The di-deoxy, four-colour method represents the most common means by which scientists today analyze their DNA sequences, as obtained from PCR or carrier-plasmid amplification. Yet it should be stressed that a fourth general class of methods is currently under development, which may even replace PCR and four-colour sequencing to some extent in the near future. Thus, many workers have made various kinds of DNA microarray or 'biochip', which can detect any short DNA sequence which may be present in a mixed chromosomal DNA sample, by means of specific Watson–Crick base-pairs or 'hybridization'; and which can also measure the relative abundance of some particular DNA or RNA molecule, in a sample which may contain all of the different DNA or RNA molecules found in a living cell.

As shown schematically in Fig. 10.5(a), those workers first prepare a glass slide or nylon filter, to which they attach chemically many different single-stranded DNA oligo-nucleotides of varied sequence in a regular microscopic array. The oligomers are typically about 25 nucleotides long. Next, they add a sample of cellular DNA or RNA, melted so that its two strands separate, and allow the two single strands to form specific Watson–Crick base-pairs with any matching oligo-nucleotides on the chip. Usually one adds the cellular sample under stringent conditions such as high temperature, and in the presence of moderate amounts of urea or formamide, to select against non-Watson–Crick base-pairs during the binding step. Finally, if each DNA or RNA sample is labelled beforehand with a fluorescent dye (or radioactive atom), by either enzymatic or chemical means, then each cell-derived sample will give a reproducible pattern of fluorescent (or radioactive) light signals on the biochip, which may be recorded under a light microscope (or by a phosphor screen).

Oligonucleotide microarrays have been useful to some extent for determining variations of base sequence between DNA from different individual samples (i.e. 'genotyping'); but they have been more useful for measuring the relative amounts of messenger-RNA present in any living cell (i.e. 'expression'). Thus, researchers can now compare the relative transcription of DNA into RNA for thousands of different genes in any complex set of biological samples: say before and after a patient has been given a therapeutic drug, or before and after certain cells have turned cancerous.

By a fifth new class of techniques, scientists have been able to speed up the process of medical diagnostics, by looking just for certain naturally occuring genetic variations in people. These are small mutations and are sometimes known as 'SNPs' (single nucleotide polymorphisms) which may sometimes be related to human disease. Once a point-mutation or SNP is determined from sequencing

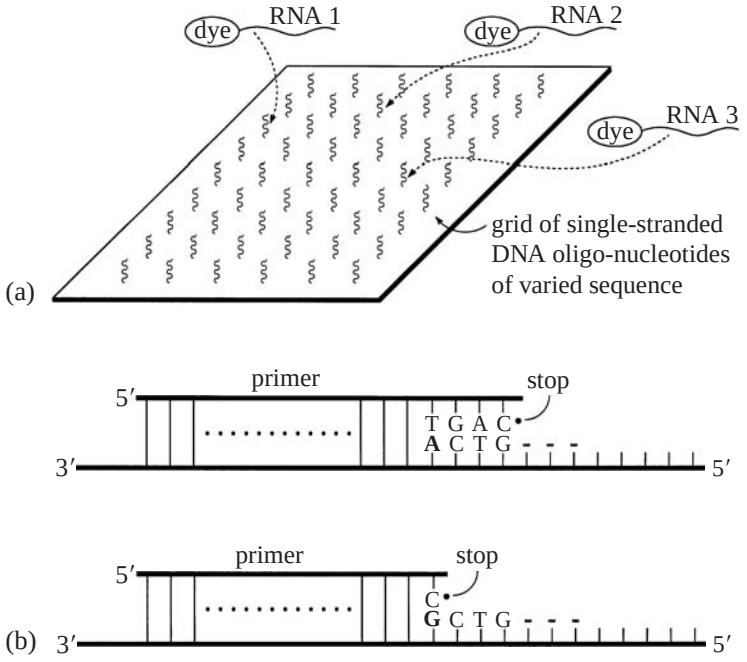

Figure 10.5 (a) A schematic illustration of how many different single-stranded DNA oligonucleotides may be attached in a regular microscopic array to a surface, so as to create a microarray or 'biochip' to which long cellular DNA or RNA molecules may bind in a specific fashion, and be detected through attached dyes. Here, only a small lattice is shown, but in practice the arrays may contain thousands of elements. (b) A hypothetical example of SNP determination: DNA polymerase will extend a short primer from its 3'-end, along a template DNA which may read as either **A**CTG or **G**CTG for the next four bases. If three deoxynucleotides T, G and A are added to the reaction, along with one di-deoxynucleotide C, that short primer will increase in length by four bases if the sequence reads **A**CTG, *versus* just one base if the sequence is instead **G**CTG. Hence the SNP A/G may be determined, by measuring an overall length for the primer once the polymerase reaction is completed.

several variants of a human genome, how do physicians find out whether their sick patient might contain such a SNP, as a possible cause of the ailment?

First one amplifies by PCR certain discrete sections of human DNA, which are thought to be of possible medical interest (typically 100 to 1000 base-pairs). Next, one can use each of those PCR products as a template for a special *polymerase extension reaction*, which includes only three of the normal deoxy-nucleotides A, T, C or G; with one of the special di-deoxy nucleotides which terminate a chain. For example as shown in Fig. 10.5(b), a specific primer of 20 bases (which has been chosen to stop just short of the suspected SNP) has been added to our PCR-derived DNA, and then a

polymerase enzyme will increase the length of that primer by different amounts, depending on whether it detects a template A or else G just after the primer 3′-end. Hence if the first few template bases to follow the primer 3′-end are **A**CTG, the polymerase will add di-deoxy-C after the primer grows longer by four nucleotides to 24 bases total. But if the first few template bases are **G**CTG, the polymerase will add di-deoxy-C after the primer grows longer by just one nucleotide, to 21 bases total.

By choosing which primer to synthesize, and which di-deoxynucleotide to add (along with the appropriate three normal deoxy-nucleotides), practically any SNP may be identified with ease. Indeed, hundreds or even thousands of polymerase extension reactions can now be analyzed rapidly through use of *time-of-flight mass spectrometry* ('MALDI-TOF'), which can measure the mass of the DNA to very high precision and with great accuracy. There the length of time required for any extended primer to reach the detector depends sensitively on its overall size, say 24 *versus* 21 nucleotides in the example given above. For scientists without access to such a costly mass spectrometry machine, an inexpensive alternative is to use two primers, each with a different terminal 3′ base, and each with a different-color dye attached to its 5′-end; then each SNP (say **A**CTG *versus* **G**CTG) can be amplified by normal PCR to yield a product of distinctive color in a capillary or gel.

So far scientists have mapped an astonishing 2 to 3 million SNPs within the human genome, and are looking for ways to use that abundant, but often meaningless, information. Most SNPs are simple changes such as T to C, or A to G (i.e. pyrimidine to pyrimidine, or purine to purine): those are called 'transitions'. Some are changes from T to G or A, or from C to G or A: those are called 'transversions'. Many or most of these changes are probably harmless, and their distribution often reflects genetic variation among different races, or among individuals within a race. Scientists are also examining polymorphisms in other organisms, such as fruit flies, to compare the naturally occuring variations within populations of breeding animals, when certain subgroups of any species cannot interbreed due to separation by geography. These have provided some insight into how genes 'flow' and change in the wild.

While most of the SNPs in humans are probably harmless, a few may be dangerous in themselves, or may be linked with defective genes. By statistical means, scientists have associated certain of those SNPs with human disease: for example, asthma, lung cancer, high blood pressure, diabetes, lupus, or migraine. Yet SNP-disease associations remain very controversial at present, since they are based only on weak statistical evidence. Furthermore, they do not

address any environmental factors which might influence whether or not someone becomes ill.

By a sixth new class of methods, scientists can now determine whether any C base in human chromosomal DNA might represent 'normal' cytosine, or else a chemically modified form known as 'methyl' cytosine. There a $-CH_3$ methyl group is added to the 5-position of the cytosine base, just as for thymine *versus* uracil (see Fig. 2.13), by special 'methylase' enzymes which are present in any cell. A full discussion of cytosine methylation will be deferred until Chapter 11.

With all of these novel technologies in hand – transgenics, PCR, four-colour sequencing, microarrays, SNPs, detection of methylation – scientists have been able to make rapid progress at finding the underlying genetic origins of many diseases. These usually involve tiny but specific defects in human chromosomal DNA. So, as the second general part (b) of our survey, let us see how the techniques described in part (a) are used in modern medicine.

Medical researchers will typically identify some small group of people who possess a particular genetic disease, and then screen their DNA within probable protein-coding regions to search for a deleterious point-mutation or slight alteration of length. Researchers can also identify a broad region of probable genetic defect under the light microscope, by studying visual abnormalities in human metaphase chromosomes. But the new molecular methods described above provide greater resolution still, of the disorders which we either can or cannot see by use of light microscopy.

Once a probable mutation is found by PCR, cloning and sequencing, the next step will often be to use rapid methods of DNA diagnostics, to determine which other members of the human population carry the same mutation. Once we examine a large number of people from diverse ethnic backgrounds, who happen to carry that same mutation (or SNP), will they appear healthy or diseased? Finally, some researchers may choose to make transgenic mice, to which a defective human gene has been added; or else make a similar deleterious mutation within the normal mouse. Will such slight genetic alterations of the mouse cause symptoms of disease identical to those seen in humans? Transgenic mice can also be used to test which drugs might cure any genetic ailment, before embarking on a human trial.

Most genetic defects as found in this way have turned out to be errors within the coding region of some essential protein, which impair its function. Sometimes a single base of the patient's DNA which has been altered by a 'point mutation', say from A to G or from C to T, makes the mutant gene code for an altered amino-acid (or else 'stop') at one location along a protein chain. In other cases,

one or several bases may be added or deleted, so that the mutant DNA codes for a dramatically altered protein by means of a 'frameshift mutation' (see Chapter 1).

By trial-and-error, scientists have slowly begun to associate various mutations in human DNA with many kinds of disease, such as cystic fibrosis, breast cancer, or neurological ailments. For example, it is now possible to predict whether a young woman will have a high chance of getting breast cancer, or whether a young man will have a high chance of getting prostate cancer, from a study of their DNA. These discoveries have prompted much debate about the ethics of medical diagnosis among physicians, as well as a keen interest in DNA from life insurance companies!

The rate of progress has been remarkable: as of 2003, over 200 different genetic or infectious diseases may be diagnosed using PCR and sequencing, by the few medical researchers working in Sydney alone. One finds for example tests for: antitrypsin deficiency, auto-immune syndrome, five different forms of cancer, Charcot-Marie-Tooth disease, chicken pox, cystic fibrosis, viral encephalitis, fragile-X syndrome, Friedrich's ataxia, haemophilia, hepatitis, and herpes simplex virus. The tissues from which DNA may be extracted to perform such tests include: blood, amniotic fluid, spinal fluid, bone marrow, mouth swabs, nasal swabs and mucus (see **www.cs.nsw.gov.au/csls**, Handbook of DNA Testing).

What a cornucopia of DNA technology, to spring up in just twenty years since the PCR method was developed! Furthermore, this abundance of medical information does not even include the numerous PCR tests used for veterinary work (e.g. dogs, sheep or horses), forensic work (e.g. rape or murder), or archeological studies (e.g. ancient humans).

Two examples of human genetic defect are shown in Figs 10.4 and 10.6. The four-color, automated sequencing plots shown previously in Fig. 10.4 were derived from the DNA samples of real patients with neurological disease. We looked previously at Fig. 10.4 in order to explain the general nature of DNA sequencing results; but now let us try to understand those results in their full medical context.

The upper sequencing plot in Fig. 10.4(a) shows DNA from a healthy patient, whereas the lower sequencing plot in Fig. 10.4(b) shows DNA from a neurologically diseased patient, who suffers from a disease called Amyotrophic Lateral Sclerosis or 'ALS1'. Thus, the single-base change shown at position 119 of Fig. 10.4(b) corresponds to a single-copy, heterozygous point-mutation (i.e. the two copies of the same gene in any cell are different): from T (thymine) normally found on both chromosomes of an individual, to T on one chromosome of the patient but G (guanine) on the other.

Because that point-mutation lies within an important protein-coding region of chromosomal DNA, namely part of the gene for superoxide dismutase 1 or 'SOD1', it changes a key amino acid at position 148 of that protein from GTA (valine) to GGA (glycine). While the patient will make both the mutated and the normal version of the enzyme – one from each chromosome – the mutant protein will disrupt the biochemistry of human nervous function, although details of the mechanism are not yet clear.

Other point-mutations within SOD-1 also cause neurological disease, typically associated with wasting of the muscles, leading to paralysis and sometimes death. In order to learn more about ALS1 or related diseases, the reader may wish to consult an excellent general reference on the Internet at **www.ncbi.nlm.nih.gov**, subsection 'OMIM' for 'Online Mendelian Inheritance in Man'. If you proceed to that site, and search for reference number 105400, it will tell you much concerning ALS and its genetic causes; plus a history of the disease and its prevelance in ethnic populations. The OMIM website is a valuable reference for doctors who treat patients with genetic disease, and use PCR and/or sequencing to diagnose inherited ailments.

Next, the electrophoretic gel shown in Fig. 10.6 reveals detailed information about another kind of neurological ailment known as Machado-Joseph disease or Spino-cerebellar ataxia 3 ('SCA-3'), which is also described on the OMIM database under reference number 109150. This particular genetic disease comes about because one small part of human DNA gets bigger or 'expands' during early human development, at a repeated sequence CAG-CAG-CAG-CAG within the gene for a protein called 'ataxin-3'. Within normal individuals the $(CAG)_n$ triplet varies in length from $n = 10$ to 40, but in affected individuals it varies in length from $n = 60$ to 80. Such dramatic expansion of coding DNA at a chromosomal level causes affected individuals to make a larger-than-normal protein, which contains 60 to 80 successive glutamine amino acids. That mutant protein then causes, through indirect means – for example, impairment of mitochondrial function – loss of balance, and wasting of muscles in the hands, feet or face.

Looking closely at Fig. 10.6, one can see a series of PCR-derived products which measure the overall length of that CAG repeat, using DNA samples as supplied by different individuals. All PCR products were labelled with radioactive phosphorous to permit easy detection in a polyacrylamide gel. Four lanes on the left show DNA sequencing reactions for G, A, T or C as produced by the di-deoxy method, but using a gel rather than a capillary: these are there to provide markers of DNA length within the gel. Nine lanes on the right show PCR products of the ataxin gene as derived from real

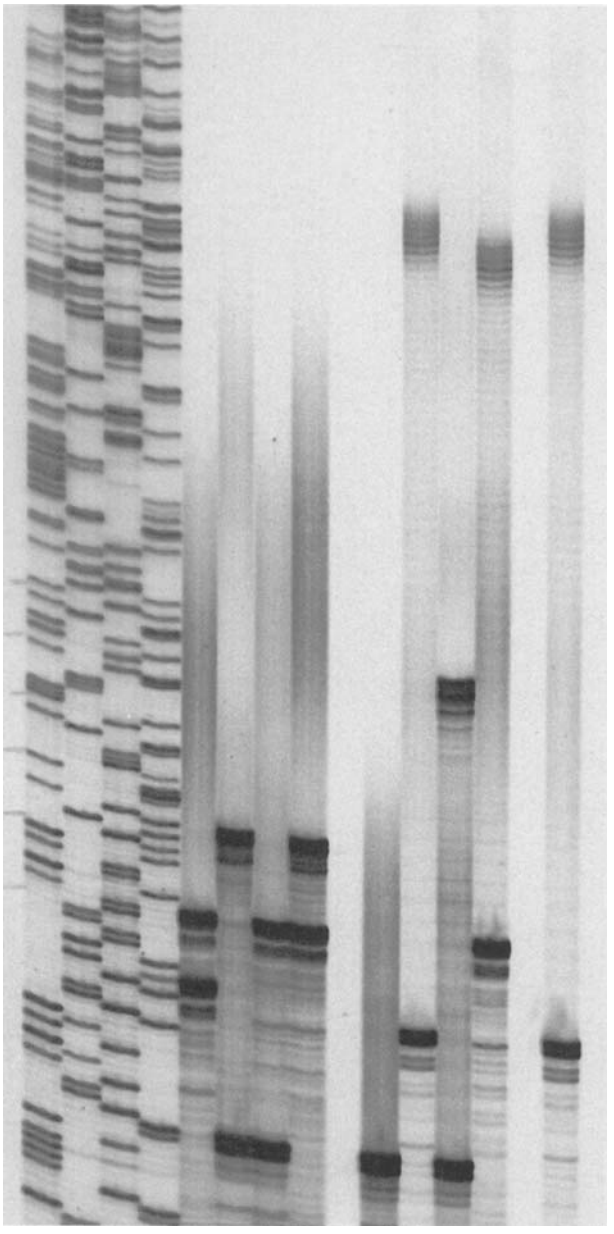

Figure 10.6 An electrophoretic gel shows large size-expansions of a triplet $(CAG)_n$ within genomic DNA, as derived from human patients, all of whom have a neurological disorder, but three of whom have the specific disease Spinocerebellar ataxia or 'SCA3'. Part of the coding region for a protein 'ataxin-3' was measured in each case, using PCR primers which anneal to base sequences on either side. Three of the patients have long $(CAG)_n$ repeats, which run slowly through the gel at top right. The remaining lanes show a normal range of lengths for $(CAG)_n$ repeats. All patients show PCR products from two homologous genes of different length, or one from each chromosome.

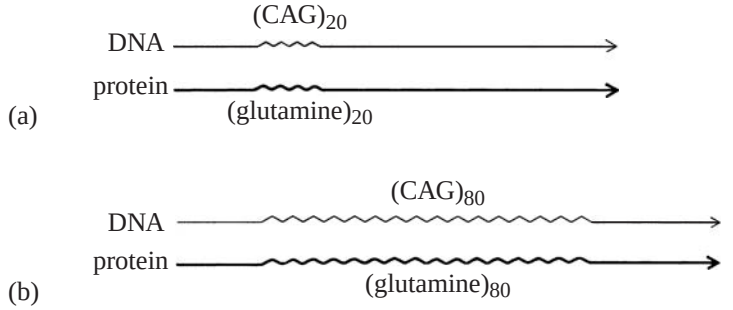

Figure 10.7 (a) A short $(CAG)_n$ repeat of $n = 20$ in healthy individuals codes for a protein containing 20 successive glutamine amino acids, whereas (b) an expanded $(CAG)_n$ repeat of $n = 80$ in diseased individuals codes for a larger protein containing 80 successive glutamine amino acids. It is that long poly-glutamine tract which seems to cause neural defects, by mechanisms which are still not clear.

patients. Out of those nine samples, six show short CAG repeats $n = 10$ to 40 typical of healthy individuals, whereas three show expanded CAG repeats $n = 60$ to 80 typical of diseased individuals.

The biochemical effects of triplet-expansion are illustrated schematically in Fig. 10.7. There we can see how a normal CAG triplet of $n = 20$ gives rise to an ataxin protein with 20 successive glutamine amino acids; whereas an expanded CAG triplet of $n = 80$ gives rise to a larger protein with 80 successive glutamine amino acids.

How many other kinds of genetic disease might be due to triplet expansion, as opposed to a simple point or frameshift mutation? So far more than a dozen different human diseases, mostly neurodegenerative, have been associated with the expansion of triplet repeats; as well as three easily breakable or 'fragile' sites within human chromosomes. For example, Huntington's disease is associated with expansion of a triplet $(CAG)_n$ in a different protein, while myotonic dystrophy is caused by expansion of a triplet $(CTG)_n$. Fragile-X syndrome is caused by expansion of a triplet $(CGG)_n$; while Friedreich's ataxia is caused by expansion of a triplet $(GAA)_n$ (to a size as large as $n = 1000$ base-pairs).

In summary, the vast majority of genetic diseases as detected within the past 20 years correspond to point or frameshift mutations, or else triplet expansions, within some part of human chromosomal DNA that codes for protein. There do exist, however, exceptions to this rule. For example, as we have seen, Spinocerebellar ataxia and Huntington's disease are both due to expansion of a triplet CAG, within the coding region for some protein. Yet myotonic dystrophy and fragile-X syndrome are due to expansion of triplets CTG or CGG respectively in *non-coding* regions of DNA, which make messenger-RNA but are not translated into protein.

How might triplet expansion at a transcribed but *non-coding* sequence influence human health? The triplet expansion of $(CGG)_n$ which causes fragile-X, a disease of mental retardation, is located on the X chromosome, and causes that chromosome to become easily broken or 'fragile'. Hence its genetic locus has been called 'FMR1' for 'Fragile-X Mental Retardation 1'. The FMR1 gene codes for a protein which is necessary for brain function (specifically, regulated translation of messenger-RNA in nerve cells); yet the protein made from FMR1 remains identical for both normal and affected individuals. Instead, it is the overall production of this protein which is lost in fragile-X syndrome. How might that be accomplished?

Near the 5′-end of FMR1, where the DNA has already started to make messenger-RNA, but is not yet coding for protein, a healthy individual contains a base sequence which repeats as $(CGG)_n$ for n from 6 to 50, as shown schematically in Fig. 10.8(a). But in mentally retarded individuals, that $(CGG)_n$ triplet expands to roughly 600 base-pairs as shown in Fig. 10.8(b). So for mentally retarded individuals, the long repeat of 600 base-pairs somehow impairs production of the FMR1 protein.

Do we have any clues as to how this impairment might work? It is usually observed that the expanded CGG triplet of FMR1 becomes modified by cellular methylase enzymes, which convert certain bases within that particular small region of DNA from normal cytosine to 5-methyl-cytosine (see Chapter 11). Since 5-methyl-cytosine can prevent the binding of proteins known as 'transcription factors' to promoter DNA, which enable the transcription of DNA into

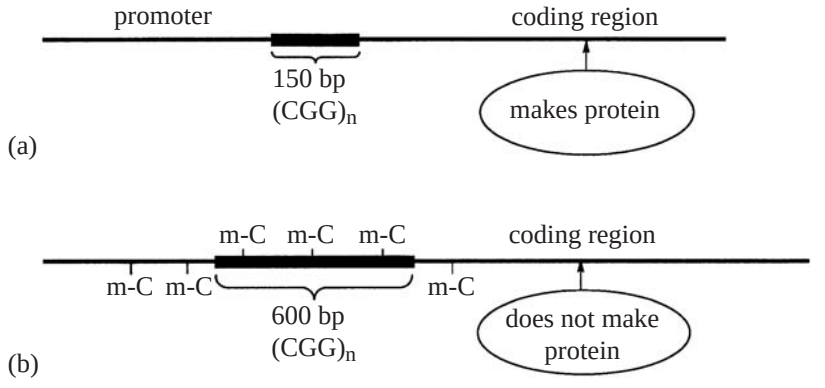

Figure 10.8 (a) When the FMR1 gene contains a repeat of $(CGG)_n$ less than 150 base-pairs long, that DNA still works normally to make RNA and protein. (b) But when the FMR1 gene contains a repeat of $(CGG)_n$ more than 600 base-pairs long, the expanded repeat somehow triggers methylation of nearby promoter DNA (i.e. conversion of cytosine to methyl-cytosine) by cellular methylase enzymes; and so the gene no longer works to make RNA or protein in the normal amounts.

RNA, this may explain how an extended CGG repeat impairs production of FMR1 protein: by blocking messenger-RNA synthesis.

Certain kinds of cancer have been associated, but only in an indirect fashion, with the expansion of *di*nucleotide repeats such as TATATA…, CACACA… or CGCGCG…. Those and other highly-repetitive sequences (such as triplets (as above) or tetramers) may be found at many different places within a human genome. In the majority of cases, however, they do not code for protein; and they seldom cause disease, apart from the few triplets mentioned above. Still, patterns of length variation in non-essential repeats or 'microsatellites' have proven useful for the genotyping (i.e. 'fingerprinting') of individuals within any species. They are now used for example in the identification of individuals in police forensic work (as mentioned earlier); the establishment of animal pedigrees; the diagnosis of colorectal cancer from stool, or liver cancer from blood; and the avoidance of genetic defects when a couple chooses *in vitro* fertilization.

The wide variety of lengths seen for repetitive microsatellites in humans, as well as for triplet-repeats which cause disease, raises a question concerning the molecular mechanism of such expansions. Why should certain trinucleotides, when repeated over and over again, expand in number during early human development so as to cause disease? Why should microsatellites vary in length throughout the chromosomes, typically in non-coding regions?

Most workers believe that certain DNA molecules of highly-repeated sequence can 'slip out' so as to form hairpin or stem-loop structures, during DNA copying or replication; which in turn may induce expansion on a molecular scale. As shown in Fig. 10.9(a), DNA polymerase will usually copy both strands of any pre-existing DNA molecule, so as to form two new double helices. Almost all polymerase enzymes work in a specific 5′to 3′direction while making new strands. Thus the lower polymerase shown in part (a) can copy the lower strand continuously, as it works from right to left; but the upper polymerase can only copy the upper strand discontinuously, in small parts of roughly 300 base-pairs. Those small parts – known as 'Okazaki fragments' – will be joined together later by a 'ligase' enzyme. Once sealed and ligated, each new double helix will go to one new cell. It may be a normal somatic cell, or else a sex cell (egg or sperm).

Now if the DNA that is being copied includes a long repeat of $(CTG)_n$, within either the new strand or the old, that particular sequence may slip out to form a hairpin as shown in Fig. 10.9(b) and (c). Suppose a CTG hairpin forms within the newly-made strand: then some CAG triplets may be copied twice as shown in