cytosine, in which the 5-carbon of normal cytosine has been replaced by nitrogen. Thus 5-aza-cytosine acts as a strong inhibitor of most of the cellular enzymes, which add methyl groups CH_3 to cytosine in that same 5-position. Now it turns out that when mammalian or plant cells are treated with 5-aza-cytosine, certain genes do indeed become de-methylated and make more messenger-RNA. Hence, loss of DNA methylation was shown in some cases to correlate with increased gene activity.

Research involving DNA methylation in the years 1980–1990 seemingly became bogged down on account of lack of technical innovation. Thus the *Cold Spring Harbor Symposium* volume 47 for 1982 shows, for example, endless patterns of *Hpa* II/*Msp* I digestion along with numerous 5-aza-cytosine experiments! Since 1990 however, that field of DNA methylation has been reborn due to many new techniques, some of which will be described below. Furthermore, when new methods in methylation have been combined with innovations from normal genetics such as PCR, four-colour sequencing or microarrays (see Chapter 10), a great variety of new approaches to the understanding of cytosine methylation and its effects in biology have become available.

The most important breakthrough came in 1992, when Marianne Frommer, Geoff Grigg, Douglas Millar (and later, Susan Clark) developed a simple chemical method for determining whether any particular C base in total genomic DNA might be normal cytosine or else 5-methyl-cytosine. First, they treated their genomic sample with a reactive chemical called sodium bisulfite or $NaHSO_3$ (used also to clean swimming pools). When applied to single-stranded DNA, sodium bisulfite adds a sulfite group (SO_3) to the 6-position of cytosine: see Fig. 11.3(a). But bisulfite does not react with methyl-cytosine on account of steric hindrance from the nearby 5-methyl group; see Fig. 11.3(b). Next, they subjected their modified sample to high pH, so that all sulfonated cytosines would convert to uracil U, which is almost like thymine T: see Figs 11.3(a) and 2.13. Finally, they amplified one small part of their converted sample by PCR using two specific primers, and analyzed its base sequence by the di-deoxy method: see Chapter 10. They found that every methylated cytosine continued to read as C, whereas every unmethylated cytosine had been converted by bisulfite from C to U, which is read as T in the PCR and sequencing reactions.

The bisulfite method has since proven to be very valuable for analyzing 5-methyl-cytosine content in clinical DNA samples, in order to determine whether or not someone has a particular disease. For example, certain genes will lose or gain a particular series of methyl groups when any cell turns cancerous, with sufficient

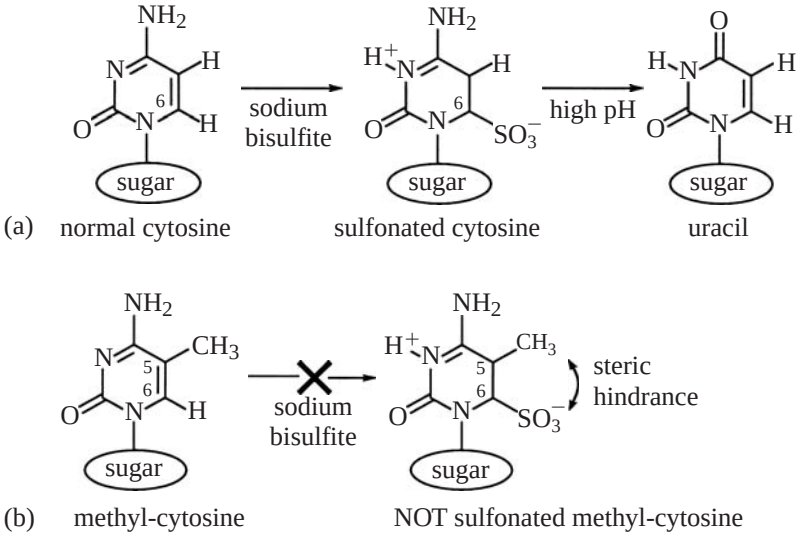


Figure 11.3 A summary of the bisulfite method that has proven very useful to distinguish normal cytosine from 5-methyl-cytosine. In (a) normal cytosine reacts well with sodium bisulfite ($NaHSO_3$) so as to leave sulfonate (SO_3) at the 6-ring position; then it converts to uracil once treated with high pH. In (b) 5-methyl-cytosine will not react with sodium bisulfite, due to steric hindrance between the 6-ring position and a large 5-methyl group.

reliability to enable this small molecular marker to be used as a diagnostic method. Soon, indeed, we should see new, non-invasive tests for many common forms of cancer (e.g. prostate, colon or breast), based on bisulfite-PCR analysis of low-level tumour DNA methylation in the blood or stool. Such DNA-based tests could eventually replace surgical or needle biopsies which, in addition to being painful, can also spread cancer cells undesirably through the lymph system when the needle disturbs the cancerous growth.

Scientists are even contemplating today an 'Epigenome Project', where cytosine methylation would be studied in many different human tissues or clinical samples, by comprehensive four-colour sequencing of both normal and bisulfite-treated DNA. Some people are actually discussing a 'methyl-SNP' project, where millions of different changes in cytosine-to-methyl-cytosine ('single-nucleotide-polymorphisms') would be used for the prediction of long-term human disease.

In addition to the chemical bisulfite method for detection and location of methyl-cytosine within a complex genome, we have also seen the discovery of more bacterial enzymes which are sensitive to 5-methyl-cytosine content in their cutting of DNA. Recall that the DNA-cutting action of *Hpa* II but not of *Msp* I is blocked by CG

methylation at CCGG tetramers (Fig. 11.2(a)). Similarly, we know now that the activity of the enzyme *Sma* I but not of *Xma* I is blocked by CG methylation at CCCGGG hexamers; while all of *Hha* I, *Xho* I and *Not* I are blocked by CG methylation at their respective tetramer, hexamer and octamer sites GCGC, CTCGAG and GCGGCCGC. These are only a handful of examples from the hundreds of different bacterial 'restriction enzymes' that have now been discovered. Many of them show a sensitivity to cytosine methylation at a particular sequence; and they may be used to probe 5-methyl-cytosine within DNA having a wide range of sequences.

As well as the large number of 'cutting' enzymes, a broad variety of bacterial enzymes called 'methylases' have also been discovered, which will add new methyl groups CH_3 onto cytosine bases at many specific sequences of DNA. Here, we shall use the convention that m-C is a methylated C base. From analyzing diverse strains of bacteria, scientists were able to discover for example: *M. Sss* I, that methylates all CG to m-CG; *M. Hpa* II, that methylates CCGG to Cm-CGG; *M. Msp* I, that methylates CCGG to m-CCGG; and *M. Hha* I, that methylates GCGC to Gm-CGC. In all cases, the methylation occurs on both strands.

Lastly, we have learned much over the past ten years, about the various enzymes which add methyl groups to cytosine bases in human, mammalian or plant cells. The most abundant of these is called Dnmt1 in animals or MET1 in plants, as shown in Fig. 11.4(a). These convert DNA which has been methylated on only one strand into DNA that has been fully methylated on both strands. But how does it come about that DNA is sometimes found to be methylated only on one strand?

After replication, all bases newly added by the polymerase enzyme to the template strand are ordinary C, A, T and G, without modification (see Fig. 10.2(b)). Consequently, although the template strand may be methylated, the new strand will not be. Such DNA is said to be 'half-methylated' or 'hemi-methylated'; and so the replication process will create two half-methylated 'daughter' DNAs from the fully methylated 'parental' DNA. The purpose of Dnmt1 (or MET1) is to restore, or 'maintain', the original pattern of methyl groups that was present before replication.

Another abundant enzyme, which is found only in plants, is called CMT3 in *Arabadopsis*, or ZMET2 in maize. It converts half-methylated CNG to fully-methylated CNG where N may be any base. All of those enzymes are called 'maintenance methylases', because they serve to maintain some pre-existing pattern of methyl groups within the DNA, but cannot start a new pattern *de novo*. Maintenance methlyases are found often in the vicinity of replication

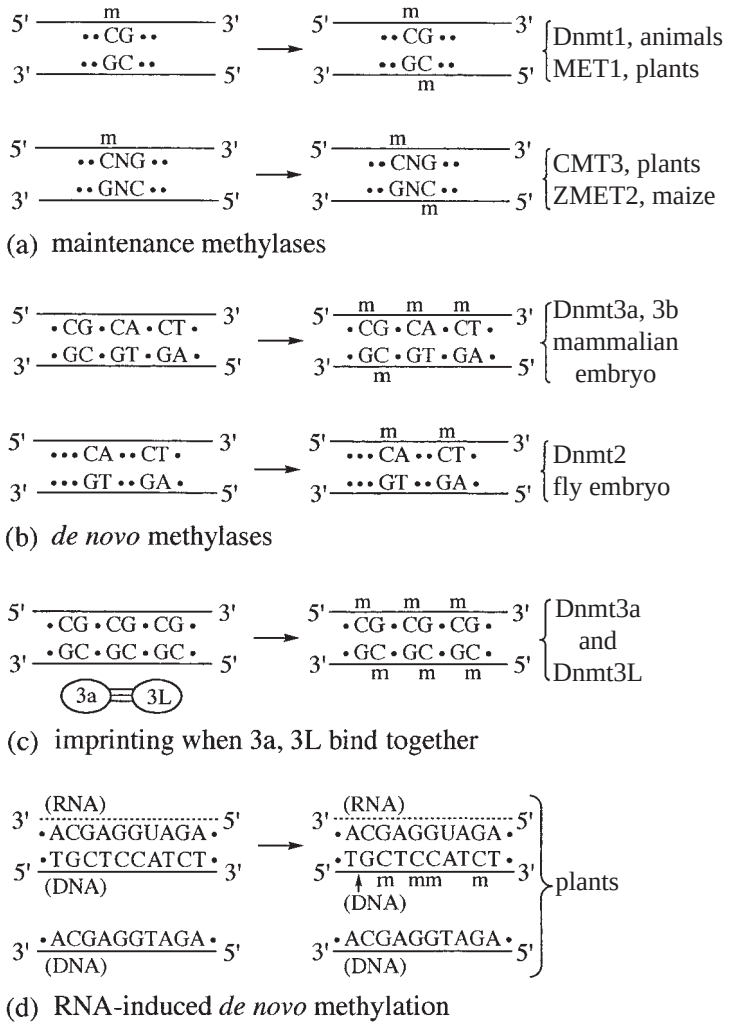


Figure 11.4 A summary of the enzymes that add methyl groups to cytosine in animals and plants. In (a) we show the 'maintenance methylases' which will add a second methyl group to the strand opposite where one already exists, but will not add two methyl groups *de novo*; these are called Dnmt1 in animals or MET1 in plants for the sequence CG; or else CMT3 in plants or ZMET2 in maize for the sequence CNG where N, N can be any base and its complement. In (b) we show the '*de novo* methylases' which will add two new methyl groups to any unmethylated site; these are called Dnmt2 in fly embryos for the sequences CA or CT, or Dnmt 3a or 3b in mammalian embryos for the sequences CG, CA or CT in order of prevalence (high on left, low on right). In (c) we show an 'imprinting protein' Dnmt3L which has no methylase activity of its own, yet helps *de novo* methylase Dnmt3a to methylate genes in mammalian embryos in particular circumstances. In (d) a short, single-stranded region of RNA binds to a complementary strand of double-helical DNA, thereby displacing the other DNA strand – here shown below the RNA-DNA hybrid; which is a signal for the methylase to methylate *every* cytosine of the bound DNA strand.

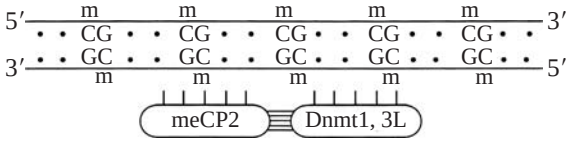
forks in dividing cells. There some old DNA strand carries an established pattern of methyl groups, whereas the newly-made strand does not, at least until a maintenance methylase goes to work on it!

Another general class of enzymes are known as *de novo* methylases, because they can create a new pattern of methyl groups within the DNA where none previously existed – for example in early embryos. As shown in Fig. 11.4(b), fly embryos contain a *de novo* methylase called Dnmt2, which methylates CA or CT rather than CG, yet has little activity in humans. Mammalian embryos and embryonic stem-cells contain two important *de novo* methlyases called Dnmt3a and Dnmt3b. Both will convert unmethylated CG to fully-methylated CG, and also will methylate CA or CT weakly. A related protein Dnmt3L shows no activity of its own, yet helps Dnmt3a to make a methylation 'imprint' in female pro-nuclear DNA (i.e. the DNA contribution from an egg cell prior to nuclear formation), as shown in Fig. 11.4(c).

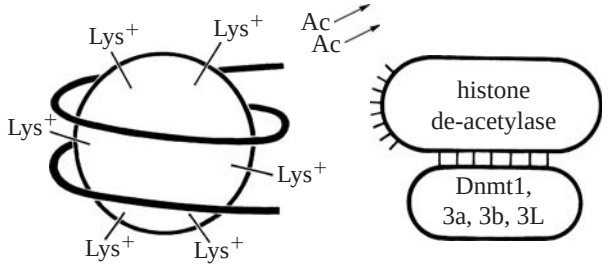
Lastly, there seems to exist in plants a special *de novo* methylase which adds methyl groups to *all* cytosine bases, within a small region of RNA-DNA binding as shown in Fig. 11.4(d). That methylase is somehow directed by a short, 25-to-50 nucleotide segment of RNA, which binds to specific sequences of DNA within a plant chromosome; yet so far its amino-acid identity remains unknown. The existence of such an RNA-directed, *de novo* activity has nevertheless enabled scientists to manipulate the DNA methylation patterns of plants in a directed fashion, using short pieces of added non-cellular RNA; but not yet in animals, where the enzymes seem to be different.

In principle, those methylation enzymes might methylate *all* CG steps in the genome; but in practice, only a subset of CG steps are actually methylated. How might these enzymes, whether maintenance or *de novo*, know where to add methyl groups within a total genome? It is now known that such methylases will be targeted or 'recruited' to particular genomic regions by other proteins which already happen to be binding there. For example, as shown in Fig. 11.5(a), the mammalian protein 'methyl-CG-2' or MeCP2 binds preferentially to CG methylated DNA of any kind; and also binds to the maintenance methylase Dnmt1 (or to the imprinting protein Dnmt3L), as a way of localizing methyl-adding activity where it is most needed.

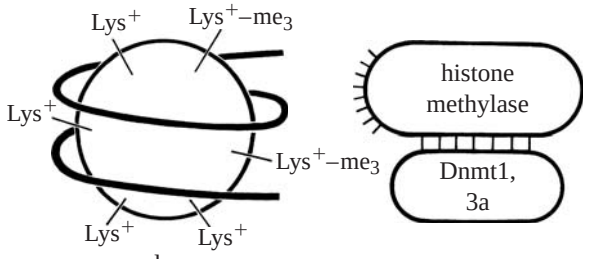
Mammalian or plant cells also contain a wide range of enzymes which can modify the histone proteins of chromatin. It is known from recent work that such histone-modifying enzymes can localize DNA methylases to particular genes. For example, the most abundant



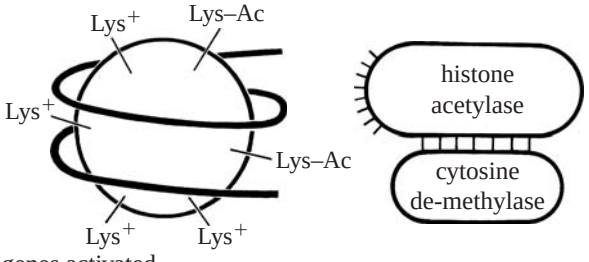
(a) genes repressed



(b) genes repressed



(c) genes repressed



(d) genes activated

Figure 11.5 A summary of protein-to-protein contacts between DNA methylases and other cellular enzymes. In (a) we see that the major methyl-cytosine-binding-protein 'methyl-CG-2' (or MeCP2) will bind to both methylases Dnmt1 or Dnmt3L, thereby linking methylation with maintenance or imprinting activities. In (b) we see that the major histone de-acetylase that removes acetyl groups from lysine on histone proteins, will bind to all of Dnmt1, 3a, 3b and 3L, thereby linking histone deacetylation with all forms of cytosine methylation. In (c) we see that histone methylase (which adds methyl groups to lysine 9 of histone H3) will bind to either Dnmt1 or Dnmt3a, thereby linking histone methylation with maintenance or imprinting. Finally in (d) we wonder whether histone acetylase (which adds acetyl groups to lysines) might bind to an unpurified cytosine de-methylase. All of phenomena (a, b, c) are associated with repressed genes, whereas (d) is associated with active genes.

cellular enzyme which *removes* an acetyl group from lysine is known as 'histone de-acetylase 1'. It binds well to all of the cytosine methylases Dnmt1, Dnmt3a, 3b and 3L, as shown in Fig. 11.5(b). Such protein-to-protein associations serve to link histone de-acetylation with cytosine DNA methylation, as two redundant factors which can both repress genes. There also exist 'histone methylases' which can add methyl groups to lysine (e.g. lysines 4 or 9 of histone H3), especially within repetitive human DNA sequences (some of which have been given names, such as 'Alu', 'SINE' or 'LINE'). Those histone methylases also bind to the Dnmt1 or Dnmt3a methylases, and repress genes similarly, as shown in Fig. 11.5(c).

When histone de-acetylase is inhibited by a small molecule such as sodium butyrate, valproate or trichostatin A, the activity of other enzymes known as 'histone acetylases' (which *add* acetyl groups to lysine) become predominant. Then we see highly-acetylated histone proteins in living cells, which readily activate genes as shown in Fig. 11.5(d). Such histone acetylases may bind in principle to (hypothetical) cytosine de-methlyase enzymes, which would remove methyl groups from nearby CG sequences.

So much for the enzymes which control DNA methylation, or histone acetylation/methylation: they seem very complicated! But what about the *effects* of such epigenetic changes on biological activity? When studied in a broad perspective, DNA methylation is now known to regulate five general kinds of biological activity in higher animals or plants: (i) formation of specific tissues in the embryo (e.g. *de novo* methylation by Dnmt3a or Dnmt3b); (ii) imprinting of genes in a newly fertilized egg cell (see Chapter 10); (iii) transcriptional repression of an entire chromosome (e.g. one of the two female X chromosomes in mammals); (iv) transcriptional repression of highly selected genes (e.g. by blocking the binding of transcription factors to a promoter, or by recruiting MeCP2); and (v) suppression of invading foreign DNA.

As shown in Fig. 11.6, the male contribution to the genome of the embryo (or 'male pronucleus') becomes actively demethylated by unknown enzymes (which we may call 'sperm de-methylases') just after fertilization, and before the male and female pronuclei join together as homologous chromosome pairs. All DNA molecules of both male and female origin then become passively demethylated by dilution of the parental strand through successive cell-divisions, for a brief period of time when the methylase enzymes remain absent. Finally, once *de novo* methylases Dnmt3a and Dnmt3b as well as maintenance methylase Dnmt1o (a variant of Dnmt1) begin to be expressed around the four-to-eight-cell stage, the total DNA of both genomes slowly begins to gain methyl groups again,

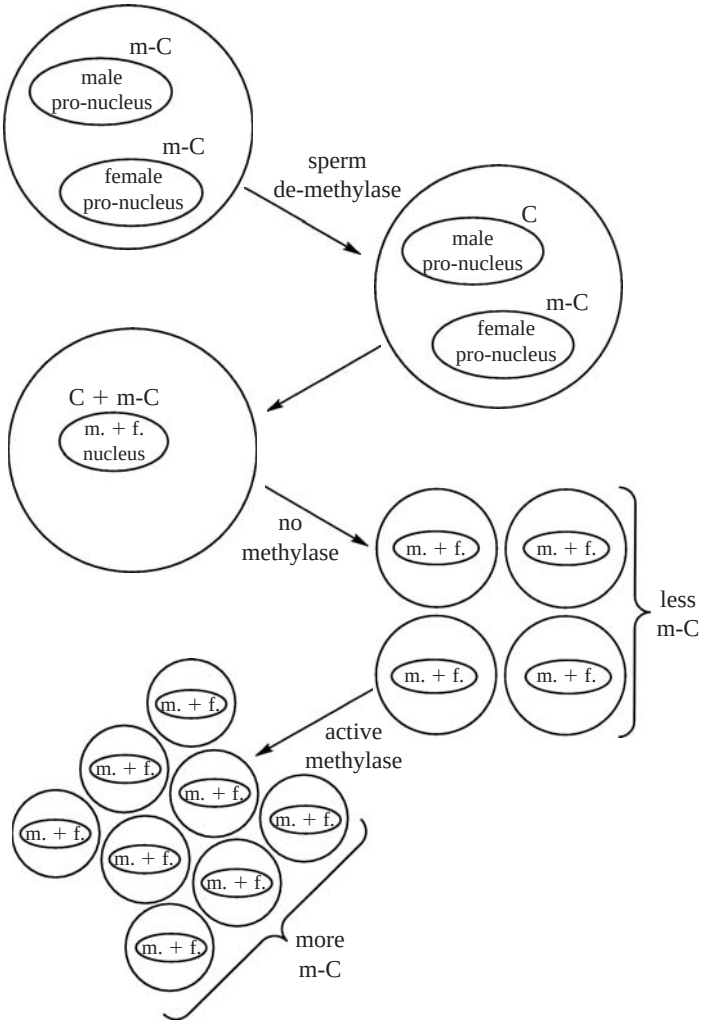


Figure 11.6 A summary of changes in DNA cytosine methylation that have been observed within early mammalian embryos. The fertilized egg (upper left) contains initially both male-origin and female-origin chromosomes from sperm and egg respectively, both of which are highly methylated. Next a sperm-demethylase activity removes most of the cytosine methylation from male-origin chromosomes but not the female, within 4 to 6 hours as shown at upper right. Then both half-sets of chromosomes merge to form homologous pairs (middle left), after which the newly-created cell undergoes limited growth and division in the absence of any methylase enzyme activity. Hence even female-origin chromosomes will become passively demethylated through dilution, when methyl-cytosine is replaced by normal cytosine during DNA replication (middle right). Finally, all of the cytosine methylases Dnmt1o (a variant of Dnmt1 found in embryos), Dnmt3a and Dnmt3b are expressed at a late stage of development: thus various genomic sites on both male and female-origin chromosomes will undergo *de novo* DNA methylation, which seems essential for the formation of correct body tissues.

thereby adopting particular patterns of methylation which may help to determine particular kinds of tissue.

'Imprinted' regions of both male- and female-origin genomes (recall Chapter 10) represent an exception to this general picture, because only male-origin genes remain active in some cases, and female-origin in others. Another exception is seen just in female embryos, where one of the two homologous X chromosomes shuts down and becomes transcriptionally inactive or 'silent'. The choice of which of the two chromosomes will be turned off appears to be random; and it is seems that once a 'choice' has been made, the closing down must spread cooperatively throughout one and only one of the two chromosomes. By a poorly understood process, the *Xist* RNA triggers DNA methylation on one copy of X but not the other (see Chapter 10).

When farm animals such as pigs, sheep or cattle are cloned by 'nuclear transfer' technology, then the combined male and female DNA contributions from a mature cell-nucleus are added all at once to an empty pig, sheep or cow egg-cell. Often the well-organized process of de-methylation and subsequent re-methylation (see Fig. 11.6) then goes badly wrong, and so the vast majority of nuclear-transfer embryos will be aborted. Obviously the male pronucleus in such a case will not be de-methylated specifically *versus* the female; while the carefully-timed expression of cytosine methylases in the four-to-eight cell embryo often becomes aberrant.

Similarly, when transgenic mice are prepared so that most of their Dnmt1 methylase activity is absent, the effects on early development are profound. Such mice appear as runts at birth; they carry chromosomal abnormalities; and by six months of age they are replete with aggressive cancers.

In adult animals or plants, many changes of DNA methylation patterns can be detected. Those changes reflect both a normal, time-dependent aging of the organism, as well as an abnormal conversion of normal cells into cancer cells. It is not certain yet whether changes of cytosine methylation are a primary event in cancer, or are secondary to other phenomena which silence genes, for example histone de-acetylation. In any case, such changes of DNA methylation seem reliable enough to diagnose cancer for medical purposes, without the need for surgical biopsies (as evidenced by ongoing clinical trials).

How precisely might changes of DNA methylation be related to aging or cancer in mammals, at a deep biochemical level? A large part of the aging and cancerous processes has to do with imperfect enzymatic repair of mutations in which a C–G base pair has become a T–G base-pair; these mutations arise naturally and frequently in methylated DNA. As mentioned already, it is well known that animal

or plant genomes, which contain mostly methyl-cytosine at their CG steps, also show a reduced incidence of those CG steps, of only 10% to 20% relative to the incidences of GC, CC or GG. Why should such a profound deficiency exist? It is simply because cytosine bases C are continually 'de-aminating' their base rings at a constant, slow rate, to yield the RNA base uracil U as shown in Fig. 11.7(a). Such ongoing de-amination changes CG steps to UG, or GC steps to GU, or CC steps to CU, once every minute or so, at some location in a large genome! But in fact, most cells contain a repair enzyme known as 'uracil glycosylase', which goes around cutting out U bases wherever it can find them, and then a repair polymerase pairs the opposite G base with C in order to fix the damage (see Fig. 11.7(c)).

Yet if the original base happens to be methyl-cytosine rather than cytosine, such deamination will leave the normal base thymine T, rather than the abnormal base uracil U as shown in Fig. 11.7(b). (This process has the same end result as that shown in Fig. 11.3(a); but without the mediation of sodium bisulfite.) De-amination will thus change any m-CG step to normal TG, which uracil glycosylase cannot repair. There does exist another repair enzyme called 'thymine glycosylase', that recognizes the base-pair mismatch of TG on one strand with CG on the other (i.e. where G pairs with T), then cuts out the T and replaces it by C with the help of a repair polymerase (see Fig. 11.7(d)). But thymine glycosylase is far less efficient than uracil glycosylase at removing defects. Hence any individual m-CG step will decay to TG in mammals with a half-life of perhaps 1000 to 10 000 years, unless that particular m-CG happens to be essential for the working of the cell. All of these processes taken together result in a long-term deficiency of CG steps within a mammalian genome, while the frequency of the decay products TG and CA will often be enhanced.

Now when human DNA from adult tissues is compared with human DNA from embryos, scientists often find numerous CG to TG step mutations that correlate with age, auto-immunity or cancer. For example, most mutations in colon or brain cancer are natural de-aminations of that kind. Other mutations may result from damage by environmental sources, for example sunlight or carcinogens. In any case, it is likely that there will soon be diagnostic assays for colon cancer, aging, or infertility, which will be based on the analysis of specific CG to TG step mutations which accumulate naturally in our tissues as we age, but at different rates in different people. If the normally unmethylated CG steps within a 'CG island' become methylated aberrantly, for example, that is usually very serious; because it may shut down some essential gene for a tumor-suppressor protein, and thereby cause cancer directly.