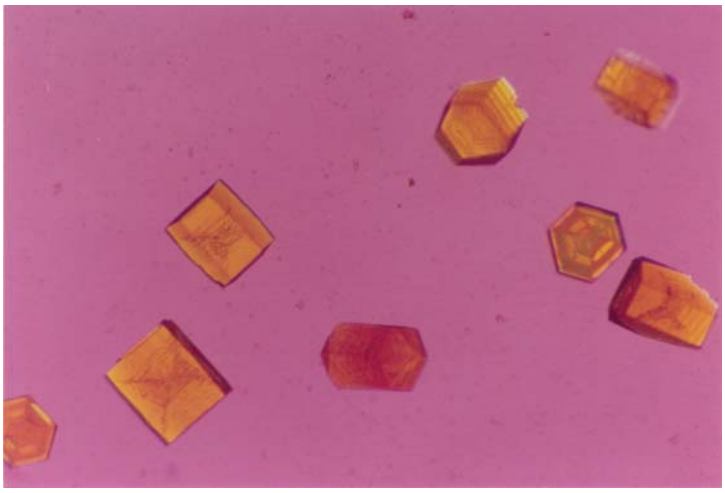

Figure 9.1 Crystals of DNA of sequence AGCATGCT in combination with the antibiotic nogalamycin. Each crystal is about 1 mm long. Courtesy of Maxine McCall and Louise Lockley.

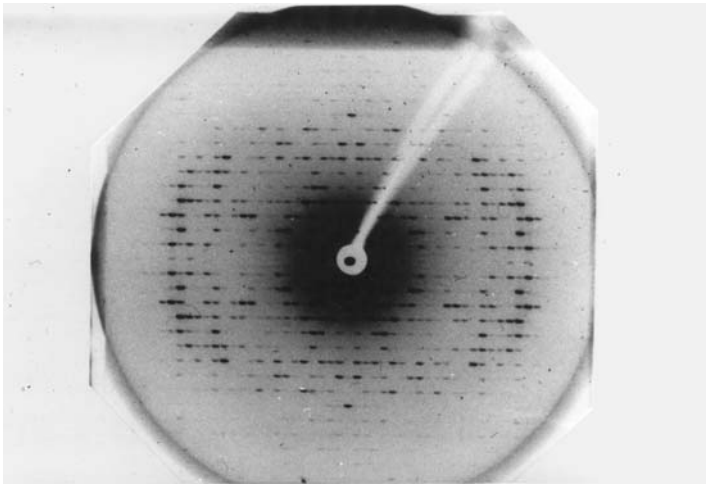

Figure 9.2 Typical X-ray diffraction photograph of a DNA crystal. The DNA molecule in the crystal which produced this picture contained 12 base-pairs of sequence CGCGAATTCGCG on each strand.

what sort of repeating array the molecules have formed. DNA can pack into a crystal having any one of 65 different kinds of three-dimensional symmetry; and so the first task is to determine which kind of symmetry is present in any particular crystal. The next job is to measure the relative intensities of the spots, for each of many photographs taken with different orientations of the crystal in the X-ray beam. In the early days this was a time-consuming task, but now it

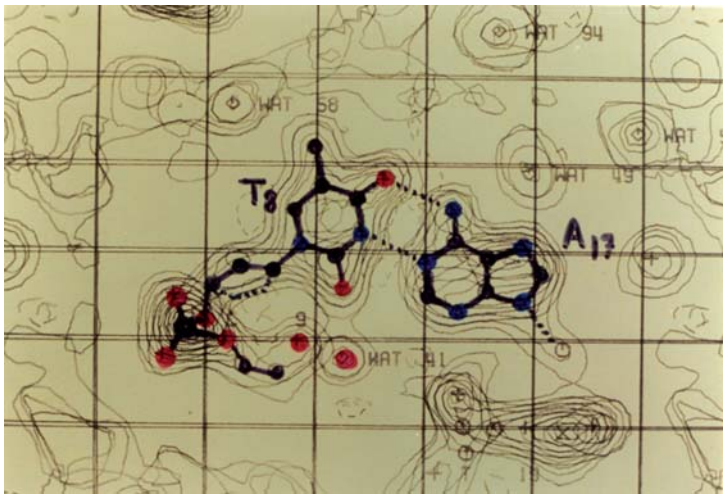

Figure 9.3 Assignment of atomic structure to part of an electron-density map as produced by X-ray diffraction methods. This part of the map shows an A–T base-pair as in Fig. 2.11(a). Water molecules are labeled 'WAT'. From Dickerson and Drew (1981) *Journal of Molecular Biology* **149**, 761–86.

can be done routinely in a short time by automated methods, and by use of a powerful X-ray beam. The final task, which is the most difficult, is to translate the relative intensities of spots into a model of the atomic structure, for DNA and any antibiotic or protein in the crystal. Each non-hydrogen atom (for example, carbon, nitrogen, oxygen, or phosphorus) can be located to an accuracy of about 0.1 to 0.2 Å in three-dimensional space if this last job is done properly; and so even the fine details of a structure can be found.

Figure 9.3 shows one small part of a completed DNA structure as determined in this way. There we can see an adenine–thymine base-pair, surrounded by many ordered water molecules. The locations of carbon, nitrogen, and oxygen atoms are identified by successive contours of increasing electron density. In fact, the X-ray scattering power of any atom is proportional to the square of its electron number; so carbon scatters X-rays as $(6)^2 = 36$, nitrogen as $(7)^2 = 49$, and oxygen as $(8)^2 = 64$. That is why hydrogen atoms cannot usually be located, because they scatter X-rays only weakly as $(1)^2 = 1$. And that is also why heavy atoms such as bromine, iodine, or platinum can be used to help solve X-ray structures, because platinum, for example, scatters as $(78)^2 = 6084$, or much more strongly than the other light atoms.

From the final assignment of atomic positions in a crystal, and after many cycles of refinement by a computer, one can obtain a highly accurate model of the whole DNA molecule (plus protein or antibiotic,

as appropriate) in three dimensions. Such three-dimensional models are usually regarded as being broadly representative of the structure in solution, on average: for if the structure in the crystal were too different from that in solution, the molecule would never have crystallized! These three-dimensional models form the whole underpinning for the science of molecular biology; and that is why we have explained how they are derived. In addition, the principles of symmetry and molecular structure which one learns during the course of an X-ray analysis are useful in understanding other, wider aspects of biology which do not have anything to do with crystals. For example, one cannot understand filaments of any sort, whether they take the form of flagella, muscles, microtubules, or DNA, without knowing something about symmetry. As J.D. Bernal once wrote, 'generalized crystallography is the key to molecular biology'.

We have not explained here how scientists actually convert the relative intensities of spots in a set of X-ray patterns to a detailed three-dimensional atomic structure. Our reason for this is that the mathematics used in the process are extremely difficult; all the student needs to know is that the final structure is built up from the superposition of many waves, and that the relative intensity of each spot on the film defines the height of one particular wave. One should also know that the mathematics are highly statistical in nature: each X-ray particle (or photon) behaves unpredictably when going through the crystal, and can emerge at any spot. Thus, from a single X-ray scattering event you cannot learn anything, but by averaging over many events, you can obtain a consistent probability of the photon's arriving at any spot, which is then proportional to its intensity.

In addition to X-ray diffraction methods for analysing atomic structure, one can also carry out *electron microscopy* experiments. Here one first lays a molecule of DNA onto a 'grid', and applies a heavy-metal stain such as uranium or platinum in order to help visualize the DNA in the electron beam. Then one surrounds the grid by a vacuum and shoots electrons through the sample. The electrons are focused, and clean pictures of DNA such as those shown in Figs 5.5 and 6.7 may be obtained. Electron microscopy experiments are easy to perform, but the pictures are lacking in atomic detail. Furthermore, the DNA – or whatever – can easily be distorted from its natural shape in the course of preparation for electron microscopy, since it must be removed from the fluid which normally surrounds it and be placed in a vacuum. The images produced by electron microscopy generally show the molecule of interest in just two dimensions, unless special care is taken to tilt the grid and shoot successive pictures from different perspectives.

Both X-ray structure analysis and electron microscopy are *direct* techniques for determining the structure of DNA. In the end, you can simply look at a three-dimensional model of DNA as determined by X-ray diffraction, or at a picture of DNA on a grid as determined by electron microscopy, and be confident that the thing you are looking at corresponds to physical reality. However, not all scientists practise these two methods, because: (a) the necessary equipment is expensive; (b) a scientist must be highly trained in order to carry out such analyses; and (c) it is often difficult to prepare a suitable crystal, or indeed a sample for electron microscopy, of a biologically interesting substance.

For those reasons, many scientists today use a variety of *indirect* methods for finding out about the structure of DNA. Most of the indirect methods are less reliable than the direct methods described above, but they are generally cheaper and simpler to perform. We shall explain about several different kinds of indirect method here. First there are the *spectral methods* such as nuclear magnetic resonance, Raman spectroscopy and circular dichroism. Then there are the *enzymatic methods* such as 'footprinting' with a DNA-cutting enzyme. Finally there are the *electrophoretic methods*, where DNA is passed through a gel in the presence of an electric field, and thereby separated according to its size, shape, and electric charge.

In the technique of *nuclear magnetic resonance* or NMR (which of late has become very expensive to perform, even more so than X-ray diffraction), a concentrated sample of DNA is brought into the presence of a strong magnetic field, so that the magnetic moments of all of the hydrogen atoms in the DNA align themselves with this major field. Next, the sample is exposed to a low-energy electromagnetic field over a range of radio frequencies; and individual protons within the nuclei of the hydrogen atoms of the DNA may absorb energy at some particular frequency, and thereby align their magnetic moments *against* the main field. The amount of energy required to flip the magnetic moment of a hydrogen atom against the main field is very sensitive to its location in the molecule, how it is chemically bonded to other atoms, and what atoms are located near it in three-dimensional space. Figure 9.4 shows part of the NMR spectrum for a 12-base-pair molecule of sequence CGCGAATTCGCG at several different temperatures. Because both ends of the molecule are related by symmetry (in other words, CGCGAA can pair to TTCGCG), there are only six peaks rather than 12 in the spectrum. Each of those peaks represents the magnetic alignment of a single kind of hydrogen atom in millions of identical DNA molecules. These particular hydrogens lie in the center of Watson–Crick base-pairs as N–H···O or N–H···N hydrogen bonds (recall Fig. 2.11). There are six peaks, at slightly

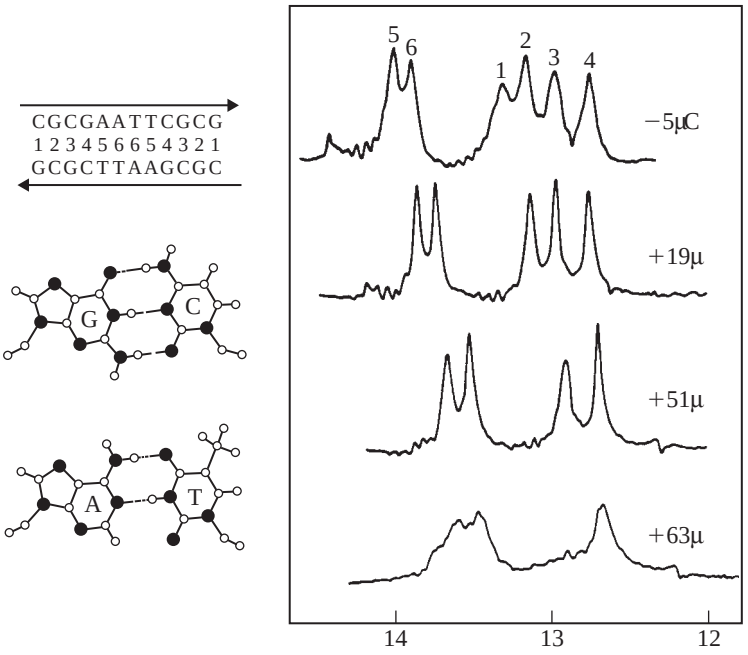

Figure 9.4 NMR spectra of a DNA molecule of sequence CGCGAATTCGCG, at four different temperatures. The arrows alongside the sequence show the (5′) to (3′) directions, and the horizontal scale under the spectra indicates radio frequency. Courtesy of Dinshaw Patel.

different frequencies of radio absorption, because there are six slightly different kinds of base-pair in different chemical environments. When the temperature of the sample is increased from −5°C to +51°C, the base-pairs on either end of the molecule (numbers 1 and 2) begin to fall apart: then the NMR peaks for their hydrogen atoms are lost, as they exchange with water. The main virtue of NMR methods is that they provide information about the dynamic structure of DNA in solution, which is not available from X-ray or electron microscopy studies. Many of the applications of NMR to DNA in solution were pioneered by Dinshaw Patel in the 1970s, and in recent years hundreds of scientists have entered the field.

In the 1980s, Kurt Wüthrich and colleagues applied a new technique in NMR spectroscopy, first developed by Richard Ernst, that allows you to measure the transfer of magnetic alignment from one hydrogen atom to any other in a chemical molecule, and then use that information to tell you something about its chemical or three-dimensional structure. The two hydrogens must be close together in space for such transfer of magnetism to occur. This technique allows one to measure approximately all of the interatomic distances between different hydrogens in biological molecules, and so

to determine their structure in solution. The method has worked well for proteins, but in practice has proved only qualitatively useful for DNA, because there are so few hydrogen atoms on the bases and sugar. But such 'transfer' methods are useful in studying complexes of DNA with antibiotics or with proteins, where they show which hydrogens on the DNA are close to which hydrogens on the antibiotic or protein. NMR methods are limited to molecules having no more than a few thousand atoms, because of the increasing complexity of the spectra for many atoms, and because very large molecules do not turn over (or tumble) rapidly enough in solution to produce a clean spectrum.

Raman spectroscopy measures the vibrational frequencies of individual bonds in the DNA, and hence it is a sensitive measure of chemical bonding and structure. *Circular dichroism spectroscopy* measures the absorption of polarized ultraviolet light by DNA, and shows whether the molecule absorbs more left-handed or right-handed polarized light. Both of these methods were used more in the past than today – for example, in 1972 by Pohl and Jovin to find evidence for left-handed DNA; so they will not be discussed further here.

A variety of *enzymes* and *chemicals* can be used for the analysis of DNA sequence and structure. Some of these will break the DNA into bits at certain short series of nucleotides such as GAATTC, or else at certain single nucleotides such as A, G, C, or T. Others will cut the DNA in practically any location, except where an antibiotic or protein has bound itself to the molecule. Still others will cut the DNA only in places where they detect an unwound single strand, instead of a double helix. What all of these methods have in common is that they use *electrophoresis in gels* to separate the fragments of DNA according to their size. So we must say something about the motion of DNA through gels, before we can explain how enzymatic or chemical methods can be used to probe the structure.

A gel is nothing more than a three-dimensional array of tiny, randomly oriented fibers, like the fibers in a grass mat that you wipe your feet on before going into the house. Most of the spaces in a gel are filled with water. For example, when you make 'jello' as a dessert at children's parties, you simply boil a small amount of gelatin powder in a large volume of water, and let it cool; then the final gel will be no more than 10% gelatin and 90% water. It is easy to see why a typical gel should be highly porous to small molecules such as DNA or protein: they can move easily through the gel by passing through the water spaces between the gel fibers.

Some small molecules can move only slowly through a gel by diffusion, if they are uncharged. But DNA and protein both carry a net electric charge, and so they can move quickly through a gel in the

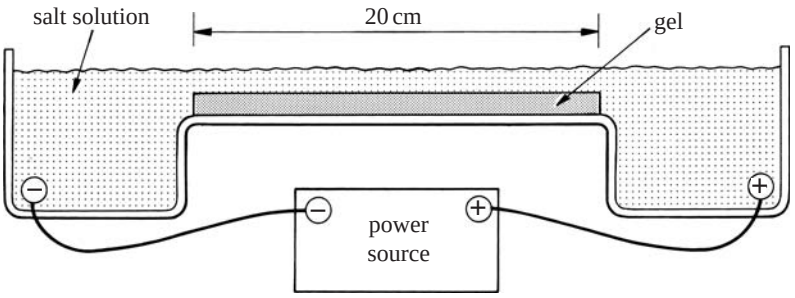

Figure 9.5 A typical gel-running apparatus. Other set-ups may be as long as 50 cm, and may be vertical as well as horizontal.

presence of an electric field. An ordinary gel as used in DNA work can be poured between two glass plates, or else as a thin slab onto a flat surface of size typically 20 cm × 20 cm, as shown in Fig. 9.5. (More recently, gel-like polymers have even been put inside narrow capillary tubes.) Positively and negatively charged electrodes can be placed at its ends in a suitable salt solution in order to impart the desired voltage gradient.

We need not concern ourselves yet with details of the gel-running experiment, such as how to choose the correct density of gel; or indeed how to describe the motion of DNA through a gel by use of mathematics. For present purposes, only two things really matter: one is that short molecules of DNA can travel through a gel more rapidly than long ones, and the other is that every kind of DNA molecule runs through the gel at a very definite, size-related speed.

Suppose that we have a pure sample of linear[1] DNA of some given length and sequence. If we load this sample into a small 'well' at one end of the gel, and turn on the voltage for a few hours, we find that the DNA migrates as a tight band towards the other end, without significant broadening or diffusion. There are two commonly used ways to locate the DNA in a gel. One is to stain the gel with a dye such as ethidium bromide (see Chapter 2), which fluoresces strongly under ultraviolet light when it is bound between two base-pairs. The other is to incorporate one or more radioactive phosphorus atoms into the DNA at its 5′- or 3′-end, or perhaps throughout the length of the molecule. Then the radioactive band of DNA will darken an ordinary photographic film after only a few hours. A third, less commonly used way to locate either DNA or protein in a gel, is to stain the gel with silver metal: then any DNA or protein within the gel binds to the silver metal and so turns the gel brown locally.

Suppose next that we take the same sample of DNA, but treat it with a 'restriction enzyme' called 'Eco RI', that cuts wherever it can find the particular sequence of nucleotides GAATTC. If there are n

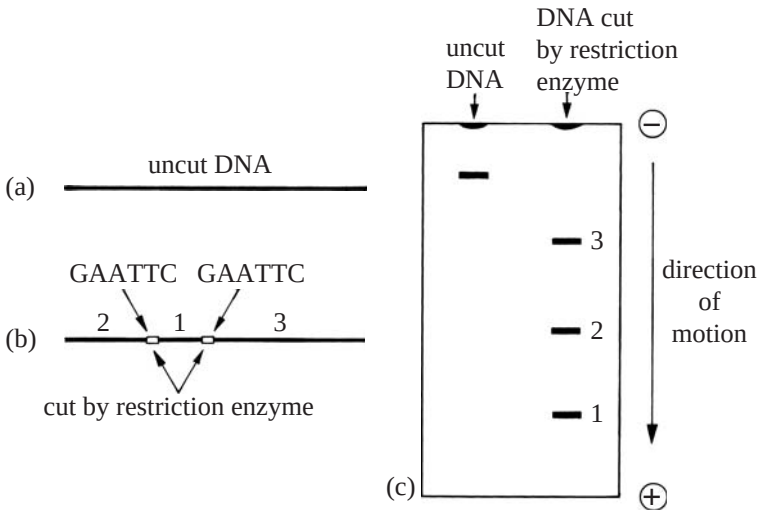

Figure 9.6 Running of DNA fragments through a gel, before and after cleavage with a restriction enzyme.

sequences of the kind GAATTC along its length, our DNA sample will now run through the gel as a series of $n + 1$ distinct bands, as shown schematically in Fig. 9.6. This scheme enables us to find out how many sequences GAATTC are contained in the DNA, and something about their location, since the smaller fragments will run further down the gel than the larger ones.

Finally, suppose we treat the same double-stranded DNA with some chemical that cuts only at the nucleotide A on any strand. Then if the products of this chemical reaction are run through a 'denaturing' gel that contains a high concentration of urea in order to separate the two strands, they will produce a 'ladder' of single-stranded fragments that mark the relative locations of all bases A in the DNA sequence, as shown on the left-hand side of Fig. 9.7. In order to locate bases G, C, and T, one can use other chemical reactions that are specific for these bases, and run these DNA fragments through the gel as well (Fig. 9.7). You can select which strand of the double helix you want to see in the gel photograph, by attaching a radioactive phosphorus to either one strand or the other; fragments from the 'other' strand will be non-radioactive and therefore invisible.

Another commonly-used method is to attach fluorescent dyes of four different colors to the four sequencing reactions A, G, C, and T. A dye will be attached only at one end of the molecule, and to terminal bases which may be distinguished by their distinct colours. All four kinds of reaction for any sequence can then be run in the same gel lane – or in the same gel-filled capillary tube – so as to improve

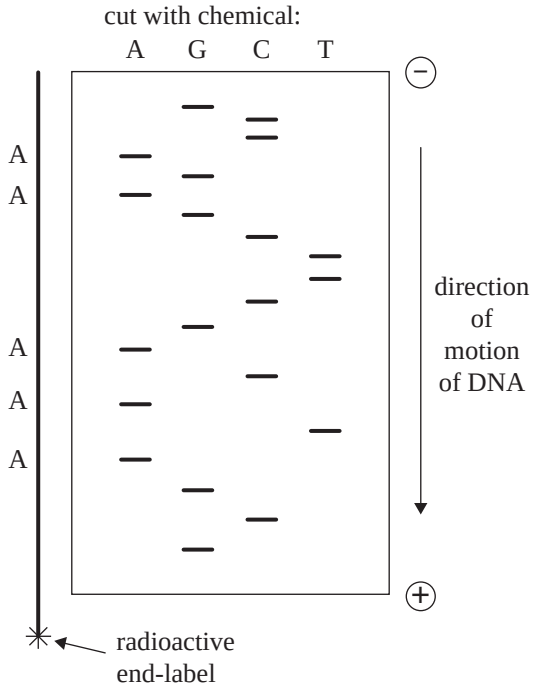

Figure 9.7 A gel for determining the sequence of a DNA molecule, which is shown in part on the left.

efficiency. We shall describe this four-colour sequencing method further in Chapter 10.

Modern sequencing methods allow one to determine the complete sequence of a DNA molecule as long as 800 to 1000 nucleotides, because single strands of length 50 to 800 nucleotides will run at slightly different speeds through a gel, or through a polymer-filled capillary, for most sequencing analyses that are done today. Strands of length greater than about 800 nucleotides tend to run at more identical speeds, and so individual lengths cannot easily be resolved from one another (say 800 from 799 or 801). In order to determine the complete sequence of a very long DNA, say from a chromosome or a virus, you have to break it into many different pieces of size about 800 base-pairs or less, and then sequence the pieces individually. By another commonly used method, one can start reading a sequence reaction on long DNA at roughly 800-base-pair intervals or less; in that case, a polymerase enzyme is used to build up many DNA sequencing strands for application to a gel or capillary. The strand to be sequenced is recognized specifically through Watson–Crick base-pairing to a series of small oligonucleotides or 'primers'.

These are not the only ways of determining the sequence of a long DNA molecule, but they are representative of the other methods.

The techniques described here have become so routine that scientists today have already completed large-scale sequencing for the complete genomes of many different animals, plants or micro-organisms – including the well-known Human Genome Project. The medical and scientific implications of these data will be discussed in Chapter 10.

Most of the enzymes and chemicals discussed so far can cut the DNA in very precise locations, according to its base sequence. Certain other enzymes and chemicals, for example DNAase I, can cut the DNA in practically any location, with only a mild specificity for the base sequence. How might such a generalized DNA-cutting activity be useful?

Suppose we have isolated from an animal or plant some important protein that affects gene activity, by binding to an unknown DNA sequence along the length of a chromosome. How can we determine where it prefers to bind to the DNA? Usually the protein will bind so tightly to the DNA that it blocks the cutting activity of an enzyme such as DNAase I; so we can locate the bound protein by looking to see where the cutting activity of DNAase I is reduced in the presence of protein. This technique is known as 'footprinting', because the regions of reduced cutting by DNAase I look like 'footprints' of the protein on the DNA when we study a gel photograph.

One example of such an experiment is shown in Fig. 9.8. There we are looking to see where a small antibiotic called 'echinomycin' binds along the DNA. Our detailed procedure is as follows: we label a DNA molecule of 200 base-pairs at either of its two 3′-ends with radioactive phosphorus atoms in separate experiments; then we add the antibiotic to each of these DNA preparations, and wait for a few minutes until the antibiotic has located its preferred binding sites; finally, we add DNAase I for a certain length of time, until some cutting has taken place at every nucleotide. When the two kinds of DNA sample are run on a urea-containing gel, we obtain the patterns shown in Fig. 9.8.

The left- and right-hand sides of this figure show the results of DNAase I cutting along either of the two strands of the double helix. The first three gel lanes on either side show how DNAase I cuts the free DNA, in the absence of echinomycin. The bands there show evidence for some cutting at every nucleotide; yet these bands are of greatly varying intensity. It seems that DNAase I prefers to cut more at some base sequences than at others. Many studies have shown that DNAase I binds across the minor groove of DNA, and only cuts well if this groove is of a correct size, and if the bond to be cut is positioned properly relative to the active site of the enzyme. Each of these structural features depends on the base sequence of the DNA, and so we see a rather complex pattern of cutting even in the absence of the antibiotic.