Here we shall survey in a concise way what is known about the workings of DNA at a medically relevant level in complex organisms such as man or mouse, subject to the limitations of knowledge as emphasized above. Our survey will cover four areas of research:

(a) important methods which have been developed since 1985, such as transgenics, the polymerase chain reaction (PCR), di-deoxy four-colour sequencing, oligo-nucleotide microarrays, and polymerase extension to measure point-mutations (SNPs);
(b) errors in DNA structure associated with human disease, for example certain point-mutations, frameshift-mutations and tri-nucleotide expansions;
(c) a type of genetic inheritance called 'imprinting', whereby the activity of a gene depends upon whether you have inherited it from mother or from father; and
(d) attempts to fix errors in human DNA that cause disease, by adding new full-length genes of the correct kind, or else by adding new DNA which may repair pre-existing genes: these strategies are known as 'gene therapy' and 'gene correction', respectively. Yet another strategy to suppress gene expression uses short double-stranded RNA; such techniques will be described in Appendix 3.

However, before proceeding to a survey of such diverse subjects, we need to explain why scientists are devoting so much effort to the study of DNA and its associated genes at a detailed level in complex organisms.

Many scientists and physicians in the early 21st century have come to the conclusion that traditional forms of medicine are inherently inadequate for dealing with certain ailments. Those traditional forms include surgical operations, where the doctor cuts away infected tissue with a knife and then repairs the damage by sewing, until nature has healed the wound – a dangerous and lengthy procedure; and the prescription of some small-chemical drug which may be ingested, inhaled, or injected into the body. Usually, the doctor hopes that such a drug will spread quickly throughout the body, so as to find some specific biological target, often either an enzyme or a signalling protein. By binding tightly to its specific target, such a drug may then inhibit the aberrant biological function of the enzyme or the signalling protein, in order to cure a disease, or at least to slow its progress.

We now know that surgical approaches are unlikely to cure most forms of cancer, or to cure infection by viruses such as hepatitis or HIV (which causes AIDS), since the cancerous cells or deadly viruses may spread rapidly throughout the body so as to become

essentially inoperable. Furthermore, a cancer-causing cell often looks much like an ordinary human cell on a molecular scale; and it seldom offers any unique biological target, against which a small-chemical drug may specifically act. This difficulty in distinguishing normal cells from cancerous and virus-infected cells also makes it hard to treat certain diseases through immunization – that is, the activation of the body's natural immune response. And most chemotherapeutic agents that target DNA have powerful but adverse side-effects on healthy cells.

By contrast, the recent development of many new and exciting techniques in molecular biology has allowed scientists and physicians to do new kinds of experiment, which may provide for entirely new forms of medical therapy in the future – as useful say as the discovery of immunization in 1800, or the discovery of penicillin in 1930. As the first part (a) of our survey, let us describe briefly some of the more important techniques which have been developed since 1985, and which have made new kinds of medicine possible, if not always practical, by 2004.

First, scientists can often determine the function of any normal human gene, or else test if a mutant human gene is defective, by injecting the complete DNA for such a gene into a mouse embryo, as shown in Fig. 10.1. The injected DNA will then be incorporated stably into a mouse chromosome by a process called 'illegitimate recombination', which is essentially a semi-random joining of injected human DNA onto pre-existing mouse DNA at rare chromosomal breaks. Thus, wherever the chromosomal DNA of a mouse breaks, infrequently and at random places in an egg cell, it will be repaired by several different enzymes, which usually include a 'ligase' or DNA-joining enzyme to re-connect the broken ends. Any injected human DNA may then take part in such mending, if each end of the human DNA is joined by a repair-ligase enzyme onto both broken ends of pre-existing mouse DNA.

When a mouse with this extra human DNA grows to an adult, it may show differences in physiology or behavior relative to a normal

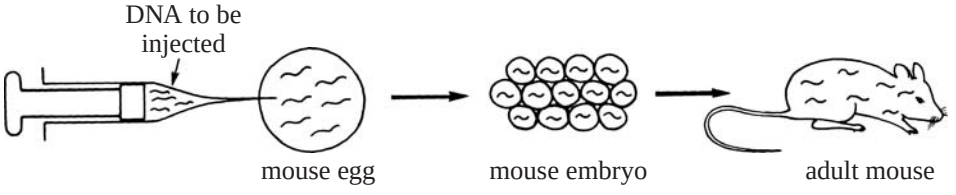

Figure 10.1 New DNA can be added in a permanent fashion to mice or other animals or plants, in order to provide extra genes for these organisms, or else to impair the function of pre-existing genes.

mouse. For example, some genetically-altered mice become especially fat, or grow to a larger size than normal, because their genes have been altered slightly. The additional human DNA may introduce new genes, or else it may impair the function of existing genes.

These altered animals are known in general as 'transgenic' or 'knockout' mice, where certain genes have been added to or deleted from, respectively, the normal mouse set of chromosomes. Often very small changes to the total DNA of an organism will produce large changes in its physical appearance, behavior, or intelligence. Transgenic mice have proved useful for creating animal models of human disease, for instance: prostate cancer, thyroid deficiency, obesity, lateral sclerosis, or Alzheimer's disease. Such genetically-altered mice may also be used to test new drugs which might potentially cure disease in humans.

Similar kinds of transgenic experiment are now being conducted extensively worldwide to make 'improved' animals or plants. For example, various genes have been added to pigs; not only to make them grow fatter or faster for agricultural purposes, but also (and somewhat incredibly) to modify cell surfaces along their bodily organs, so that pig organs will not be rejected by the human immune system, if they are transplanted by surgery into human patients! Genetically-modified pig hearts have already been tested successfully in monkeys, as a step before giving them to humans who might need an organ transplant.

Progress has also been made at cloning adult pigs, sheep and cows from their mature cellular tissues, rather than by normal reproduction using sex cells. Using a microscope, the scientist simply transfers by micro-dissection (or by disruption of the membranes by an electric field, a process known as 'electroporation') a nucleus from a mature cell into an empty egg-cell, while trying not to disrupt any of the mature chromosomes along the way.

Additions of extra DNA to plants have provided for fruit or vegetables of improved quality, for example: tomatoes or strawberries which ripen more slowly than normal; cotton which is more resistant to insects; rice with extra iron as a nutritional supplement; strawberries with extra vitamin C as an antioxidant supplement; and even carrots which carry a measles protein, as a possible and inexpensive way of 'vaccinating' people against the measles virus.

By a second new class of techniques, scientists can now amplify and determine the base sequence of practically any DNA or RNA molecule which might be found within a cell, starting from just tiny amounts of material (say 1 to 100 individual molecules). This means that they can easily isolate new, full-length genes of therapeutic or industrial interest. Also, they can quickly detect aberrant changes in

DNA or RNA sequence as a diagnostic for cancer, or for genetically-inherited diseases. Finally, they can detect low-level infection by micro-organisms such as viruses, bacteria or mycoplasmae; because such micro-organisms will add their own distinctive DNA to the cellular mix, which would not be present in a healthy individual. The genetic composition of the micro-organisms can also be analyzed for drug-resistance profiles, which permits the tailored use of theraupeutic compounds such as antibiotics.

These new amplification and sequencing methods have proven useful as well for the identification of individuals in police forensic work. Some of the DNA within our chromosomes is quite specific to certain individuals, just as are fingerprints. Hence those highly-variable parts of human DNA (known as 'micro-satellites') may be used to identify suspects in a criminal case, from trace amounts of blood, hair, semen or skin left at the scene of a crime.

How exactly do these new amplification methods work? Why are they so sensitive to small amounts of starting material, yet so easy to carry out? In the past, in order to determine the base sequence of any small fragment of DNA, scientists first had to 'clone' that fragment into a circular carrier-molecule or 'plasmid'. Next, the plasmid and its extra DNA would be copied many times over, by the process of DNA replication within a fast-growing culture of bacterial cells. Finally, a large culture of cells would yield large amounts of the desired plasmid, from which the extra DNA insert could be obtained in sufficient amounts, to sequence by chemical or enzymatic means: see Fig. 10.2(a).

That carrier-plasmid method is still used today, when tens of micrograms or even milligram quantities of DNA are desired. Yet it was never reliable for cloning trace amounts of DNA, on the order of 1 to 100 molecules. Futhermore, it was always quite tedious to grow such carrier-plasmids overnight in bacterial cells, and then to extract and purify their small DNA insert, for sequence analysis by chemical or enzymatic means.

Hence the major breakthrough was to do all of that copying and amplification of DNA outside of the cell and in a test-tube. A schematic overview of the new, improved method is shown in Fig. 10.2(b). To any single DNA molecule in solution, one simply adds a heat-stable polymerase enzyme (e.g. *Taq* as taken from the bacterium *Thermus aquaticus*, which grows in hot springs); along with two short pieces of chemically-synthesized DNA which act as 'primers' for the polymerase; plus an abundance of four nucleotide *tri*-phosphates A, T, C, and G. So far in this book, we have described only the nucleotide *mono*-phosphates, which are the units that make up a completed DNA chain. But the preliminary units that

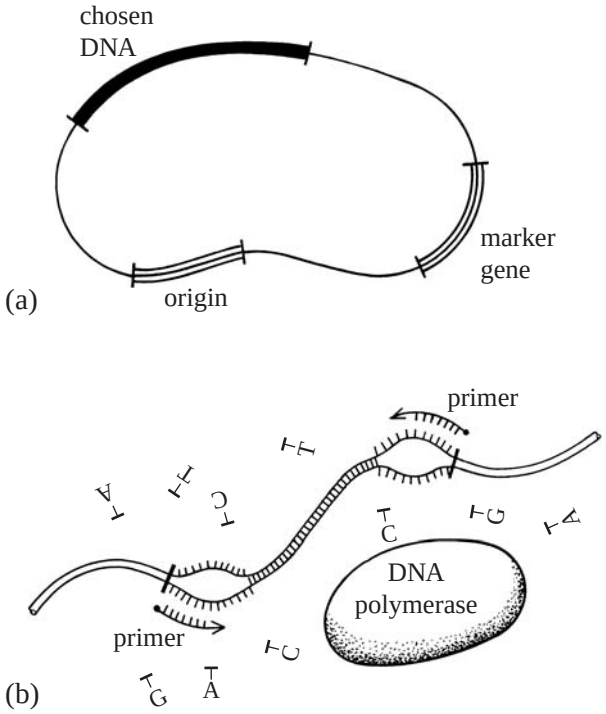

Figure 10.2 (a) Any small piece of DNA may be ligated into a carrier plasmid, that may be prepared on a large scale if bacterial cells which contain the plasmid divide and copy all of their DNA repeatedly. Most plasmids contain an 'origin of replication' to assist in DNA copying by polymerase enzymes; and also a 'marker gene' for resistance to some antibiotic such as ampicillin, chloramphenicol, kanamycin or tetracyclin. This marker gene ensures that all cells will contain the plasmid, when grown in a medium that includes antibiotic. (b) Any small piece of DNA may also be prepared on a large scale, outside of the cell and in a test-tube, by copying it repeatedly with heat-stable DNA polymerase enzymes – a process known as the polymerase chain reaction or 'PCR'. Two short pieces of chemically-synthesized DNA of selected base sequence may be chosen as 'primers', to tell the polymerase where to start copying along each strand.

the polymerase adds successively to any growing chain actually contain *three* phosphate groups, two of which are cut off during polymerization.

When the starting DNA molecule is heated strongly at 95°C, so as to separate its two strands, and then cooled to 55°C, the two short pieces of 'primer' DNA will bind or 'anneal' specifically at each end of the long, starting DNA molecule. Those primers will then form two specific binding-sites for the polymerase enzyme, with one adhering to each of the separated strands. Such double-stranded complexes between short primer DNA and long template DNA will next inititate the synthesis of more double-stranded DNA, when the

reaction is warmed to 72°C in the presence of *Taq* DNA polymerase and the four deoxy-nucleotides A, T, C, and G. Those deoxy-nucleotides will rapidly 'extend' or add sequentially to the 3′-end of each primer, until the 5′-end of each strand of the template DNA is reached.

Hence, our heat-stable polymerase enzyme will make a new, full-length double helix from each of the two single strands which were present in the original sample after heating. Now when those two double-stranded products of the polymerase reaction are heated again to 95°C, then cooled to 55°C and warmed to 72°C, they will yield four single strands in total, which may in turn act as templates for further DNA synthesis. Next, when those four products of the second polymerase reaction are heated and cooled and warmed, they will yield eight single strands which may serve as templates for making more DNA, and so on. After 20 cycles of heating-cooling-warming and polymerase action, our starting DNA molecule will have been amplified from just one to nearly 10 million copies (i.e. by a factor of 2^{20}), or from picogram to microgram amounts.

Because such cyclic amplification of DNA in the test-tube resembles a chain reaction in physics, this new method is called the 'polymerase chain reaction' or 'PCR'. It reminds one of the old story about a king who loses his kingdom, by promising a loyal servant one kernel of corn for the first space on a chessboard, then two kernels of corn for the second space, four kernels for the third space, and so on for all 64 spaces on the board. By the twentieth space, the king has to give his servant more than a million kernels of corn; or by the fortieth space more than a million million, or enough to fill several large ships!

The two primers for each PCR reaction must be chosen carefully, so that they will each contain a short sequence of bases (typically 15 to 30) which binds specifically to two unique parts of the total chromosomal DNA for any organism, using Watson–Crick base-pairs. The template fragments of DNA which may be amplified by this method range in length from several hundred to several thousand base-pairs; or even to tens of thousands of base-pairs using a special method known as 'long PCR'.

This PCR method was the first to be developed, around 1985, and is still the most versatile; yet other useful refinements to methods of test-tube DNA amplification have been developed recently. Two of these are called SDA ('strand displacement amplification') and MDA ('multiple displacement amplification'). Neither of these requires any heating or cooling cycles; they can be performed isothermally at 55°C or 30°C respectively, which may offer some advantages under certain circumstances.

Next, following amplication of some biologically relevant DNA to microgram amounts by the PCR method (or the carrier-plasmid method, as described above), one would traditionally analyze its base sequence by chemical or enzymatic means; and then display or 'read' that base sequence using an electrophoretic polyacrylamide-urea gel, as discussed in Chapter 9. However, most scientists today use a third general class of methods: they send their amplified DNA to a commercial sequencing facility, where special machines are able to analyze hundreds or even thousands of different DNA fragments overnight.

The basic idea behind this procedure is to create, for each sample supplied by the user, a large number of DNA molecules of different length, and then to separate them out by electrophoresis, as explained in Chapter 9. Suppose, for the sake of definiteness, that a sample of 500 base-pairs is to be sequenced. Then the plan is to create a range of samples starting at the same end, and of length 1, 2, 3, … up to 500 base-pairs. Each of these samples must end or 'terminate' in one of A, T, C or G; and if the end-letter can be identified on the gel 'ladder', then the sequence can, in principle, be read off.

But how can the various samples be made to stop at all of these different lengths? And how can the end-letter be identified? The key to the situation is to look at the DNA's sugar–phosphate 'backbone', as shown in Fig. 2.8(b). There we saw the chemical structure of a piece of backbone containing bases and sugar rings, joined by phosphate groups. Now the central picture of Fig. 10.3(a) shows one base in the same sort of backbone, but with two carbon atoms of the sugar ring labelled 2′ and 3′, respectively. Also, on the left is shown the corresponding portion of an RNA backbone; the only difference being that the RNA sugar ring has an OH group attached at position 2′. DNA is called *Deoxy*-ribo-Nucleic Acid precisely because it lacks this oxygen, which is a feature of Ribo-Nucleic Acid's sugar ring.

Finally, look at the right-hand picture of Fig. 10.3(a). This shows the special *di-deoxy* form of the sugar ring: it has no oxygen atom either at position 2′ or at position 3′. In this case the lower phosphate group is not shown, precisely because it cannot attach itself to the ring if an oxygen is absent from position 3′. In other words, the sugar–phosphate chain cannot continue to extend further, once a di-deoxy base has been added by the polymerase enzyme.

In this method, therefore, the surrounding medium in which the polymerase operates contains not only the usual C, A, T and G deoxy nucleotides, just as in the PCR method, but also some of these special di-deoxy nucleotides of all four kinds. And like PCR, the sequencing process requires a short 'primer' DNA for the polymerase to start the copying. Thus, when the polymerase is proceeding along a

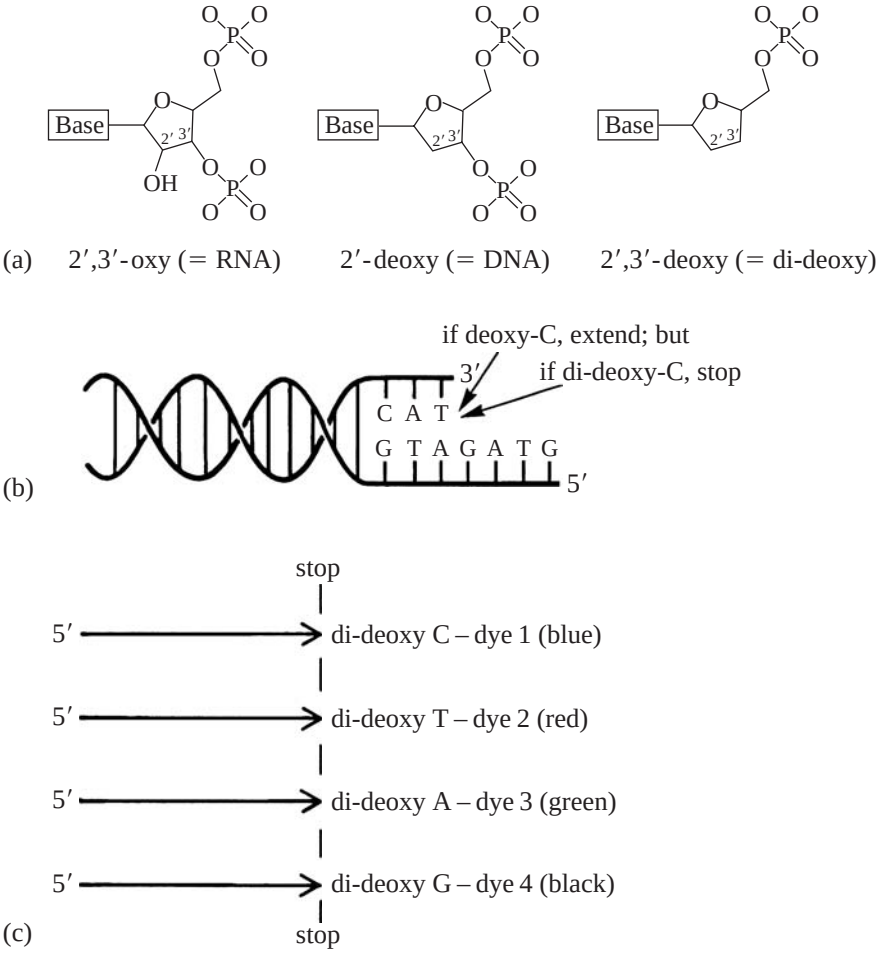

Figure 10.3 Key concepts from the di-deoxy, four-colour method of DNA sequencing: (a) comparison of three forms of nucleotide on an atomic scale; (b) comparison of chain extension by 2′-deoxy cytosine *versus* 2′,3′-di-deoxy cytosine under the action of DNA polymerase; the chain extends normally if 2′deoxy-C is added (and pairs with the guanine on the template), but stops or terminates if 2′,3′-di-deoxy-C is added; (c) fluorescent dyes of four different colours can be attached to the four different di-deoxy nucleotides C, T, A and G, so that any particular product of chain termination will show just one color by capillary electrophoresis.

particular DNA molecule it will occasionally, and at random, add a di-deoxy nucleotide; and the chain-building will abruptly terminate right there. This is shown schematically in Fig. 10.3(b): the polymerase will stop if a di-deoxy C happens to be added, but not if an ordinary, deoxy C is put there. Now in a large sample, we can rely on this random event – the addition of a di-deoxy nucleotide – occurring

at every possible stopping-point between 1 and 500, and hence a full set of truncated DNA molecules being created.

The only problem remaining is to identify the four kinds of di-deoxy nucleotide. This is done by chemically modifying these nucleotides with fluorescent dyes of different colour – typically rhodoamine or fluorescein derivatives – as shown in Fig. 10.3(c). These dyes are attached to the 7-position of A and G or to the 5-position of C and T, through long flexible linkers, so that these large and bulky compounds do not interfere with the ability of the polymerase to add each unit to the end of the terminated chain (see Fig. 11.1 for the numbering convention of the 7- and 5-positions). After a mixture of new chains has been made, the DNA-sequencing machine will separate them by size, using capillary electrophoresis. Separation of DNA chains by size in a capillary filled with some gel-like polymer, is precisely analogous to separating DNA fragments by size using a large polyacrylamide or agarose gel; but for a tiny capillary much higher voltages may be applied (without excess heating), which speeds the process greatly.

Finally, the DNA-sequencing machine uses a fluorescent light-source and detector, to look for the distinctive 'colour' which is associated with the last base of each chain , as it flows through the capillary under the influence of an electric field.

Some typical results of di-deoxy, four-colour DNA sequencing are shown in Fig. 10.4. Part (a) shows the partial sequence of a PCR-derived fragment. It reads TGGAA-GTCGT-TTGGC-TTGTG-GTGTA-ATTGG-G going left to right across the page. C (cytosine) bases are shown in blue; T (thymine) bases are shown in red; A (adenine) bases are shown in green; and G (guanine) bases are shown in black. Below each base is a curve which tells how strongly each di-deoxy-fluorescent signal was detected, when any particular DNA fragment proceeded through the capillary under the influence of an electric field.

Figure 10.4(b) shows a similar result for a PCR-derived fragment which is slightly different from that of (a). Here a single location 119 on the right-hand side has been read by the computer as 'N' to signify 'uncertain base'. Such a result could be due conceivably to experimental error. However, in this case we know that the PCR-derived fragment shown in (b) contains a mixture of two different bases at position 119, namely T (red) and G (black), owing to a genetic mutation on one of a patient's two homologous chromosomes. What a boon for medical science, to be able to detect single-copy (i.e. heterozygous) mutations in DNA which can cause disease, within a hospital laboratory in just a short period of time, from inconsequential amounts of a patient's blood!

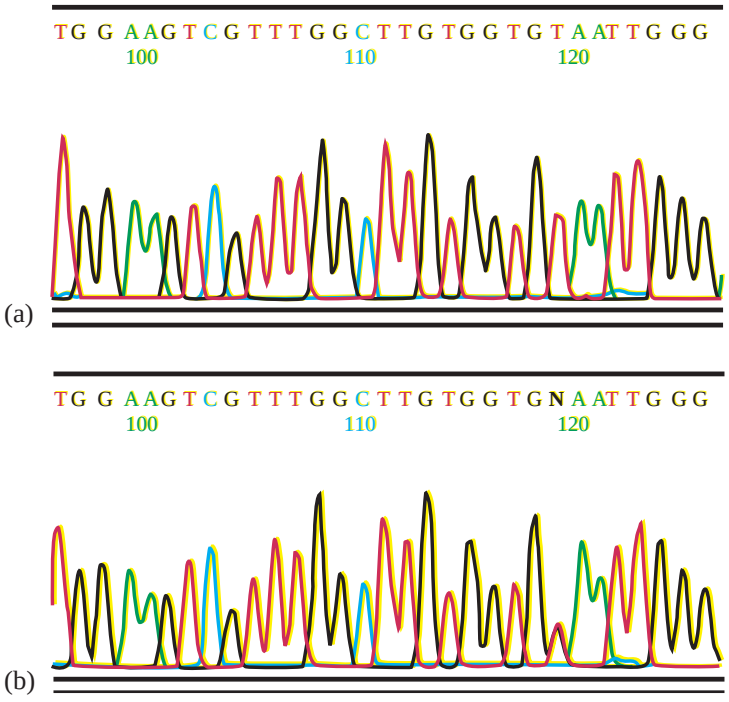

Figure 10.4 Two examples of di-deoxy, four-colour sequencing using PCR fragments as amplified from the DNA of two human patients. The sequencing plot shown in (a) comes from a healthy individual, whereas the sequencing plot shown in (b) comes from an individual affected with the neurological disease Amyotrophic Lateral Sclerosis or 'ALS'. A point-mutation, shown by 'N' on the right-hand side of (b) at position 119, changes 'T' on both homologous chromosomes to 'T' on one chromosome but 'G' on the other; and thereby changes amino-acid 148 in the protein 'superoxide dismutase' from valine to glycine, so as to produce this ailment. Courtesy of G. Nicholson.

In fact, the four-colour method for determining the base sequence of DNA is so powerful and fast, that it has enabled the full sequence determination of many different bacterial, plant and animal genomes, including human. Such data provide an excellent 'map' of the genome to facilitate genetic or medical research. Most of these fast-accumulating results (20 million bases per day at large facilities) can be accessed *via* the Internet at **www.ncbi.nlm.nih.gov**. Who would have thought, in 1960, that medical and biological scientists would spend much of their time accessing DNA sequences from a worldwide computer network? In the future, those extensive sequence data may provide for new cloned proteins as therapeutic compounds; already they have proven useful for the diagnosis of many kinds of disease.