochemistry* **30**, 11018-11025.

Cheng, X., Kumar, S., Posfai, J., Pflugrath, J. W. and Roberts, R.J. (1993). *Cell.* **74**, 299–307.

Church, G. M. and Gilbert, W. (1984). *Proc. Natl. Acad. Sci. USA* **81**, 1991-1995.

Clark, S. J., Harrison, J. and Frommer, M. (1995) *Nature genetics* **10**, 20-27.

Comb, M. and Goodman, H. M. (1990). *Nucleic Acids Res.* **18**, 3975-3982.

Cooper,D. N. and Youssoufian, M. (1988). *Hum Genet.* **78**, 151-155.

Cross, S., Kovarik, P., Schmidtke, J. and Bird, A. P. (1991) *Nucleic Acids Res.* **19**, 1469-1474.

Deumling, B. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 338-342.

Donald, R. G. K. and Cashmore, A. R. (1990) *EMBO. J.* **9**, 1717-1726.

Doskocil, J. and Sormova, Z. (1965) *Biochem. Biophys. Acta.* **95**, 513-515.

Dretzen, G., Ballard, M., Sassone-Corsi, P. and Chambon, P. (1981) *Anal. Biochem.* **112**, 295-298.

Dunn, D. B. and Smith, J. D. (1955) *Biochem. J.* **68**, 627–636.

Ehrlich, K. C. (1993) *Biochem. Biophys. Res Commun.*, **1172**, 108–116.

Ehrlich, M., Gama-Sosa, M. A., Huang, L. H., Midgett, R. M., Kuo, K. C., McCune, R. A. and Gehrke, C. (1982). *Nucleic Acids Res.* **10**, 2709-2721.

Eick, D., Fritz, H. J. and Doerfler, W. (1983). *Anal. Biochem.* **135**, 165-171.

Erlanson, D. A.,Chen, L. and Verdine, G. L. (1993) *J. Am. Chem. Soc.* **115**, 12583-12584.

Fedoroff, N. V. (1989) *Cell* **56**, 181-191.

Felsenfeld, G., Nickol, J., Behe, M., McGhee, J. and Jackson, D. (1983). *Cold Spring Harbor Symp. Quant. Biol.* **47**, 577-584.

Finnegan, E. J. and Dennis, E.S. (1993). *Nucl. Acids Res.* **21**, 2383–2388.

Finnegan, E. J., Brettell, R.I.S. and Dennis, E.S. (1993) The role of DNA methylation in the regulation of plant gene expression in *DNA methylation: Molecular biology and biological significance.* (Jost, J.P. and Saluz, H.P., eds), Birkhaüser Verlag, Basel, pp 218–261.

Ford, J. P., Coca-Prados, M. and Hsu, M.-T. (1980). *J. Biol. Chem.* **255**, 7544-7547.

Ford, K., Taylor, C., Connolly, B. and Hornby, D.P. (1993). *J. Mol. Biol.* **230**, 779–786.

Frank, D., Keshet, I., Shani, M., Levine, A., Razin, A. and Cedar, H. (1991). *Nature* **351**, 239-241.

Friedman, S. and Ansari, N. (1992). *Nucleic Acids Res.* **21**, 3241-3248

Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L. and Paul, C. L. (1992). *Proc. Natl. Acad. Sci. USA* **89**, 1827-1831.

Gama-Sosa, M. A., Midgett, R. M., Slagel, V. A., Githens, S., Kuo, K. C., Gehrke, C. and Ehrlich, M. (1983a). *Biochim. Biophys. Acta* **740**, 212-219.

Gama-Sosa, M. A., Wang, R. Y. H., Kuo, K. C., Gehrke, C. W. and Ehrlich, M. (1983b) *Nucleic Acids Res.* **11**, 3087-3097.

Gardiner-Garden, M., Sved, J.A. and Frommer, M. (1992). *J. Mol. Evol.* **34**, 231–245.

Gey, G. O., Coffman, W. D. and Kubicek, M. T. (1952). *Cancer Res.* **12**, 264-265.

Gierl, A., Lutticke, S. and Saedler, H. (1988) *EMBO. J.* 7, 4045–4053.

Gilmartin, P.M., Sarokin, L., Memelink, J. and Chua, N-H. (1990) *Plant cell* 2, 369–378.

Giordano, M., Mattachini, M.E., Cella, R. and Pedrali-Noy, G. (1991). *Biochem. Biophys, Res. Comm.*, **177**, 711–719.

Glickman, J. F. and Reich, N. O. (1995) *Biochem. Biophys. Res Commun.*, **204**, 1003–1008.

Goldsbrough, P.B., Ellis, T.H.N. and Lomonossoff, G.P. (1982). *Nucleic Acids Res.* **10**, 4501–4514.

Gopal, J., Yebra, M. J. and Bhagwat, A. S. (1994) *Nucleic Acids Res.* **21**, 4482-4488

Gruenbaum, Y., Naveh-Many, T., Cedar, H. and Razin, A. (1981). *Nature*, **292**, 860–862.

Hank, T.,Schmidt, S. and Fritz, H-J. (1993). *Nucleic Acids Res.* **22**, 303-309.

Hansen, R. S., Ellis, N. A. and Gartler, S. M. (1988). *Mol. Cell. Biol.* **8**, 4692-4699.

Hepburn, A. G., Belanger, F. C. and Mattheis, J. R. (1987). Devel. Genetics **8**, 475–493.

Herman, G. E. and Modrich, P. (1977) *J. Biol. Chem.* **257**, 2605-2612.

Hershkovitz, M., Gruenbaum, Y., Renbaum, P., Razin, A. and Loyter, A. (1990). *Gene* **94**, 189–193.

Higurashi, M. and Cole, R. D. (1991). *J. Biol. Chem.* **266**, 8619-8625.

Holler, M., Westin, G., Jiricny, J. and Schaffner,W. (1988) *Genes Devel.* **2**, 1127-1135.

Hotchkiss, R. D. (1948). *J. Biol. Chem.* **168**, 315-332.

Houlston, C.E., Lindsay, H., Pradhan, S. and Adams, R.L.P. (1993). *Biochem. J.* **293**, 617–624.

Howlett, S. K. and Reik, W. (1991). *Development* **113**, 119-127.

Hubrich-Kühner, K., Buhk, H.-J., Wagner, W., Kröger, H. and Simon, D. (1989). *Biochem. Biophys. Res Commun.*, **160**, 1175–1182.

Iguchi-Ariga, S. M. M. and Schaffner, W. (1989). *Genes Dev.* **3**, 612-619.

Jeppesen, P. and Turner, B. M. (1993). *Cell* **74**, 281-289.

Johnson, C., Goddard, J. P. and Adams, R. L. P. (1995). *Biochem. J.***305**, 791-798.

Jones, P. A. (1984). In: DNA methylation: Biochemistry and Biological significance, A. Razin, H. Cedar and A. D. Riggs (ed) pp. 165-187. Springer-Verlag, NewYorK

Jones, P. A. and Taylor, S. M. (1980). *Cell* **20**, 85-93.

Jost, J.-P. (1993). *Proc. Natl. Acad. Sci. USA* **90**, 4684-4688.

Juttermann, R., Li, E. and Jaenisch, R. (1994). *Proc. Natl. Acad. Sci. USA* **91**, 11797–11801.

Kadonaga, J.T. and Tjian, R. (1986). *Proc. Natl. Acad. Sci. USA* **83**, 5889–5893.

Kass, S. U., Goddard, J. P. and Adams, R. L. P. (1993). *Mol. Cell. Biol.* **13**, 7372-7379

Keshet, I., Lieman-Hurwitz, J. and Ceder, H. (1986) *Cell* **44**, 535-543.

Klimasauskas, S., Kumar, S., Roberts, R. J., and Cheng, X. (1994). *Cell* **76**, 357- 369

Klimasauskas, S., Nelson, J.L. and Roberts, R. J. (1991). *Nucleic Acids Res.* **19**, 6183-6190

Klimasauskas, S., Timinskas, A,. Menkevicius, S.,Butkiene, D.,Butkus, V. and Janulaitis, A. A. (1989). *Nucleic Acids Res.* **17**, 9823-9832.

Knight, M. R. and Jenkins, G. I. (1992) *Plant Mol. Biol.* **18**, 567-579.

Kølsto, A. B., Kollias, G., Giguere, V., Isobe, K. I., Prydz, H. and Grosveld, F. (1986). *Nucleic Acids Res.* **14**, 9667-9678.

Kozack, M. (1986) *Cell* **44**, 283-

Kumar, S., Cheng, X., Klimasauskas, S., Sha, M., Posfai, J., Roberts, R. J. and Wilson, G. G. (1994). *Nucleic Acids Res.* **22**, 1-10.

Lammli, U. K. (1970) *Nature* **227**, 680-685.

Langdale, J.A., Taylor, W.C. and Nelson, T. (1991). *Mol. Gen. Genet.* **225**, 49–55.

Lauster, R. Trautner, T. A. and Noyer-weidner, M. (1989). *J. Mol. Biol.* **206**, 305-312

Leonhardt, H., Page, A. W. Weier, H. -U. and Bester, T. H. (1992) *Cell* **71**, 865- 873.

Leutwilder, L. S., Hough-Evans, B. R. and Meyerowitz, E. M. (1984). *Theor. Appl. Genet.* **194**, 15-23

Levine, A., Cantoni, G. L. and Razin, A. (1991). *Proc. Natl. Acad. Sci. USA* **88,** 6515-6518.

Levine, A., Yeivin, A., Ben-Asher, E., Aloni, Y. and Razin, A. (1993). *J. Biol. Chem.* **268**, 21754-21759.

Lewis, J. and Bird, A. P. (1991) DNA methylation and chromatin structure. *FEBS Lett,* **285**, 155–159.

Lewis, J. D., Mechan, R. R., Henzel, W. J., Maurer-Fogy, I., Jeppesen, P., Klein, F. and Bird, A. (1992). *Cell* **69**, 905-914.

Li, E., Bestor, T. H. and Jaenisch, R. (1992). *Cell* **69**, 915-926.

Lindahl, T. (1974) *Proc. Natl. Acad. Sci. USA* **71**, 3649–3653.

Mansour, S. L., Thomas, K. R. and Capecchi, M. R. (1988) *Nature* **336**, 348-352.

Maxam, A. M. and Gilbert, W. (1980). *Methods Enzymol.* **65**, 499-560.

Meehan, R. R., Lewis, J. D. and Bird, A. P. (1992). *Nucleic Acids Res.* **20**, 5085-5092.

Messeguer, R., Ganal, M.W., Steffens, J.C. and Tanksley, S.D. (1991) *Plant Mol. Biol.* **16**, 753–770.

Meyer, P., Niedenhof, I. and ten Lohuis, M. (1994). *EMBO J.* **13**, 2084–2088.

Mi, S. and Roberts, R. J. (1992). *Nucleic Acids Res.* **20**, 4811-4866.

Mi, S. and Roberts, R. J. (1993). *Nucleic Acids Res.* **21**, 2459-2464.

Miller, O. J., Schnedl, W., Allen, J. and Erlanger, B. F. (1974). *Nature* **251**, 636- 637.

Montero, L.M., Filipski, J., Gil, P., Capel, J., Martinez-Zapater, J.M. and Salinas, J. (1992). *Nucleic Acids Res.* **20**, 3207–3210.

Morgan, B. D.,MacDonald, M. H., Leggewie, G. and Hunt, A. G. (1992) *Mol. Cell. Biol.* **12**, 5406-5414.

Murray, E. J. and Grosveld, F. (1987). *EMBO J.* **6**, 2329-2335.

Murray, M.G. and Thompson, W.F. (1980). *Nucl. Acids Res.* **16**, 4321–4325.

Ngernprasirtsiri, J., Chollet, R., Kobayashi, H.,Sugiama, T. and Akazawa, T. (1989) *J. Biol. Chem.* **264**, 8241–8248

Nick, H., Bowen, B., Ferl, R.J. and Gilbert, W. (1986) . *Nature* **319**, 243–246.

Noyer-Wreidner, M. and Trautner, T. A. (1993) Regulation of *de novo* methylation in *DNA methylation: Molecular biology and biological significance.* (J. P. Jost and H. P. Saluz, eds), 39-108, Birkhäuser, Basel.

Oberle, I., Rousseau, F., Heitz, D. Kretz, C.,Devys, D., Hanauer, A., Boue, J., Bertheas, M. F. and Mandel, J. L. (1991). *Science.* 252, 1097-1102.

Oesterman, D. G., Depillis, G. D., Wu, J. C., Matsuda, A. and Shanti, D. V. (1988). *Biochemistry* **27**, 5204-5210.

Pfeifer, G. P. and Drahovsky, D. (1986) *Biochem. Biophys. Acta.* **868**, 238-242.

Pfeifer, G. P., Grunwald, S., Boehm, T. L. and Drahovsky, D. (1986) *Biochem. Biophys. Acta.* **740**, 323-330

Pfeifer, G. P., Steigerwald, S. D., Mueller, P. R., Wold, B. and Riggs, A. D. (1989). *Science* **246**, 810-813.

Pollack, Y., Kasir, J., Shemer, T., Metzger, S. and Szyf, M. (1984). *Nucleic Acids Res.* **12**, 4811-4824.

Pósfai, J., Bhagwat, A.S., Pósfai, G and Roberts, R.J. (1989). *Nucl. Acids Res.* **17**, 2421–2435.

Pradhan, S. and Adams, R. L. P. (1995). *The Plant Journal.* 7. 471-481.

Prendergast, G. C., Lawe, D. and Ziff, E. B. (1991). *Cell* **65**, 395-407.

Proffitt, J. H., Davie, J. R., Swinton, D. and Hattman, S. (1984). *Mol. Cell. Biol.* **4**, 985-988.

Razin, A. and Cedar, H. (1977). *Proc. Natl. Acad. Sci. USA* **74**, 2725-2728.

Razin, A. and Cedar, H. (1991) DNA methylation and gene expression. *Microbiolog. Rev.*, **55**, 451–458.

Razin, A. and Cedar, H. (1991) DNA methylation and gene expression. *Microbiolog. Rev.*, **55**, 451–458.

Razin, A., Szyf, M., Kafri, T., Rool, M., Giloh, H., Scarpa, S., Carotti, D. and Cantoni, G. L. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 2827-2831.

Reik, W., Collick, A., Norris, M., Barton, S., and Surani, M. A. (1987). *Nature*, **328**. 248-251

Rideout, W. M., Coetzee, G. A., Olumi, A. F. and Jones, P. A. (1990) *Science*. **249**, 1288-1290.

Riggs, A. D. (1990). *Mol. Cell. Biol.* **10**, 4987-4989.

Rogers, S.O. and Bendich, A.J. (1988) Extraction of DNA from plant tissues in *Plant Molecular Biology Manual* (Gelvin, S.B., Schilperoort, R.A. and Verma, D.P.S. eds) p A6:1–10. Kluwer Acad. Pubs., Dordrecht.

Rouleau, J., Tanigawa, G. and Szyf, M. (1992) *J. Biol. Chem.* **267**, 7368-7377.

Rubin, R. A. and Modrich, P. (1977) *J. Biol. Chem.* **252**, 7265-7272.

Saluz, H. P. and Jost, J. P. (1993). In: *DNA Methylation: Molecular Biology and Biological Significance*, J. P. Jost and H. P. Saluz (ed.), pp. 11-26, Birkhäuser Verlag, Basel.

Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989). *Molecular cloning: a laboratory manual*, 2nd ed. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

Santoro, R., D'Erme, M., Mastrantonio, S., Reale, A.,Merenzi, S., Saluz, H. -P., Strom, R. and Caiafa, P. (1995) *Biochem. J.* **305**, 739-744.

Sasaki, T., Hansen, R. S. and Gartler, S. M. (1992). *Mol. Cell. Biol.* **12**, 3819-3826.

Scheidt, G., Weber, H. Graessmann, M. and Graessmann, A. (1994). *Nucleic Acids Res.*, **22**, 953–958.

Selker, E. U., Fritz, D. Y. and Singer, M. J. (1993) *Science* **262**, 1724-1728.

Shapiro, H.S. (1975) Content of 6-methylaminopurine and 5-methylcytosine in DNA. CRC Handbook of Biochemistry and Molecular Biology. CRC Press. Boca Raton, FL.

Shemer, R., Walsh, A., Eisenberg, S., Breslow, J. and Razin, A. (1990). *J. Biol. Chem.* **265**, 1010-1015.

Shen, J. C., Rideout, W. M. and Jones, P. A. (1992). Cell. 71, 1073-1080.

Singer-Sam, J., Grant, M., LeBon, J. M., Okuyama, K., Chapman, V., Monk, M. and

Smith, S.S., Kaplan, B.E., Sowers, L.C. and Newman, E.M. (1992). *Proc. Natl. Acad. Sci. USA* **89**, 4744–4748.

Som, S., Bhagwat, A. S. and Fredman, S. (1987). *Nucleic Acids Res.* **15**, 313-323.

Staiger, D. H., Kaulen, H. and Schell, J. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 6930–6934.

Strand, M., Prolla, T.A., Liskay, R.M. and Petes, T.D. (1993) *Nature* **365**, 274–276.

Swain, J. L., Stewart, T. A. and Leder, P. (1987). *Cell.* **50**, 719-727.

Szilak, L., Venetianer, P. and Kiss, A. (1990) *Nucleic Acids Res.* **18**, 2287-2291.

Tate, P. H. and Bird, A. P. (1993). *Cur. Opin. Gen. Dev.* **3**, 226-231.

Tautz, D., Trick, M. and Dover, G.A. (1986) *Nature* **322**, 652–656.

Taylor, C., Ford, K., Connolly, B.A. and Hornby, D.P. (1993). *Biochem.J.* **291**, 493–504.

Theiss, G., Schleicher, R., Schimpff-Weiland, R. and Follmann, H (1987). *Eur. J. Biochem.* **167**, 89–96.

Tollefsbol, T. O. and Hutchison III, C. A. (1995) *J. Biol. Chem.*-Press

Trautner, T.A., Balganesh, T., Wilke, K., Noyer-Weidner, M., Rauhut, E., Lauster, R., Behrens, B and Pawlek, B. (1988). *Gene.* **74**, 267–270.

Trautner, T.A., Balganesh, T., Wilke, K., Noyer-Weidner, M., Rauhut, E., Lauster, R., Behrens, B and Pawlek, B. (1988). *Gene*. **74**, 267–270.

Tribioli, C., Tamanini, F., Patrosso, C., Milanesi, L., Villa, A., Pergolizzi, R., Maestrini, E., Rivella, S., Bione, S., Mancini, M., Vezzoni, P. and Toniolo, D. (1992). *Nucleic Acids Res*. **20**, 727-733.

Turnbull, J. F. and Adams, R. L. P. (1976) *Nucleic Acids Res*. **3**, 677-695.

Urieli-Shoval, S., Gruenbaum, Y., Sedat, J. and Razin, A. (1982) *FEBS Lett*. **146**, 148-152

Van der Krol, A. R. and Chua, N. -H. (1991) *Plant Cell* **3**, 667-675.

Vongs, A., Kakutani, T., Martienssen, R.A. and Richards, E.J. (1993). *Science*. **260**, 1926–1928.

Walter,.J., Noyer-Weidner, M. and Trautner, T. A. (1990) *EMBO J*. **9**, 1007-1013.

Weber, H., Ziechmann, C. and Graessmann, A. (1990). *EMBO. J*. **9**, 4409–4415.

Wike, K., Rauhut, E., Noyer-Weidener, M., Lauster, R., Pawlek, B., Behrens, B. and Trautner, T. A. (1988) *EMBO J*. **7**, 2601-2609.

Williams, M.E., Foster, R. and Chua, N-H. (1992) *Plant Cell* **4**, 485–496.

Wolf, S. F., Dintzis, S., Toniolo, D., Persico, G., Lunnen, K. D., Axelman, J. and Migeon, B. R. (1984). *Nucleic Acids Res*. **12**, 9333-9348.

Wu, J.C. and Santi, D V. (1987). *J. Biol. Chem*. **262**, 4778–4786.

Wyszynski, M. W., Gabbara, S. and Bhagwat, A. S. (1992). *Nucleic Acids Res*. **21**, 319-326

Wyszynski, M. W., Gabbara, S. Kubareva, E. A., Romanova, E. A., Oretskaya, T. S., Gromova, E. S., Shabarova, Z. A. and Bhagwat, A. S. (1993). *Nucleic Acids Res*. **21**, 295-301

Yen, C. R. -W., Vertino, P. M., Nelkin, B. D., Yu, J. J., El-Deiry, W., Cumaraswamy, A., Lennon, G. G., Trask, B. J., Celano, P. and Baylin, S. B. (1992) *Nucleic Acids Res.* **20**, 2287-2291.

Yesufu, H. M. I., Hanley, A., Rinaldi, A. and Adams, R.L.P. (1991). *Biochem. J.* **273**, 469–475.

Yuan, R. and Hamilton, D. L. (1984) Type I and type III restriction-modification enzymes in: DNA methylationBiochemistry and Biological significance, pp11-37. Eds. A. Razin, H. Cedar and A. D. Riggs. Springer-Verlag, New York.

Appendix-I

PTF sequences from pea nuclear genome

ptf6 Length: 113

```
  1  caggatagct tctgtcgctg tactatatgg tattctggaa ttactaaagt
 51  aggattcatt ggaagaattt cgattctttg aaagattgga aatgggctct
101  gataagcacc ctg
```

ptf17 Length: 139

```
  1  cagctgcgga agaagagatt gatgaatcta attccgataa ttcagattca
 51  gatgagtggg agttgagcaa tgagagcaag aagaggaaga agggaagaaa
101  tgctatggaa tttaggacac agcgaggttg caagtgctg
```

ptf21 Length: 147

```
  1  caggtgaaag tcttcagggg gacttttgtt ctgttcactt taaggtaaaa
 51  gagaaatgtg taatagagag agttgactaa cctcaggcgg gttgacccac
101  atctgtccat cataccatcc actttcttca taacgaagga atgactg
```

ptf31 Length: 92

```
  1  cagttgtctt tctgtttgga attgaataag gcaagctctt gtaaataaac
 51  cccaaacaat ggggaaggga taagagggaa gacactcatc tg
```

ptf102 Length: 98

```
  1  cagattggag accaagggtc tttccaggag gctcgggtga gaggggtgtg
 51  atgaatagag atttcccacc agcagcttct cctatcactt tacttctg
```

ptf104 Length: 98

```
  1  cagaagtaaa gtgataggag aagctgctgg tgggaaatct ctattcatca
 51  caccctctc acccgaggct cctggaaaga cccttggtct ccaatctg
```

ptf105 Length: 117

```
   1  cagcacagat gaacgacttc tcaccaacac tctgtgacct gcgaacgaca
  51  acaaactgag ccgaaattcc aaaacaaata aacgacatcg aaaccgaacc
 101  tccatcctcc aggcctg
```

ptf106 Length: 142

```
   1  caggacttga tcaagtggcc ttatgcattg aaggtttgga gcttgttatg
  51  agttatgttt atggcttgtt tagaatgcat cgttgaaatg ctaatggaga
 101  taattcttca ttgttgaatt gcagatgaaa ggttcgtgtc tg
```

ptf110 Length: 88

```
   1  cagagcaacc aaaccgccaa ctcaacctgc attccaacac tgcttcttcg
  51  cgcaatggcc atgaaatttc ataagcttga tccatctg
```

ptf111 Length: 257

```
   1  cagggtttag tatttggaag gaaggagctt cacttctcat tctgaaaatt
  51  atgcaggatt gtcttggtgt gctcaaatta tggtgttgct tggaatcatg
 101  gactgtgttg aataaaattg ggaaatgggt tgaagcttag atataggact
 151  gatgtaccaa tgcatttatg aagcttcact aaggttggtt ggctatgaca
 201  ttccggtgaa atgcacgatt cttaaaggcg taccatgagg gtgatgatgg
 251  ttggctg
```

ptf113 Length: 101

```
   1  cagtgtgtta atatattacc tgttgaaaac cagttggttg ttaatggtct
  51  agctacaagc tgaatgttaa tgaatggata aaatcccgtt aaacgaccct
 101  g
```

ptf115 Length: 160

```
   1  caggggaaaa gattcaggag ttcatagtgt aacacctgct gttggaatta
  51  ctcaagaagc ttttttttt cgtccgatga agctgttcag actgctagta
 101  gtttggccat atcttcgaga ttggttcagg tccctttcc aaagtaggca
 151  ttgttagctg
```

ptf118 Length: 142

```
   1  cagatgagtg tcttcccttct tatcccttcc ccattgtttg ggtttaattt
  51  caagagcttg ccttagtatg aatcccatct acggctgagt caatagtacc
 101  ctccaacttt aggtgattca attgcaaaca gaaagacaac tg
```

ptf119 Length: 141 ..

```
   1  cagttgatct ttctgtttgc aattgaatca ctaaagttgg agggtgctat
  51  ggactcagcc gtagatggga ttcatactaa ggcaagctct tgaaattaaa
 101  cccaacaatg gggaagggat aagagggaag acactcatct g
```

ptf121 Length: 142

```
   1  cagttgagct ctctgttcgc aatagaatca ctaaagttgg agggtgctat
  51  ggactcagcc gctgatggtt tacatactaa ggcaatctct ggaacttata
 101  cccaaacgct ggtgaagggg taatacggaa gtaactcacc tg
```

ptf125 Length: 140

```
   1  cagattcaat gctttatggt tgcaattgat gcattgatag gttcatatgt
  51  gttgcattat gaatttgttg gacttgcagg atttgggctt tatagttgcc
 101  tcattgctta ctgtctacaa aaggttggtc aggagtgctg
```

ptf126 Length: 180

```
   1  cagtgaatgg tatagacctt ggcacgctat actggactga taagttagta
  51  acgacggagt cagatgatcc acctgacatc agatgctcca aatcaagctg
 101  taatcttatg caacgtaatc agtattctcc aaatcaaact gcaatcttat
 151  gcaagggggc agattctcca aatcaggctg
```

ptf127 Length: 98

```
   1  cagattggag accaagggtc tttccaggag gctcgggtga gaggggtgtg
  51  atgaatagag atttcccacc agcagcttct cctatcactt tacttctg
```

ptf129 Length: 88

1 cagacggcag gcagatcgtt ggcagggcta aatctgaagc aaataactat

51 gacaggtttg ctctatttgc ttgcaacctt tatttctg

ptf130 Length: 98

1 cagattggag accaagggtc tttccaggag gctcgggtga gaggggtgtg

51 atgaatagag atttcccacc agcagcttct cctatcactt tacttctg

ptf131 Length: 175

1 caggaccaac gtacacagga taatgtcagc attataattg ccgatctagg

51 gtatgtattt tgtcaatgta attatgattt tggacttcaa ttaggttcta

101 ttactggcta tcgctatcct gttgtgtctt atattgagtg cttgcaggaa

151 gacagattgg aagaatgtgc cactg

ptf133 Length: 153

1 cagaggctag caagtttaca gtaacacagc tcatccaggg caatgtttgg

51 ttgaagaaac acaggggtag ccttcattga aggcctgtag aaatcggctt

101 cagcaggcgt gtactattat gtttctgtaa aaaaaaggta tcatttgaaa

151 ctg

ptf134 Length: 166

1 cagcaaggtg gtacagtgag ccagccctca ctcatttatt gtggctcttc

51 aaaatcattt taccactctt catatatcta ccaatagtca ctacaggagc

101 aatgtcctcg ccggaagtgg tgaaaccgct tatggtctct gagaattaca

151 tcccggttgt aaactg

ptf136 Length: 177

1 cagcatgact gagaaacttt ctgcttccat ggctgatgcg gttagcaggg

51 ctcattggaa ttattagaag gattcggcag gtaatgtcgt cctgtaatgg

101 aggttgactt catcatggga acttgaaggg gtctgttatt gaattcaaag

151 tgggcgagag aaggtgttag aatactg

ptf137 Length: 206

```
   1  cagtgctttg ttgttgttat taatgcttag ttaagtgtat tgttttgatg
  51  tacagattat ctgctaaatt atattattgt tgaagaactt cctgcactgg
 101  ttaaggaaga atatttgaca aatacactgc actgaaagcg gcatcatgta
 151  gtaggtggga aatggcattt tacacaaaaa tttggatagc aaatgcaaaa
 201  actctg
```

ptf138 Length: 148

```
   1  cagattcatt ctcacgacaa gatctaatgg aatggggaac tcaggttgga
  51  gtatgggttt ttctcttgga taattgctca ggtaagctga tttaaaacga
 101  caaattattc agggcacttt tgtaaatatg catgaagaaa aacttctg
```

ptf139 Length: 141

```
   1  cagatgagtg tcttcctct tatcccttcc ccattgtttg ggtttaattt
  51  caagagcttg ccttagtatg aatcccatct acggctgagt caatagcacc
 101  ctccaacttt agtgattcaa ttgcaaacag aaagacaact g
```

ptf140 Length: 106

```
   1  cagaaggaac tctagatttt ttttggagtt cctgcaaaat ggatcttacc
  51  ctttgtttta ctggaaggtg ctttgtcatt tatcaatagc ttggtccatg
 101  aggctg
```

ptf141 Length: 135

```
   1  cagagaattg ggtagaccat cgtccggaaa cgacaagaga gatggaattg
  51  aatggagttc caagtgttga agacaagtgc tcataaacat tttaggtgga
 101  gacaacatcc taaatttctg tatttttaca tgctg
```

ptf145 Length: 98

```
   1  cagaagtaaa gtgataggag aagctgctgg tgggaaatct ctattcatca
  51  cacccctctc acccgagtct cctggaaaga cccttggtct ccaatctg
```

ptf147 Length: 102

```
   1  cagtactatg gagaccaagg gtctccccag gaggctcggt tgaaaggagt
  51  gtgctgattc aagatttccg acactgaagc tactcccatc aatatatttc
 101  tg
```

ptf150 Length: 80

```
   1  cagtctctga ttcctgtggc ggctcttcag cttccgctat ggcagtttcc
  51  agcttcagca actgtctcga gcagcaactg
```

ptf151 Length: 166

```
   1  cagaatccgt gggggggggat tgcttttttcc gttctctgag tctcgagctc
  51  attacgagtc cgcggcagga gatacagaag tatcaggatt atggtgtatt
 101  cgatcacagt gtgtctcgtc agggggggctc acgtggagca caatccatag
 151  agctcggaag gctctg
```

ptf152 Length: 299

```
   1  cagataagaa taaggataac cttgtaaagg agtctcctta gcttgcaaag
  51  caggcctggt aaggagtgaa agagagaaga ttgagtagcc gggtgaaata
 101  tccttagcta ttgaatttta agaagctgtg gctctttcaa taagatcttg
 151  gactgggata ctaagcaaga gagccaaagg ggcgaatcga gatatagatg
 201  actggatcaa tctggataca ataccttgga aaaggggata agccctggga
 251  cttcttttttt tagttagagc ttctcacttc attgggtcag tcatagctg
```

ptf153 Length: 114

```
   1  caggttaaac cactatctcg gcttaggggg gtcctagggt tgccaccagg
  51  agggtatctg cctggacgac caggatggac ctcgaggctt accataatga
 101  gggatagggt cctg
```

ptf154 Length: 307

```
   1  cagtcaagac tctttcaaga gcttcttcaa gggccttctt tccggttata
  51  gtttcttttg catgcctcgt gtaccctttt caaaatatcc atacgcgatg
```

```
   101  aatatgcttc atcaacttca agcatagcat atgctactct cttcttagga
   151  ctgtccttct ggggctcgag cttcacgggt cagagtaaca tactcatcaa
   201  ttggaaagcg taatttcctt gcgctttggc gtacactctc ttggtgaaac
   251  ccctgtgggt tagagttaga caaaagccgt tatcttcgac aagggtagca
   301  gattctg

ptf155  Length: 66
     1  cagatactat tgccaggtgt ctatacgccc ttgtaatgtg cgtggcatcc
    51  catgtgcacg ccactg

ptf156  Length: 98
     1  cagggtcgaa aacaaagatt cttgatataa tgtcgcgaca ttcaggagat
    51  atttgaacac agtctggtac ggaatactgg acgctgagca accctctg

ptf157  Length: 98
     1  cagattagag accaagggtc tttccaggag gctcgggtga gaggggtatg
    51  atgaataggg atatcccacc aggaggttct cctatcactt taattctg

ptf158  Length: 144
     1  cagaaaaaac caaaaaggag tggggaaaaa aaatgaaata cacattttca
    51  ttatatattg ctcacacttg tcttgttttc tttgcatcta actcactctt
   101  gttgtgtaat cattttgttt tgcatttttg tgtgtttctt tctg

ptf159  Length: 141
     1  cagttgtctt tctgtttgga attgaatcac taaagttgga gggtgatatt
    51  gactcagccg tagatgggat tcatactaag gtaagctctt gaaattaaac
   101  ccaaacaatg gggaagggat aagagggaag acactcatct g

ptf160  Length: 164
     1  ctgagtacat gagcgaattc tcaaacctgg caaagcaggg atgttgtaag
    51  tccaccgaca atggcattca caatcaaggt catatcaggt acaatcaacg
   101  caagaagtgc cattgggact atattcccat cagtccattg tagctgtaac
```

151 aaatcaggtt gcag

ptf165 Length: 170

1 cagtttacaa ccggggatgt aattctcaga gagcataagc ggtttcacca

51 cttcccgcgg agggacattg ctcctgtagt gactattggt aggatatatt

101 aagagtggta aaatgattgt ggagagccac aataaattag ttagggctgg

151 ttcactatac caccctgctg

ptf166 Length: 161

1 gagtgtctgt gtggaatatt cgcagctagc ggatatctcg cacgattatt

51 gtgattgcat agatggaggt ctatttggaa ttataggtag cttgcaagta

101 tccaacagct aatggttgtg ggtgtttgtg atgaagatac tcgatcgcca

151 gcggatagct g

ptf167 Length: 141

1 cagttgtctt tctgtttgca attgaatcac taaagttgga gggtggtatt

51 gactcaggcg tagatgggat tcatactaag gaaagctctt gaaattaaac

101 ccaaacaatg gggaggggat aagagggaag acactcatct g

ptf169 Length: 233

1 cagttggatg aatgactact gtggcatcaa agtccgcctt tgttaatcca

51 gctttaatac caactgtaaa cctctgtata acaagcatct ttacagttac

101 gttaattaga ccacatggat cactatcaga gtcaatctaa ggagaatatc

151 ttaggaaata gcatcatcct gtgcaatttc agaaggatct ttctccacac

201 atgttggaat ccaagcacac aattcgtgtt ctg

ptf170 Length: 144

1 cagtcattcc ttatgaagga aagtggatgg tatgatggac agatgtgggt

51 caagacgcct gaggttagtc aactctctct attacacatt tctcttttac

101 cttaaagtga acagaacaaa agtccccctg taagacttca cctg

ptf173 Length: 196

```
     1  caggaattcc caaaccgatg ggccagctta catttggctg tatgtagaca
    51  ataattgtta aggatatcaa ctgagtcaag agtgaaagtg aaatagtgac
   101  caagtgaagg aagctattgt tcccctcttt ccccgaatcc gtgttgagga
   151  ttaaattgaa cagatccaaa tgctaagtta cattgtcgga tccctg
```

ptf174 Length: 126

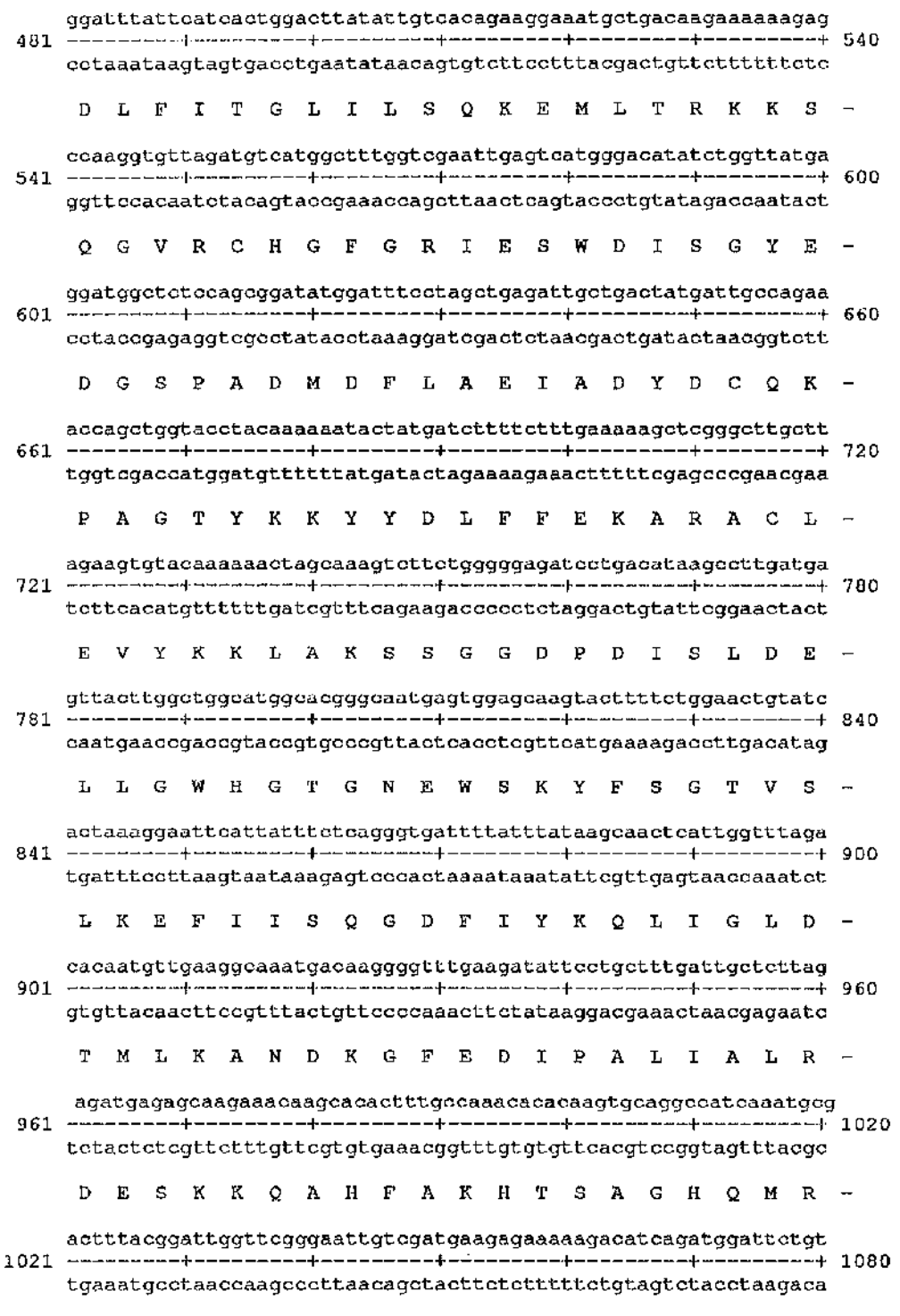
```
     1  cagacccaaa tcgacacatc tcgttggata aggtcctcag gtaaagtagc
    51  actattgcat cccaagtcat taagtaggcc aactgctgtg ctctttaatc
   101  gagatatatc ctcgtccatt ggactg
```

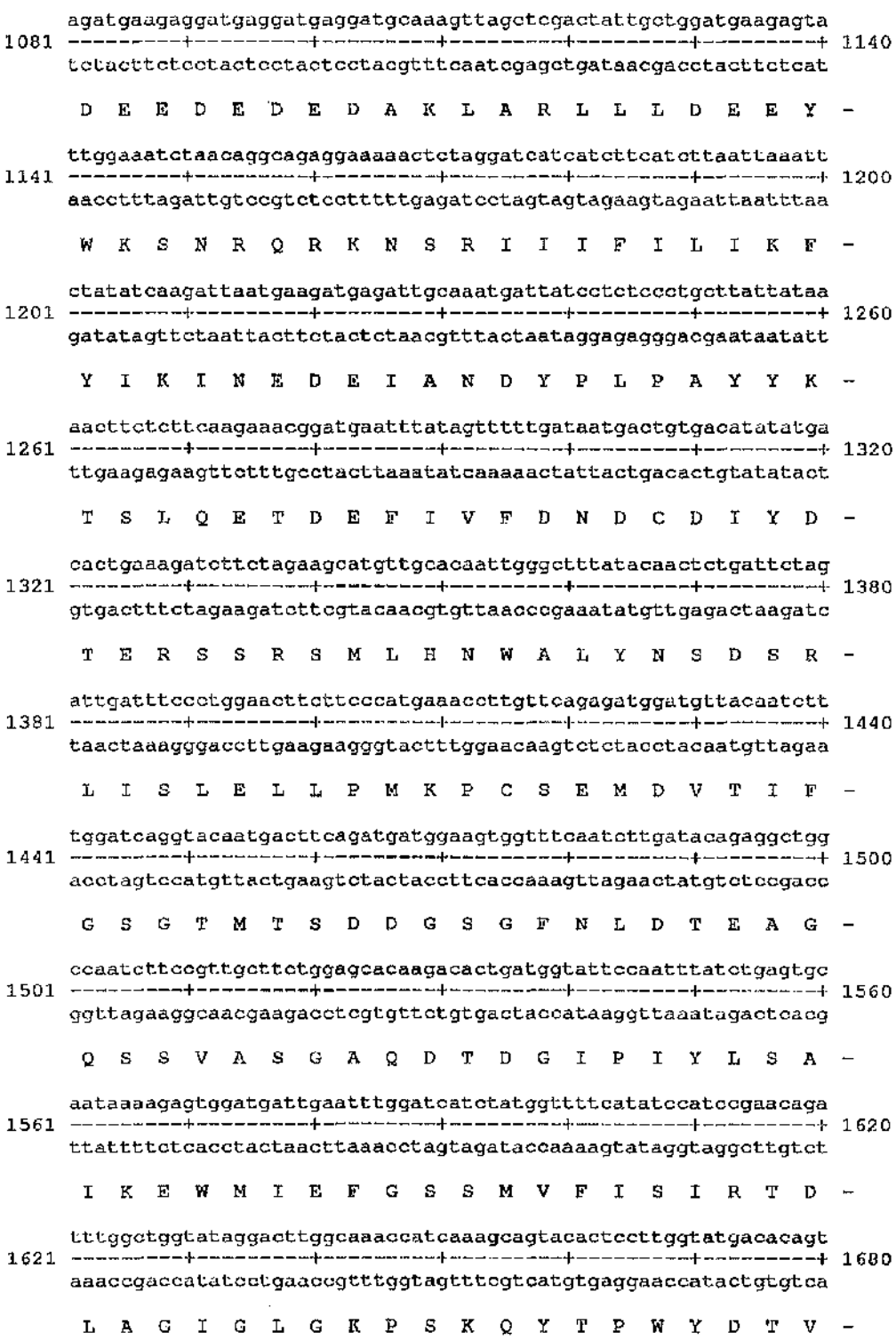
Appendix II

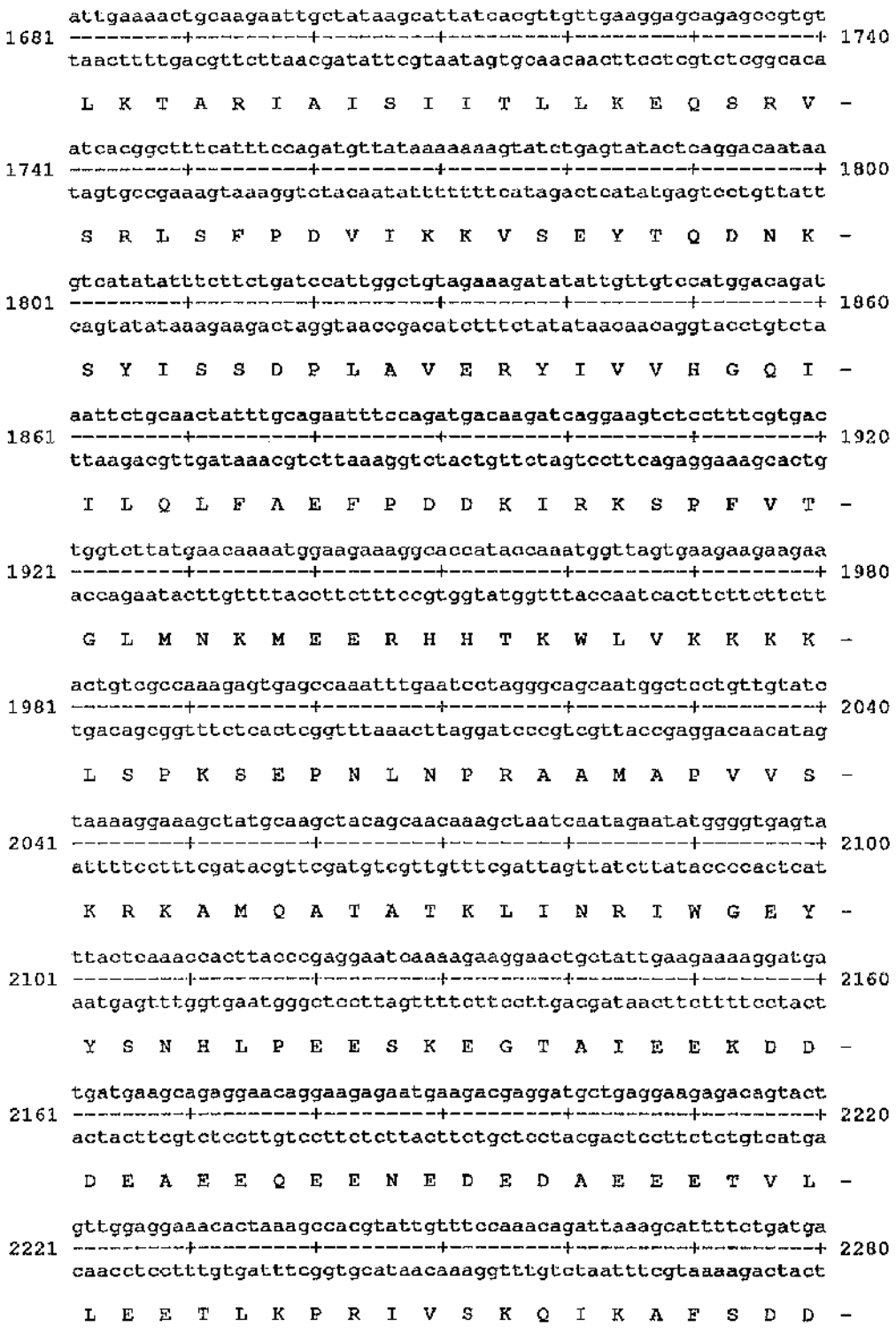
Assembled sequence for pea DNA methyltransferase

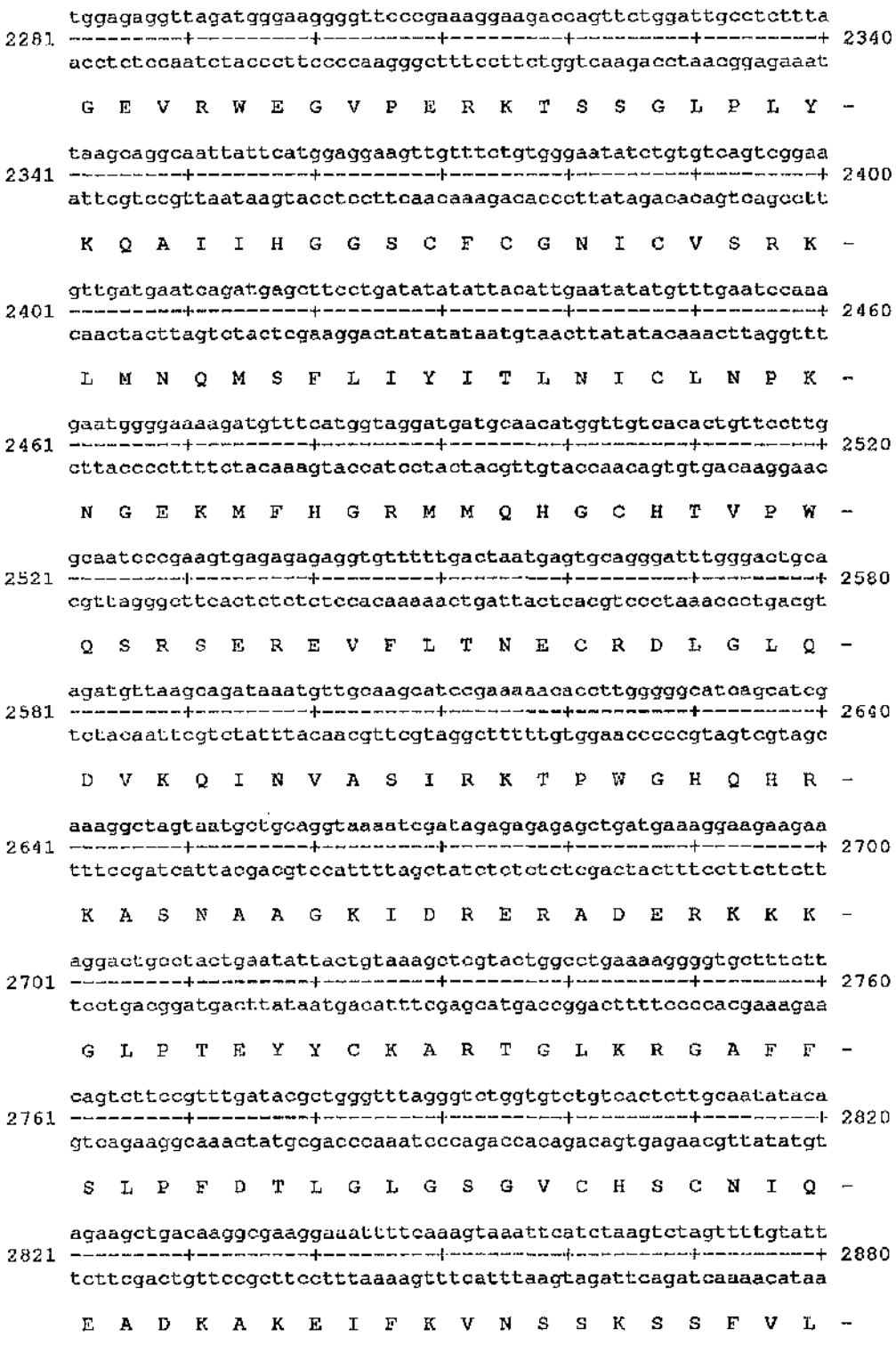
pmte4.map
(Linear) MAP of: pmte4 check: 2774 from: 1 to: 4987

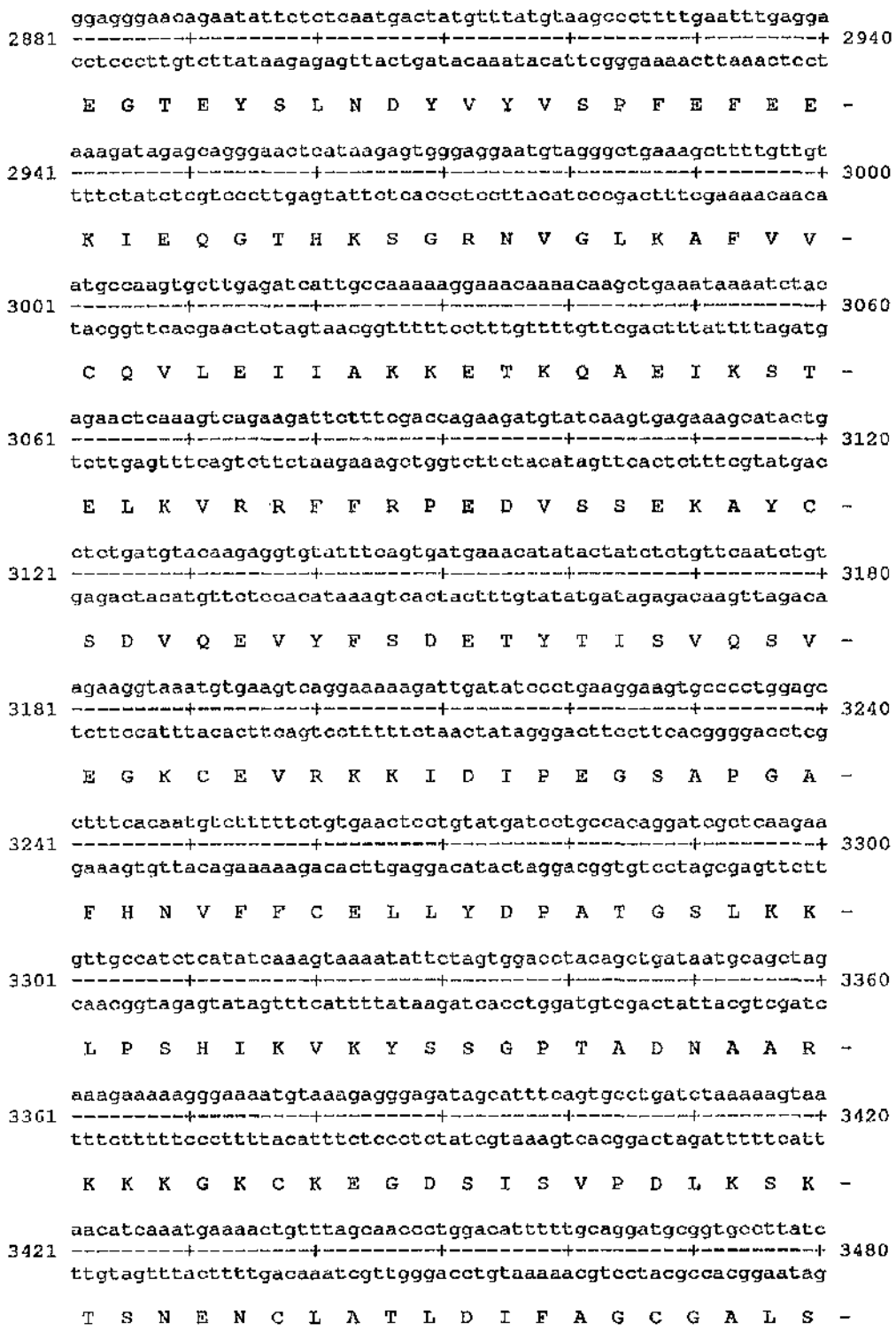
```
   caaaacgtttctcactcccatttcttcagatctacaacccgcgttttggatacaaggtaa
 1 ---------+---------+---------+---------+---------+---------+ 60
   gttttgcaaagagtgagggtaaagaagtctagatgttgggcgcaaaacctatgttccatt

   aattttccaactcatgggttccgcttcgcttttgaatccctccgattcgtctctaccggg
61 ---------+---------+---------+---------+---------+---------+ 120
   ttaaaaggttgagtacccaaggcgaagcgaaaacttagggaggctaagcagagatggccc

                M  G  S  A  S  L  L  N  P  S  D  S  S  L  P  G  -

   tggcaaggacagcacgagtaaagaagagcctgtttcaaacactgaaggggaagttatggc
121 ---------+---------+---------+---------+---------+---------+ 180
   accgttcctgtcgtgctcatttcttctcggacaaagtttgtgacttcccccttcaataccg

    G  K  D  S  T  S  K  E  E  P  V  S  N  T  E  G  E  V  M  A  -

   tggtggtaagcaaaagaagcgaagtttgtcagagagcagtgagcagcctgctcctactcg
181 ---------+---------+---------+---------+---------+---------+ 240
   accaccattcgttttcttcgcttcaaacagtctctcgtcactcgtcggacgaggatgagc

    G  G  K  Q  K  K  R  S  L  S  E  S  S  E  Q  P  A  P  T  R  -

   gaaagtggcgaaacgatctgcaagtgcagcaagtaaaaatttggaggaggaagtcttttt
241 ---------+---------+---------+---------+---------+---------+ 300
   ctttcaccgctttgctagacgttcacgtcgttcatttttaaacctcctccttcagaaaaa

    K  V  A  K  R  S  A  S  A  A  S  K  N  L  E  E  E  V  F  F  -

   ccatatcttgataagtcttgtcttgttgaaactaagaaggatcaggttgcagaaggagaa
301 ---------+---------+---------+---------+---------+---------+ 360
   ggtatagaactattcagaacagaacaactttgattcttcctagtccaacgtcttcctctt

    H  I  L  I  S  L  V  L  L  K  L  R  R  I  R  L  Q  K  E  N  -

   ttgctcgcagtccgctatgactgctggacaagaggatgaccgcccaaatagaagacttac
361 ---------+---------+---------+---------+---------+---------+ 420
   aacgagcgtcaggcgatactgacgacctgttctcctactggcgggtttatcttctgaatg

    C  S  Q  S  A  M  T  A  G  Q  E  D  D  R  P  N  R  R  L  T  -

   agactttatccttcatgatgaaagtggtgcagcacaggcacttgagatgcttgaaatcaa
421 ---------+---------+---------+---------+---------+---------+ 480
   tctgaaataggaagtactactttcaccacgtcgtgtccgtgaactctacgaactttagtt

    D  F  I  L  H  D  E  S  G  A  A  Q  A  L  E  M  L  E  I  K  -
```

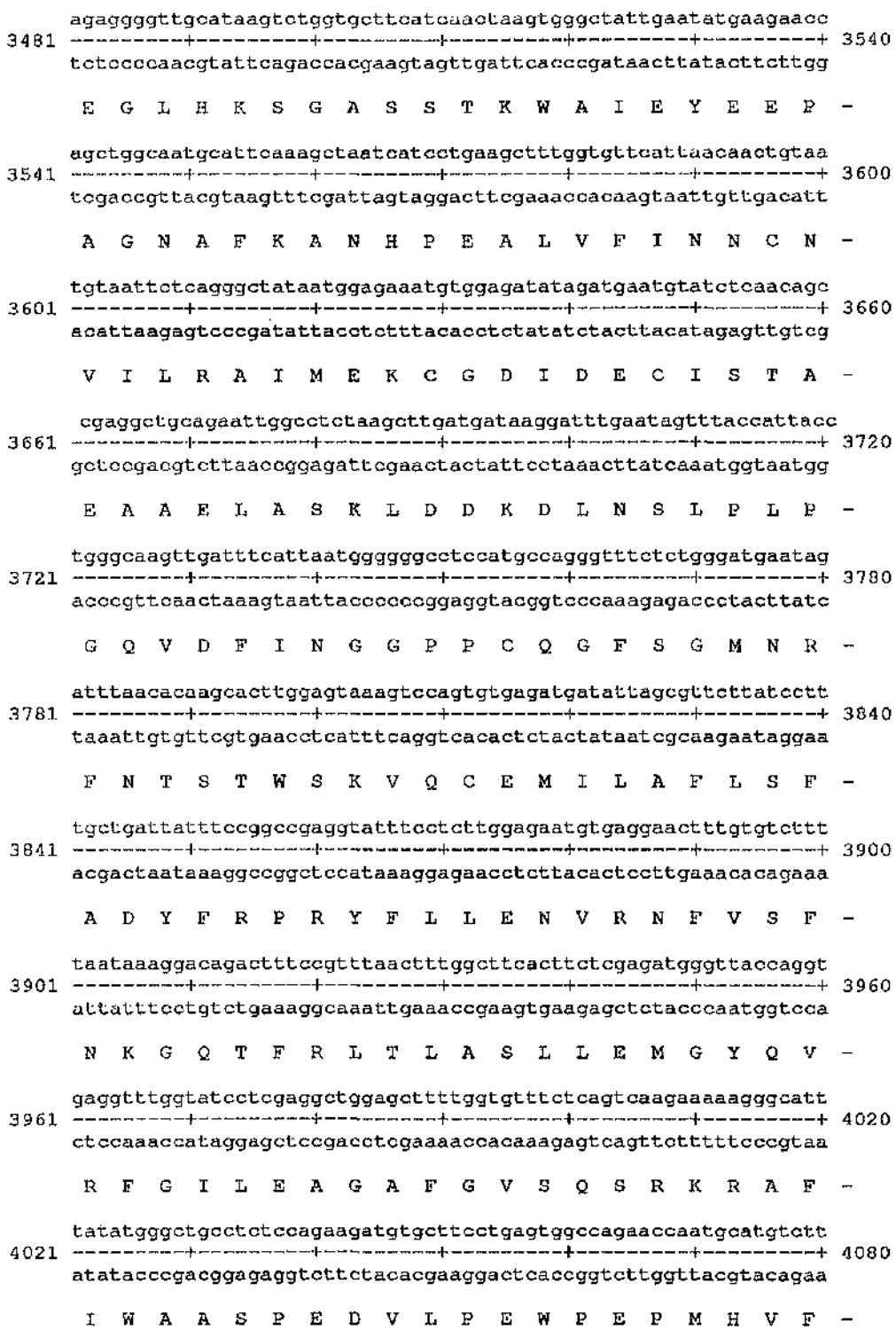

ggatttattcatcactggacttatattgtcacagaaggaaatgctgacaagaaaaaagag
481
540
cctaaataagtagtgacctgaatataacagtgtcttcctttacgactgttcttttttctc
D L F I T G L I L S Q K E M L T R K K S -
ccaaggtgttagatgtcatggctttggtcgaattgagtcatgggacatatctggttatga
541
600
ggttccacaatctacagtaccgaaaccagcttaactcagtaccctgtatagaccaatact
Q G V R C H G F G R I E S W D I S G Y E -
ggatggctctccagcggatatggatttcctagctgagattgctgactatgattgccagaa
601
660
cctaccgagaggtcgcctatacctaaaggatcgactctaacgactgatactaacggtctt
D G S P A D M D F L A E I A D Y D C Q K -
accagctggtacctacaaaaaatactatgatcttttctttgaaaaagctcgggcttgctt
661
720
tggtcgaccatggatgttttttatgatactagaaaagaaacttttcgagcccgaacgaa
P A G T Y K K Y Y D L F F E K A R A C L -
agaagtgtacaaaaaactagcaaagtcttctgggggagatcctgacataagccttgatga
721
780
tcttcacatgttttttgatcgtttcagaagacccccctctaggactgtattcggaactact
E V Y K K L A K S S G G D P D I S L D E -
gttacttggctggcatggcacgggcaatgagtggagcaagtacttttctggaactgtatc
781
840
caatgaaccgaccgtaccgtgcccgttactcacctcgttcatgaaaagaccttgacatag
L L G W H G T G N E W S K Y F S G T V S -
actaaaggaattcattatttctcagggtgattttatttataagcaactcattggtttaga
841
900
tgatttccttaagtaataaagagtcccactaaaataaatattcgttgagtaaccaaatct
L K E F I I S Q G D F I Y K Q L I G L D -
cacaatgttgaaggcaaatgacaaggggtttgaagatattcctgctttgattgctcttag
901
960
gtgttacaacttccgtttactgttccccaaacttctataaggacgaaactaacgagaatc
T M L K A N D K G F E D I P A L I A L R -
agatgagagcaagaaacaagcacactttgccaaacacacaagtgcaggccatcaaatgcg
961
1020
tctactctcgttctttgttcgtgtgaaacggtttgtgtgttcacgtccggtagtttacgc
D E S K K Q A H F A K H T S A G H Q M R -
actttacggattggttcgggaattgtcgatgaagagaaaaagacatcagatggattctgt
1021
1080
tgaaatgcctaaccaagcccttaacagctacttctcttttctgtagtctacctaagaca
L Y G L V R E L S M K R K R H Q M D S V -

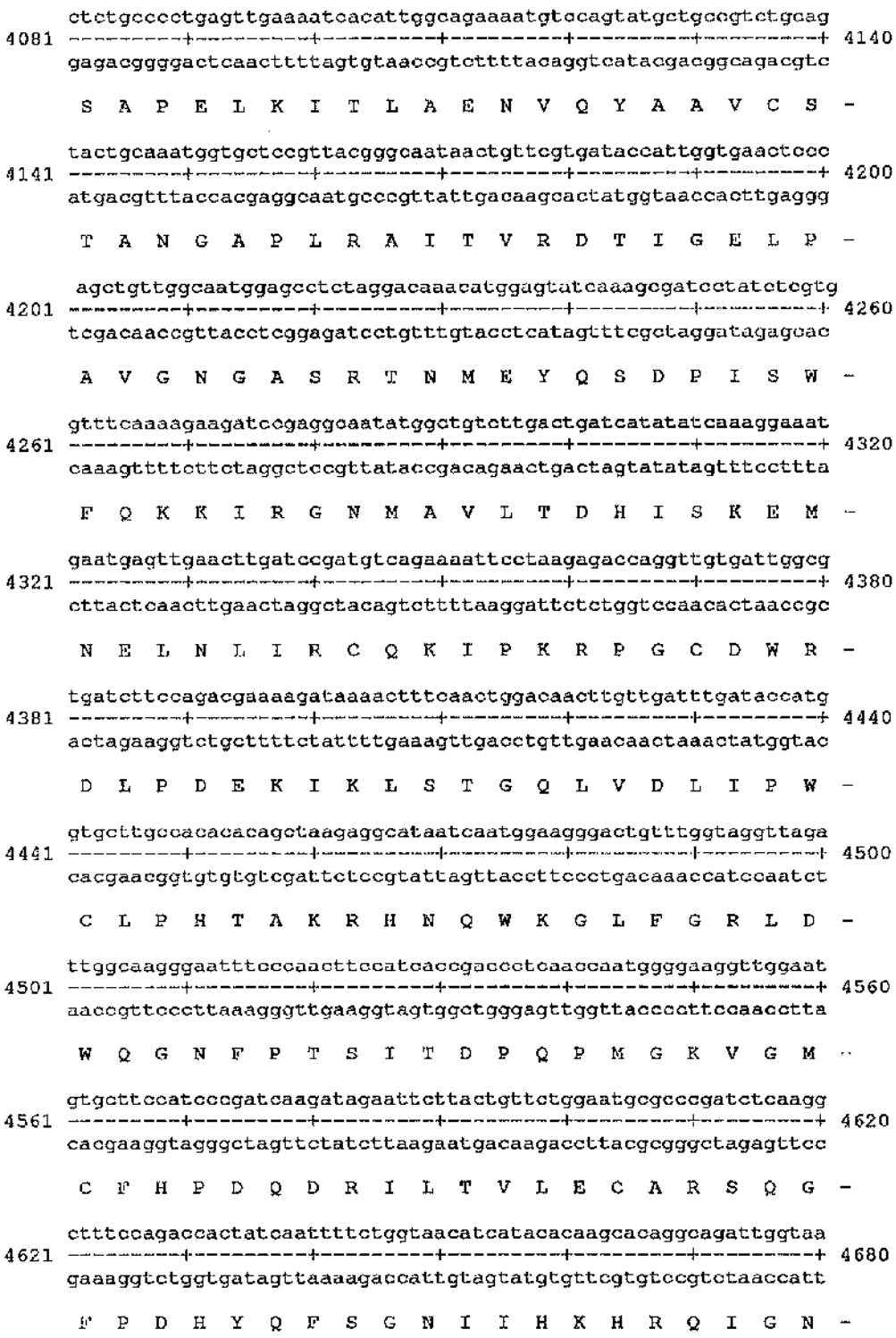

agatgaagaggatgaggatgaggatgcaaagttagctcgactattgctggatgaagagta
1081 ----------+---------+---------+---------+---------+---------+ 1140
tctacttctcctactcctactcctacgtttcaatcgagctgataacgacctacttctcat
D E E D E D E D A K L A R L L L D E E Y -
ttggaaatctaacaggcagaggaaaaactctaggatcatcatcttcatcttaattaaatt
1141 ----------+---------+---------+---------+---------+---------+ 1200
aacctttagattgtccgtctcctttttgagatcctagtagtagaagtagaattaatttaa
W K S N R Q R K N S R I I I F I L I K F -
ctatatcaagattaatgaagatgagattgcaaatgattatcctctccctgcttattataa
1201 ----------+---------+---------+---------+---------+---------+ 1260
gatatagttctaattacttctactctaacgtttactaataggagagggacgaataatatt
Y I K I N E D E I A N D Y P L P A Y Y K -
aacttctcttcaagaaacggatgaatttatagtttttgataatgactgtgacatatatga
1261 ----------+---------+---------+---------+---------+---------+ 1320
ttgaagagaagttctttgcctacttaaatatcaaaaactattactgacactgtatatact
T S L Q E T D E F I V F D N D C D I Y D -
cactgaaagatcttctagaagcatgttgcacaattgggctttatacaactctgattctag
1321 ----------+---------+---------+---------+---------+---------+ 1380
gtgactttctagaagatcttcgtacaacgtgttaacccgaaatatgttgagactaagatc
T E R S S R S M L H N W A L Y N S D S R -
attgatttccctggaacttcttcccatgaaaccttgttcagagatggatgttacaatctt
1381 ----------+---------+---------+---------+---------+---------+ 1440
taactaaagggaccttgaagaagggtactttggaacaagtctctacctacaatgttagaa
L I S L E L L P M K P C S E M D V T I F -
tggatcaggtacaatgacttcagatgatggaagtggtttcaatcttgatacagaggctgg
1441 ----------+---------+---------+---------+---------+---------+ 1500
acctagtccatgttactgaagtctactaccttcaccaaagttagaactatgtctccgacc
G S G T M T S D D G S G F N L D T E A G -
ccaatcttccgttgcttctggagcacaagacactgatggtattccaatttatctgagtgc
1501 ----------+---------+---------+---------+---------+---------+ 1560
ggttagaaggcaacgaagacctcgtgttctgtgactaccataaggttaaatagactcacg
Q S S V A S G A Q D T D G I P I Y L S A -
aataaaagagtggatgattgaatttggatcatctatggttttcatatccatccgaacaga
1561 ----------+---------+---------+---------+---------+---------+ 1620
ttatttttctcacctactaacttaaacctagtagataccaaaagtataggtaggcttgtct
I K E W M I E F G S S M V F I S I R T D -
tttggctggtataggacttggcaaaccatcaaagcagtacactccttggtatgacacagt
1621 ----------+---------+---------+---------+---------+---------+ 1680
aaaccgaccatatcctgaaccgtttggtagtttcgtcatgtgaggaaccatactgtgtca
L A G I G L G K P S K Q Y T P W Y D T V -


attgaaaactgcaagaattgctataagcattatcacgttgttgaaggagcagagccgtgt
1681 ---------+---------+---------+---------+---------+---------+ 1740
taacttttgacgttcttaacgatattcgtaatagtgcaacaacttcctcgtctcggcaca
L K T A R I A I S I I T L L K E Q S R V -
atcacggctttcatttccagatgttataaaaaaagtatctgagtatactcaggacaataa
1741 ---------+---------+---------+---------+---------+---------+ 1800
tagtgccgaaagtaaaggtctacaatatttttttcatagactcatatgagtcctgttatt
S R L S F P D V I K K V S E Y T Q D N K -
gtcatatatttcttctgatccattggctgtagaaagatatattgttgtccatggacagat
1801 ---------+---------+---------+---------+---------+---------+ 1860
cagtatataaagaagactaggtaaccgacatctttctatataacaacaggtacctgtcta
S Y I S S D P L A V E R Y I V V H G Q I -
aattctgcaactatttgcagaatttccagatgacaagatcaggaagtctcctttcgtgac
1861 ---------+---------+---------+---------+---------+---------+ 1920
ttaagacgttgataaacgtcttaaaggtctactgttctagtccttcagaggaaagcactg
I L Q L F A E F P D D K I R K S P F V T -
tggtcttatgaacaaaatggaagaaaggcaccataccaaatggttagtgaagaagaagaa
1921 ---------+---------+---------+---------+---------+---------+ 1980
accagaatacttgttttaccttctttccgtggtatggtttaccaatcacttcttcttctt
G L M N K M E E R H H T K W L V K K K K -
actgtcgccaaagagtgagccaaatttgaatcctagggcagcaatggctcctgttgtatc
1981 ---------+---------+---------+---------+---------+---------+ 2040
tgacagcggtttctcactcggtttaaacttaggatcccgtcgttaccgaggacaacatag
L S P K S E P N L N P R A A M A P V V S -
taaaaggaaagctatgcaagctacagcaacaaagctaatcaatagaatatggggtgagta
2041 ---------+---------+---------+---------+---------+---------+ 2100
attttcctttcgatacgttcgatgtcgttgtttcgattagttatcttataccccactcat
K R K A M Q A T A T K L I N R I W G E Y -
ttactcaaaccacttacccgaggaatcaaaagaaggaactgctattgaagaaaaggatga
2101 ---------+---------+---------+---------+---------+---------+ 2160
aatgagtttggtgaatgggctccttagttttcttccttgacgataacttcttttcctact
Y S N H L P E E S K E G T A I E E K D D -
tgatgaagcagaggaacaggaagagaatgaagacgaggatgctgaggaagagacagtact
2161 ---------+---------+---------+---------+---------+---------+ 2220
actacttcgtctccttgtccttctcttacttctgctcctacgactccttctctgtcatga
D E A E E Q E E N E D E D A E E E T V L -
gttggaggaaacactaaagccacgtattgtttccaaacagattaaagcattttctgatga
2221 ---------+---------+---------+---------+---------+---------+ 2280
caacctcctttgtgatttcggtgcataacaaaggtttgtctaatttcgtaaaagactact
L E E T L K P R I V S K Q I K A F S D D -


tggagaggttagatgggaaggggttcccgaaaggaagaccagttctggattgcctcttta
2281 2340
acctctccaatctaccctccccaagggctttccttctggtcaagacctaacggagaaat
G E V R W E G V P E R K T S S G L P L Y -
taagcaggcaattattcatggaggaagttgtttctgtgggaatatctgtgtcagtcggaa
2341 2400
attcgtccgttaataagtacctccttcaacaaagacacccttatagacacagtcagcctt
K Q A I I H G G S C F C G N I C V S R K -
gttgatgaatcagatgagcttcctgatatatattacattgaatatatgtttgaatccaaa
2401 2460
caactacttagtctactcgaaggactatatataatgtaacttatatacaaacttaggttt
L M N Q M S F L I Y I T L N I C L N P K -
gaatggggaaaagatgtttcatggtaggatgatgcaacatggttgtcacactgttccttg
2461 2520
cttacccctttttctacaaagtaccatcctactacgttgtaccaacagtgtgacaaggaac
N G E K M F H G R M M Q H G C H T V P W -
gcaatcccgaagtgagagagaggtgtttttgactaatgagtgcagggatttgggactgca
2521 2580
cgttagggcttcactctctctccacaaaaactgattactcacgtccctaaaccctgacgt
Q S R S E R E V F L T N E C R D L G L Q -
agatgttaagcagataaatgttgcaagcatccgaaaaacaccttgggggcatcagcatcg
2581 2640
tctacaattcgtctatttacaacgttcgtaggcttttgtggaacccccgtagtcgtagc
D V K Q I N V A S I R K T P W G H Q H R -
aaaggctagtaatgctgcaggtaaaatcgatagagagagagctgatgaaaggaagaagaa
2641 2700
tttccgatcattacgacgtccattttagctatctctctctcgactactttccttcttctt
K A S N A A G K I D R E R A D E R K K K -
aggactgcctactgaatattactgtaaagctcgtactggcctgaaaaggggtgctttctt
2701 2760
tcctgacggatgacttataatgacatttcgagcatgaccggacttttccccacgaaagaa
G L P T E Y Y C K A R T G L K R G A F F -
cagtcttccgtttgatacgctgggtttagggtctggtgtctgtcactcttgcaatataca
2761 2820
gtcagaaggcaaactatgcgacccaaatcccagaccacagacagtgagaacgttatatgt
S L P F D T L G L G S G V C H S C N I Q -
agaagctgacaaggcgaaggaaattttcaaagtaaattcatctaagtctagttttgtatt
2821 2880
tcttcgactgttccgcttcctttaaaagtttcatttaagtagattcagatcaaaacataa
E A D K A K E I F K V N S S K S S F V L -


ggagggaacagaatattctctcaatgactatgtttatgtaagccctttgaatttgagga
2881 2940
cctcccttgtcttataagagagttactgatacaaatacattcgggaaaacttaaactcct
E G T E Y S L N D Y V Y V S P F E F E E -
aaagatagagcagggaactcataagagtgggaggaatgtagggctgaaagcttttgttgt
2941 3000
tttctatctcgtcccttgagtattctcaccctccttacatcccgactttcgaaaacaaca
K I E Q G T H K S G R N V G L K A F V V -
atgccaagtgcttgagatcattgccaaaaaggaaacaaaacaagctgaaataaaatctac
3001 3060
tacggttcacgaactctagtaacggtttttcctttgttttgttcgactttatttagatg
C Q V L E I I A K K E T K Q A E I K S T -
agaactcaaagtcagaagattctttcgaccagaagatgtatcaagtgagaaagcatactg
3061 3120
tcttgagtttcagtctttctaagaaagctggtcttctacatagttcactctttcgtatgac
E L K V R R F F R P E D V S S E K A Y C -
ctctgatgtacaagaggtgtatttcagtgatgaaacatatactatctctgttcaatctgt
3121 3180
gagactacatgttctccacataaagtcactactttgtatatgatagagacaagttagaca
S D V Q E V Y F S D E T Y T I S V Q S V -
agaaggtaaatgtgaagtcaggaaaaagattgatatccctgaaggaagtgcccctggagc
3181 3240
tcttccatttacacttcagtcctttttctaactatagggacttccttcacggggacctcg
E G K C E V R K K I D I P E G S A P G A -
ctttcacaatgtcttttttctgtgaactcctgtatgatcctgccacaggatcgctcaagaa
3241 3300
gaaagtgttacagaaaaagacacttgaggacatactaggacggtgtcctagcgagttctt
F H N V F F C E L L Y D P A T G S L K K -
gttgccatctcatatcaaagtaaaatattctagtggacctacagctgataatgcagctag
3301 3360
caacggtagagtatagtttcattttataagatcacctggatgtcgactattacgtcgatc
L P S H I K V K Y S S G P T A D N A A R -
aaagaaaaagggaaaatgtaaagagggagatagcatttcagtgcctgatctaaaaagtaa
3361 3420
tttcttttccctttttacatttctccctctatcgtaaagtcacggactagatttttcatt
K K K G K C K E G D S I S V P D L K S K -
aacatcaaatgaaaactgtttagcaaccctggacatttttgcaggatgcggtgccttatc
3421 3480
ttgtagtttacttttgacaaatcgttgggacctgtaaaaacgtcctacgccacggaatag
T S N E N C L A T L D I F A G C G A L S -

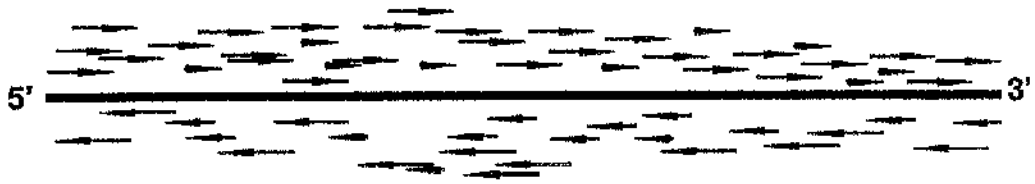

agaggggttgcataagtctggtgcttcatcaactaagtgggctattgaatatgaagaacc
3481
3540
tctccccaacgtattcagaccacgaagtagttgattcacccgataacttatacttcttgg
E G L H K S G A S S T K W A I E Y E E P -
agctggcaatgcattcaaagctaatcatcctgaagctttggtgttcattaacaactgtaa
3541
3600
tcgaccgttacgtaagtttcgattagtaggacttcgaaaccacaagtaattgttgacatt
A G N A F K A N H P E A L V F I N N C N -
tgtaattctcagggctataatggagaaatgtggagatatagatgaatgtatctcaacagc
3601
3660
acattaagagtcccgatattacctctttacacctctatatctacttacatagagttgtcg
V I L R A I M E K C G D I D E C I S T A -
cgaggctgcagaattggcctctaagcttgatgataaggatttgaatagtttaccattacc
3661
3720
gctccgacgtcttaaccggagattcgaactactattcctaaacttatcaaatggtaatgg
E A A E L A S K L D D K D L N S L P L P -
tgggcaagttgatttcattaatgggggcctccatgccagggtttctctgggatgaatag
3721
3780
acccgttcaactaaagtaattacccccggaggtacggtcccaaagagaccctacttatc
G Q V D F I N G G P P C Q G F S G M N R -
atttaacacaagcacttggagtaaagtccagtgtgagatgatattagcgttcttatcctt
3781
3840
taaattgtgttcgtgaacctcatttcaggtcacactctactataatcgcaagaataggaa
F N T S T W S K V Q C E M I L A F L S F -
tgctgattatttccggccgaggtatttcctcttggagaatgtgaggaactttgtgtctttt
3841
3900
acgactaataaaggccggctccataaaggagaacctcttacactccttgaaacacagaaa
A D Y F R P R Y F L L E N V R N F V S F -
taataaaggacagactttccgtttaactttggcttcacttctcgagatgggttaccaggt
3901
3960
attatttcctgtctgaaaggcaaattgaaaccgaagtgaagagctctacccaatggtcca
N K G Q T F R L T L A S L L E M G Y Q V -
gaggtttggtatcctcgaggctggagcttttggtgtttctcagtcaagaaaaagggcatt
3961
4020
ctccaaaccataggagctccgacctcgaaaaccacaaagagtcagttcttttttcccgtaa
R F G I L E A G A F G V S Q S R K R A F -
tatatgggctgcctctccagaagatgtgcttcctgagtggccagaaccaatgcatgtctt
4021
4080
atatacccgacggagaggtcttctacacgaaggactcaccggtcttggttacgtacagaa
I W A A S P E D V L P E W P E P M H V F -


ctctgcccctgagttgaaaatcacattggcagaaaatgtccagtatgctgccgtctgcag
4081 4140
gagacggggactcaacttttagtgtaaccgtcttttacaggtcatacgacggcagacgtc
S A P E L K I T L A E N V Q Y A A V C S -
tactgcaaatggtgctccgttacgggcaataactgttcgtgataccattggtgaactccc
4141 4200
atgacgtttaccacgaggcaatgcccgttattgacaagcactatggtaaccacttgaggg
T A N G A P L R A I T V R D T I G E L P -
agctgttggcaatggagcctctaggacaaacatggagtatcaaagcgatcctatctcgtg
4201 4260
tcgacaaccgttacctcggagatcctgtttgtacctcatagtttcgctaggatagagcac
A V G N G A S R T N M E Y Q S D P I S W -
gtttcaaaagaagatccgaggcaatatggctgtcttgactgatcatatatcaaaggaaat
4261 4320
caaagttttcttctaggctccgttataccgacagaactgactagtatatagtttccttta
F Q K K I R G N M A V L T D H I S K E M -
gaatgagttgaacttgatccgatgtcagaaaattcctaagagaccaggttgtgattggcg
4321 4380
cttactcaacttgaactaggctacagtcttttaaggattctctggtccaacactaaccgc
N E L N L I R C Q K I P K R P G C D W R -
tgatcttccagacgaaaagataaaactttcaactggacaacttgttgatttgataccatg
4381 4440
actagaaggtctgcttttctattttgaaagttgacctgttgaacaactaaactatggtac
D L P D E K I K L S T G Q L V D L I P W -
gtgcttgccacacacagctaagaggcataatcaatggaagggactgtttggtaggttaga
4441 4500
cacgaacggtgtgtgtcgattctccgtattagttaccttccctgacaaaccatccaatct
C L P H T A K R H N Q W K G L F G R L D -
ttggcaagggaatttcccaacttccatcaccgaccctcaaccaatggggaaggttggaat
4501 4560
aaccgttcccttaaagggttgaaggtagtggctgggagttggttacccattccaaccttta
W Q G N F P T S I T D P Q P M G K V G M -
gtgcttccatcccgatcaagatagaattcttactgttctggaatgcgcccgatctcaagg
4561 4620
cacgaaggtagggctagttctatcttaagaatgacaagaccttacgcgggctagagttcc
C F H P D Q D R I L T V L E C A R S Q G -
ctttccagaccactatcaattttctggtaacatcatacacaagcacaggcagattggtaa
4621 4680
gaaaggtctggtgatagttaaaagaccattgtagtatgtgttcgtgtccgtctaaccatt
F P D H Y Q F S G N I I H K H R Q I G N -

```
     cgcggttcctcctcctctggcatttgcattaggaaggaaactcaaggaagcattggatag
4681 ---------+---------+---------+---------+---------+---------+ 4740
     gcgccaaggaggaggagaccgtaaacgtaatccttcctttgagttccttcgtaacctatc

      A  V  P  P  P  L  A  F  A  L  G  R  K  L  K  E  A  L  D  S  -

      taagagcgccaattagaggattagggcgcatctttcaaaaagcatctttttatcatatag
4741 ---------+---------+---------+---------+---------+---------+ 4800
     attctcgcggttaatctcctaatcccgcgtagaaagtttttcgtagaaaaatagtatatc

      K  S  A  N  *

     ttttgtctttcagtgttctggaaacaacccaacccttgtatatagttgttttcttggcta
4801 ---------+---------+---------+---------+---------+---------+ 4860
     aaaacagaaagtcacaagacctttgttgggttgggaacatatatcaacaaaagaaccgat


     tttttcttagtttaatcaattctttgtttaaaaggattgatggaatggattatgctataa
4861 ---------+---------+---------+---------+---------+---------+ 4920
     aaaaagaatcaaattagttaagaaacaaatttttcctaactaccttacctaatacgatatt


     aactcatttttttctatcaaattggggtttcacttggttaggtgataaaaaaaaaaaaaaa
4921 ---------+---------+---------+---------+---------+---------+ 4980
     ttgagtaaaaaagatagtttaaccccaaagtgaaccaatccactatttttttttttttttt


     aaaaaaa
4981 ------- 4987
     ttttttt
```

Appendix-III

Sequencing map of the pea cytosine MTase cDNA*



Scale-500bp

***The final sequencing map is reconstructed from RP23-1, Rp12-4 and A17 clones**