

Introduction and Overview

CONTENTS

1.1	The Biological Importance of DNA	1
1.2	The Origins of Nucleic Acids Research	2
1.3	Early Structural Studies on Nucleic Acids	2
1.4	The Discovery of the Structure of DNA	4
1.5	The Advent of Molecular Biology	7
1.6	The Partnership of Chemistry and Biology	8
1.7	Frontiers in Nucleic Acids Research	10
	References	11

1.1 THE BIOLOGICAL IMPORTANCE OF DNA

From the beginning, the study of nucleic acids has drawn together, as though by a powerful unseen force, a galaxy of scientists of the highest ability.^{1,2} Striving to tease apart its secrets, these talented individuals have brought with them a broad range of skills from other disciplines while many of the problems they have encountered have proved to be soluble only by new inventions. Looking at their work, one is constantly made aware that scientists in this field appear to have enjoyed a greater sense of excitement in their work than is given to most. Why?

For over 60 years, such men and women have been fascinated and stimulated by their awareness that the study of nucleic acids is central to the knowledge of life. Let us start by looking at Fred Griffith, who was employed as a scientific civil servant in the British Ministry of Health investigating the nature of epidemics. In 1923, he was able to identify the difference between a virulent, *S*, and a non-virulent, *R*, form of the pneumonia bacterium. Griffith went on to show that this bacterium could be made to undergo a permanent, heritable change from non-virulent to virulent type. This discovery was a bombshell in bacterial genetics.

Oswald Avery and his group at the Rockefeller Institute in New York set out to identify the molecular mechanism responsible for the change Griffith had discovered, now technically called **bacterial transformation**. They achieved a breakthrough in 1940 when they found that non-virulent *R* pneumococci could be transformed *irreversibly* into a virulent species by treatment with a pure sample of high molecular weight DNA.³ Avery had purified this DNA from heat-killed bacteria of a virulent strain and showed that it was active at a dilution of 1 part in 10⁹.

Avery concluded that '*DNA is responsible for the transforming activity*' and published that analysis in 1944, just 3 years after Griffith had died in a London air-raid. The staggering implications of Avery's work

turned a searchlight on the molecular nature of nucleic acids and it soon became evident that ideas on the chemistry of nucleic acid structure at that time were wholly inadequate to explain such a momentous discovery. As a result, a new wave of scientists directed their attention to DNA and discovered that large parts of the accepted tenets of nucleic acid chemistry had to be set aside before real progress was possible. We need to examine some of the earliest features of that chemistry to fully appreciate the significance of later progress.

1.2 THE ORIGINS OF NUCLEIC ACIDS RESEARCH

Friedrich Miescher started his research career in Tübingen by looking into the physiology of human lymph cells. In 1868, seeking a more readily available material, he began to study human pus cells, which he obtained in abundant supply from the bandages discarded from the local hospital. After defatting the cells with alcohol, he incubated them with a crude preparation of pepsin from pig stomach and so obtained a grey precipitate of pure cell nuclei. Treatment of this with alkali followed by acid gave Miescher a precipitate of a phosphorus-containing substance, which he named **nuclein**. He later found this material to be a common constituent of yeast, kidney, liver, testicular and nucleated red blood cells.⁴

After Miescher moved to Basel in 1872, he found the sperm of Rhine salmon to be a more plentiful source of nuclein. The pure nuclein was a strongly acidic substance, which existed in a salt-like combination with a nitrogenous base that Miescher crystallized and called protamine. In fact, his nuclein was really a nucleoprotein and it fell subsequently to Richard Altman in 1889 to obtain the first protein-free material, to which he gave the name **nucleic acid**.

Following William Perkin's invention of mauveine in 1856, the development of aniline dyes had stimulated a systematic study of the colour-staining of biological specimens. Cell nuclei were characteristically stained by basic dyes, and around 1880, Walter Flemming applied that property in his study of the rod-like segments of chromatin (called so because of their colour-staining characteristic), which became visible within the cell nucleus only at certain stages of cell division. Flemming's speculation that the chemical composition of these **chromosomes** was identical to that of Miescher's nuclein was confirmed in 1900 by E.B. Wilson who wrote

Now chromatin is known to be closely similar to, if not identical with, a substance known as nuclein which analysis shows to be a tolerably definite chemical compound of nucleic acid and albumin. And thus we reach the remarkable conclusion that inheritance may, perhaps, be affected by the physical transmission of a particular compound from parent to offspring.

While this insight was later to be realized in Griffith's 1928 experiments, all of this work was really far ahead of its time. We have to recognize that, at the turn of the century, tests for the purity and identity of substances were relatively primitive. Emil Fischer's classic studies on the chemistry of high molecular weight, polymeric organic molecules were in question until well into the twentieth century. Even in 1920, it was possible to argue that there were only two species of nucleic acids in nature: animal cells were believed to provide **thymus nucleic acid** (DNA), while nuclei of plant cells were thought to give **pentose nucleic acid** (RNA).

1.3 EARLY STRUCTURAL STUDIES ON NUCLEIC ACIDS

Accurate molecular studies on nucleic acids essentially date back to 1909 when Levene and Jacobs began a reinvestigation of the structure of **nucleotides** at the Rockefeller Institute. Inosinic acid, which Liebig had isolated from beef muscle in 1847, proved to be hypoxanthine-ribose 5'-phosphate. Guanylic acid, isolated from the nucleoprotein of pancreas glands, was identified as guanine-ribose 5'-phosphate (Figure 1.1). Each of these nucleotides was cleaved by alkaline hydrolysis to give phosphate and the corresponding **nucleosides**, inosine and guanosine, respectively. Since then, all nucleosides are characterized as the condensation products of a pentose and a nitrogenous base while nucleotides are the phosphate esters of one of the hydroxyl groups of the pentose.

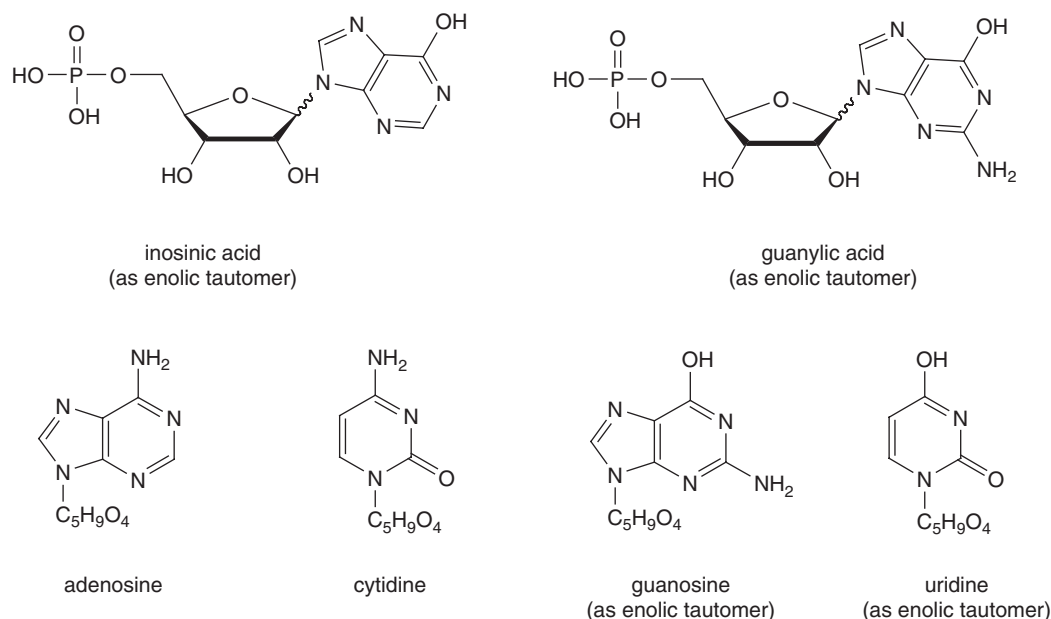


Figure 1.1 Early nucleosides and nucleotide structures (using the enolic tautomers originally employed). Wavy lines denote unknown stereochemistry at C-1'

Thymus nucleic acid, which was readily available from calf tissue, was found to be resistant to alkaline hydrolysis. It was only successfully degraded into deoxynucleosides in 1929 when Levene adopted enzymes to hydrolyse the deoxyribonucleic acid followed by mild acidic hydrolysis of the deoxynucleotides. He identified its pentose as the hitherto unknown 2-deoxy-D-ribose. These deoxynucleosides involved the four heterocyclic bases, adenine, cytosine, guanine and thymine, with the latter corresponding to uracil in ribonucleic acid.

Up to 1940, most groups of workers were convinced that hydrolysis of nucleic acids gave the appropriate four bases in **equal relative proportions**. This erroneous conclusion probably resulted from the use of impure nucleic acid or from the use of analytical methods of inadequate accuracy and reliability. It led, naturally enough, to the general acceptance of a **tetranucleotide hypothesis** for the structure of both thymus and yeast nucleic acids, which materially retarded further progress on the molecular structure of nucleic acids.

Several of these tetranucleotide structures were proposed. They all had four nucleosides (one for each of the bases) with an arbitrary location of the two purines and two pyrimidines. They were joined together by four phosphate residues in a variety of ways, among which there was a strong preference for phosphodiester linkages. In 1932, Takahashi showed that yeast nucleic acid contained neither pyrophosphate nor phosphomonoester functions and so disposed of earlier proposals in preference for a neat, cyclic structure which joined the pentoses exclusively using phosphodiester units (Figure 1.2). It was generally accepted that these bonded 5'- to 3'-positions of adjacent deoxyribonucleosides, but the linkage positions in ribonucleic acid were not known.

One property stuck out like a sore thumb from this picture: the molecular mass of nucleic acids was greatly in excess of that calculated for a tetranucleotide. The best DNA samples were produced by Einar Hammarsten in Stockholm and one of his students, Torbjörn Caspersson, who showed that this material was greater in size than protein molecules. Hammarsten's DNA was examined by Rudolf Signer in Bern whose flow-birefringence studies revealed rod-like molecules with a molecular mass of $0.5\text{--}1.0 \times 10^6$ Da. The same material provided Astbury in Leeds with X-ray fibre diffraction measurements that supported Signer's conclusion. Finally, Levene estimated the molecular mass of native DNA to be between 200,000 and 1×10^6 Da, based on ultracentrifugation studies.

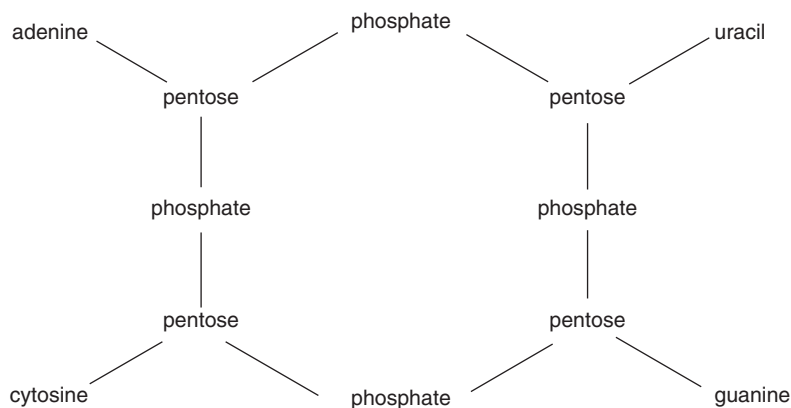


Figure 1.2 *The tetranucleotide structure proposed for nucleic acids by Takahashi (1932)*

The scientists compromised. In his Tilden Lecture of 1943, Masson Gulland suggested that the concept of nucleic acid structures of polymerized, uniform tetranucleotides was limited, but he allowed that they could ‘form a practical working hypothesis’.

This then was the position in 1944 when Avery published his great work on the transforming activity of bacterial DNA. One can sympathize with Avery’s hesitance to press home his case. Levene, in the same Institute, and others were strongly persuaded that the tetranucleotide hypothesis imposed an invariance on the structure of nucleic acids, which denied them any role in biological diversity. In contrast, Avery’s work showed that DNA was responsible for completely transforming the behaviour of bacteria. It demanded a fresh look at the structure of nucleic acids.

1.4 THE DISCOVERY OF THE STRUCTURE OF DNA

From the outset, it was evident that DNA exhibited greater resistance to selective chemical hydrolysis than did RNA. So, the discovery in 1935 that DNA could be cut into **mononucleotides** by an enzyme doped with arsenate was invaluable. Using this procedure, Klein and Thannhauser obtained the four crystalline deoxyribonucleotides, whose structures (Figure 1.3) were later put beyond doubt by total chemical synthesis by Alexander Todd⁵ and the Cambridge school he founded in 1944. Todd established the D-configuration and the glycosylic linkage for ribonucleosides in 1951, but found the chemical synthesis of the 2′-deoxyribonucleosides more taxing. The key to success for the Cambridge group was the development of methods of phosphorylation, for example for the preparation of the 3′- and 5′-phosphates of deoxyadenosine⁶ (Figure 1.4).

All the facts were now available to establish the primary structure of DNA as a **linear polynucleotide** in which each deoxyribonucleoside is linked to the next by means of a 3′- to 5′-phosphate diester (see Figure 2.15). The presence of only diester linkages was essential to explain the stability of DNA to chemical hydrolysis, since phosphate triesters and monoesters, not to mention pyrophosphates, are more labile. The measured molecular masses for DNA of about 1×10^6 Da meant that a single strand of DNA would have some 3000 nucleotides. Such a size was much greater than that of enzyme molecules, but entirely compatible with Staudinger’s established ideas on macromolecular structure for synthetic and natural polymers. But by the mid-twentieth century, chemists could advance no further with the primary structure of DNA. Neither of the key requirements for sequence determination was to hand: there were no methods for obtaining pure samples of DNA with homogeneous base sequence nor were methods available for the cleavage of DNA strands at a specific base residue. Consequently, all attention came to focus on the secondary structure of DNA.

Two independent experiments in biophysics showed that DNA possesses an ordered secondary structure. Using a sample of DNA obtained from Hammarsten in 1938, Astbury obtained an X-ray diffraction pattern

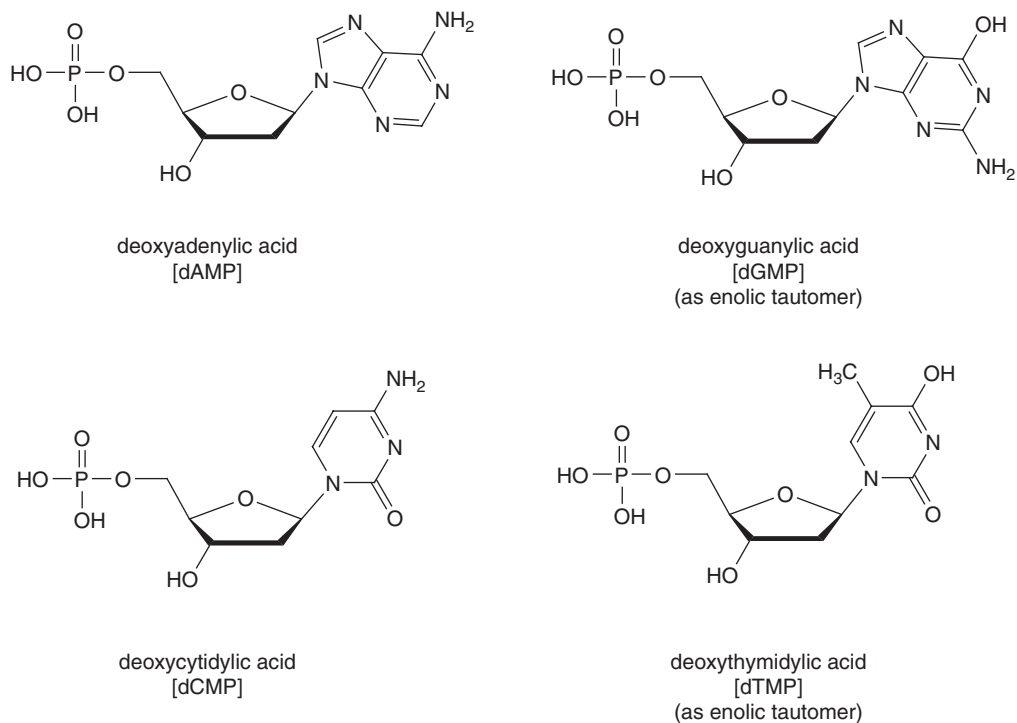


Figure 1.3 Structures of 5'-deoxyribonucleotides (original tautomers for dGMP and dTMP)

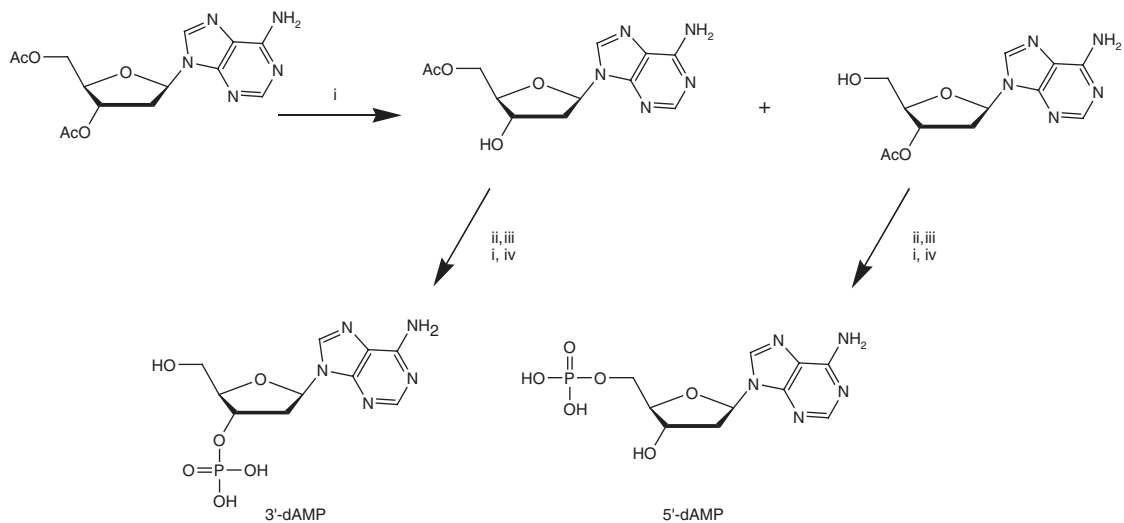


Figure 1.4 Todd's synthesis of deoxyadenosine 3'- and 5'-phosphates Reagents: (i) MeOH, NH₃ (ii) (PhO)₂P(O)OP(H)(O)OCH₂Ph (iii) N-chlorosuccinimide (iv) H₂/PdC (D.H. Hayes, A.M. Michelson and A.R. Todd, *J. Chem. Soc.*, 1955, 808–815)

from stretched, dry fibres of DNA. From the rather obscure data he deduced ‘... A spacing of 3.34 Å along the fibre axis corresponds to that of a close succession of flat or flattish nucleotides standing out perpendicularly to the long axis of the molecule to form a relatively rigid structure.’ These conclusions roundly contradicted the tetranucleotide hypothesis.

Some years later, Gulland studied the viscosity and flow-birefringence of calf thymus DNA and thence postulated the presence of hydrogen bonds linking the purine–pyrimidine **hydroxyl** groups and some of the amino groups. He suggested that these hydrogen bonds could involve nucleotides either in adjacent chains or within a single chain, but he somewhat hedged his bets between these alternatives. Sadly, Astbury returned to the investigation of proteins and Gulland died prematurely in a train derailment in 1947. Both of them left work that was vital for their successors to follow, but each contribution contained a misconception that was to prove a stumbling block for the next half-a-dozen years. Thus, Linus Pauling’s attempt to create a helical model for DNA located the pentose-phosphate backbone in its core and the **bases pointing outwards** – as Astbury had decided. Gulland had subscribed to the wrong tautomeric forms for the heterocyclic bases thymine and guanine, believing them to be **enolic** and having hydroxyl groups. The importance of the true **keto forms** was only appreciated in 1952.

Erwin Chargaff began to investigate a very different type of order in DNA structure. He studied the base composition of DNA from a variety of sources using the new technique of paper chromatography to separate the products of hydrolysis of DNA and employing one of the first commercial ultraviolet spectrophotometers to quantify their relative abundance.⁷ His data showed that there is a variation in base composition of DNA between species that is overridden by a universal 1:1 ratio of adenine with thymine and guanine with cytosine. This meant that the proportion of purines, (A + G), is always equal to the proportion of pyrimidines, (C + T). Although the ratio (G + C)/(A + T) varies from species to species, different tissues from a single species give DNA of the same composition. Chargaff’s results finally discredited the tetranucleotide hypothesis, because it called for equal proportions of all four bases in DNA.

In 1951, Francis Crick and Jim Watson joined forces in the Cavendish Laboratory in Cambridge to tackle the problem of DNA structure. Both of them were persuaded that the model-building approach that had led Pauling and Corey to the α -helix structure for peptides should work just as well for DNA. Almost incredibly, they attempted no other line of direct experimentation but drew on the published and unpublished results of other research teams in order to construct a variety of models, each to be discarded in favour of the next until they created one which satisfied all the facts.^{8,9}

The best X-ray diffraction results were to be found in King’s College, London. There, Maurice Wilkins had observed the importance of keeping DNA fibres in a moist state and Rosalind Franklin had found that the X-ray diffraction pattern obtained from such fibres showed the existence of an A-form of DNA at low humidity, which changed into a B-form at high humidity. Both forms of DNA were highly crystalline and clearly helical in structure. Consequently, Franklin decided that this behaviour required the phosphate groups to be exposed to water on the **outside** of the helix, with the corollary that the bases were on the **inside** of the helix.

Watson decided that the number of nucleotides in the unit crystallographic cell favoured a double-stranded helix. Crick’s physics-trained mind recognized the symmetry implications of the space-group of the A-form diffraction pattern, monoclinic $C2$. There had to be local twofold symmetry axes normal to the helix, a feature, which called for a double-stranded helix, whose two chains must run in opposite directions.

Crick and Watson thus needed merely to solve the final problem: how to construct the core of the helix by packing the bases together in a regular structure. Watson knew about Gulland’s conclusions regarding hydrogen bonds joining the DNA bases. This convinced him that the crux of the matter had to be a rule governing hydrogen bonding between bases. Accordingly, Watson experimented with models using the **enolic** tautomeric forms of the bases (Figure 1.3) and pairing like with like. This structure was quickly rejected by Crick because it had the wrong symmetry for B-DNA. **Self-pairing** had to be rejected because it could not explain Chargaff’s 1:1 base ratios, which Crick had perceived were bound to result if you had **complementary base pairing**.

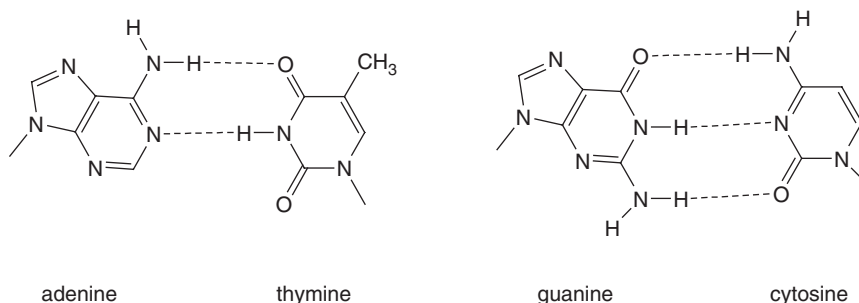


Figure 1.5 Complementary hydrogen-bonded base-pairs as proposed by Watson and Crick (thymine and guanine in the revised keto forms). The G...C structure was later altered to include three hydrogen bonds

On the basis of the advice from Jerry Donohue in the Cavendish Laboratory, Watson turned to manipulating models of the bases in their **keto forms** and paired adenine with thymine and guanine with cytosine. Almost at once, he found a compellingly simple relationship involving two hydrogen bonds for an A...T pair and two or three hydrogen bonds for a G...C pair. The special feature of this base-pairing scheme is that the relative geometry of the bonds joining the bases to the pentoses is virtually identical for the A...T and G...C pairs (Figure 1.5). It follows that if a purine always pairs with a pyrimidine then an irregular sequence of bases in a single strand of DNA could nevertheless be paired regularly in the centre of a double helix and without loss of symmetry.¹⁰

Chargaff's 'rules' were straightaway revealed as an obligatory consequence of a double-helical structure for DNA. Above all, since the base sequence of one chain automatically determines that of its partner, Crick and Watson could easily visualize how one single chain might be the template for creation of a second chain of complementary base sequence.

The structure of the core of DNA had been solved and the whole enterprise fittingly received the ultimate accolade of the scientific establishment when Crick, Watson and Wilkins shared the Nobel prize for chemistry in 1962, just 4 years after Rosalind Franklin's early death.

1.5 THE ADVENT OF MOLECULAR BIOLOGY

It is common to describe the publication of Watson and Crick's paper in *Nature* in April 1953 as the end of the 'classical' period in the study of nucleic acids, up to which time basic discoveries were made by a few gifted academics in an otherwise relatively unexplored field. The excitement aroused by the model of the double helix drew the attention of a much wider scientific audience to the importance of nucleic acids, particularly because of the biological implications of the model rather than because of the structure itself. It was immediately apparent that locked into the irregular sequence of nucleotide bases in the DNA of a cell was all the information required to specify the diversity of biological molecules needed to carry out the functions of that cell. The important question now was what was the key, the **genetic code**, through which the sequence of DNA could be translated into protein?¹¹

The solution to the coding problem is often attributed to the laboratories in the USA of Marshall Nirenberg and of Severo Ochoa who devised an elegant cell-free system for translating enzymatically synthesized polynucleotides into polypeptides and who by the mid-1960s had established the genetic code for a number of amino acids.^{12,13} In reality, the story of the elucidation of the code involves numerous strands of knowledge obtained from a variety of workers in different laboratories. An essential contribution came from Alexander Dounce in Rochester, New York, who in the early 1950s postulated that RNA, and not DNA, served as a template to direct the synthesis of cellular proteins and that a sequence of three nucleotides might specify a single amino acid. Sydney Brenner and Leslie Barnett in Cambridge, later (1961) confirmed the code to be both triplet and non-overlapping. From Robert Holley in Cornell University, New York, and

Hans Zachau in Cologne, came the isolation and determination of the sequence of three **transfer RNAs** (tRNA) ‘adapter’ molecules that each carry an individual amino acid ready for incorporation into protein and which are also responsible for recognizing the triplet code on the **messenger RNA** (mRNA). The mRNA species contain the sequences of individual genes copied from DNA (see Chapter 6). Gobind Khorana and his group in Madison, Wisconsin, chemically synthesized all 64 ribotrinucleoside diphosphates and, using a combination of chemistry and enzymology, synthesized a number of polyribonucleotides with repeating di-, tri-, and tetranucleotide sequences.¹⁴ These were used as synthetic mRNA to help identify each triplet in the code. This work was recognized by awarding the Nobel Prize for Medicine in 1968 jointly to Holley, Khorana and Nirenberg.

Nucleic acid research in the 1950s and 1960s was preoccupied by the solution to the coding problem and the establishment of the biological roles of tRNA and mRNA. This was not surprising bearing in mind that at that time the smaller size and attainable homogeneity made isolation and purification of RNA a much easier task than it was for DNA. It was clear that in order to approach the fundamental question of what constituted a **gene** – a single heritable element of DNA that up to then could be defined genetically but not chemically – it was going to be necessary to break down DNA into smaller, more tractable pieces in a specific and predictable way.

The breakthrough came in 1968 when Meselson and Yuan reported the isolation of a **restriction enzyme** from the bacterium *Escherichia coli*. Here at last was an enzyme, a nuclease, which could recognize a defined sequence in a DNA and cut it specifically (see Section 5.3.1). The bacterium used this activity to break down and hence inactivate invading (*e.g.* phage) DNA. It was soon realized that this was a general property of bacteria, and the isolation of other restriction enzymes with different specificities soon followed. But it was not until 1973 that the importance of these enzymes became apparent. At this time, Chang and Cohen at Stanford and Helling and Boyer at the University of California were able to construct in a test tube, a biologically functional DNA that combined genetic information from two different sources. This **chimera** was created by cleaving DNA from one source with a restriction enzyme to give a fragment that could then be joined to a carrier DNA, a **plasmid**. The resultant **recombinant DNA** was shown to be able to replicate and express itself in *E. coli*.¹⁵

This remarkable demonstration of genetic manipulation was to revolutionize biology. It soon became possible to dissect out an individual gene from its source DNA, to amplify it in a bacterium or other organism (**cloning**, see Section 5.2), and to study its expression by the synthesis first of RNA and then of protein (see Chapters 6 and 7). This single advance by the groups of Cohen and Boyer truly marked the dawn of modern molecular biology.

1.6 THE PARTNERSHIP OF CHEMISTRY AND BIOLOGY

In the 1940s and 1950s, the disciplines of chemistry and biology were so separate that it was a rare occurrence for an individual to embrace both. Two young scientists who were just setting out on their careers at that time were exceptional in recognizing the potential of chemistry in the solution of biological problems and both, in their different ways, were to have a substantial and lasting effect in the field of nucleic acids.

One was Frederick Sanger, a product of the Cambridge Biochemistry School, who in the early 1940s set out to determine the sequence of a protein, insulin. This feat had been thought unattainable, since it was widely supposed that proteins were not discrete species with defined primary sequence. Even more remarkably, he went on to develop methods for sequence determination first of RNA and then of DNA (see Section 5.1). These methods involved a subtle blend of enzymology and chemistry that few would have thought possible to combine.¹⁶ The results of his efforts transformed **DNA sequencing** in only a few years into a routine procedure. In the late 1980s, the procedure was adapted for use in automated sequencing machines and the 1990s saw worldwide efforts to sequence whole organism genomes. In 2003, exactly 50 years after the discovery of the structure of the DNA double helix, it was announced that the human genome sequence had been completed. The award of two Nobel prizes to Sanger (1958 and 1980) hardly seems recognition enough!

The other scientist has already been mentioned in connection with the elucidation of the genetic code. Not long after his post-doctoral studies under George Kenner and Alexander Todd in Cambridge, Gobind Khorana was convinced that chemical synthesis of polynucleotides could make an important contribution to the study of the fundamental process of information flow from DNA to RNA to protein. Having completed the work on the genetic code in the mid-1960s and aware of Holley's recently determined (1965) sequence for an alanine tRNA, he then established a new goal of total synthesis of the corresponding DNA duplex, the gene specifying the tRNA. Like Sanger, he ingeniously devised a combination of nucleic acid chemistry and enzymology to form a general strategy of **gene synthesis**, which in principle remains unaltered to this day (see Section 5.4).¹⁷ Knowledge became available by the early 1970s about the signals required for gene expression and the newly emerging recombinant DNA methods of Cohen and Boyer allowed a second synthetic gene, this time specifying the precursor of a tyrosine suppressor tRNA (Figure 1.6) to be cloned and shown to be fully functional.

It is ironic that even up to the early 1970s many biologists thought Khorana's gene syntheses unlikely to have practical value. This view changed dramatically in 1977 with the demonstration by the groups of Itakura (a chemist) and Boyer (a biologist) of the expression in a bacterium of the hormone somatostatin (and later insulin A and B chains) from a chemically synthesised gene.¹⁸ This work spawned the biotechnology

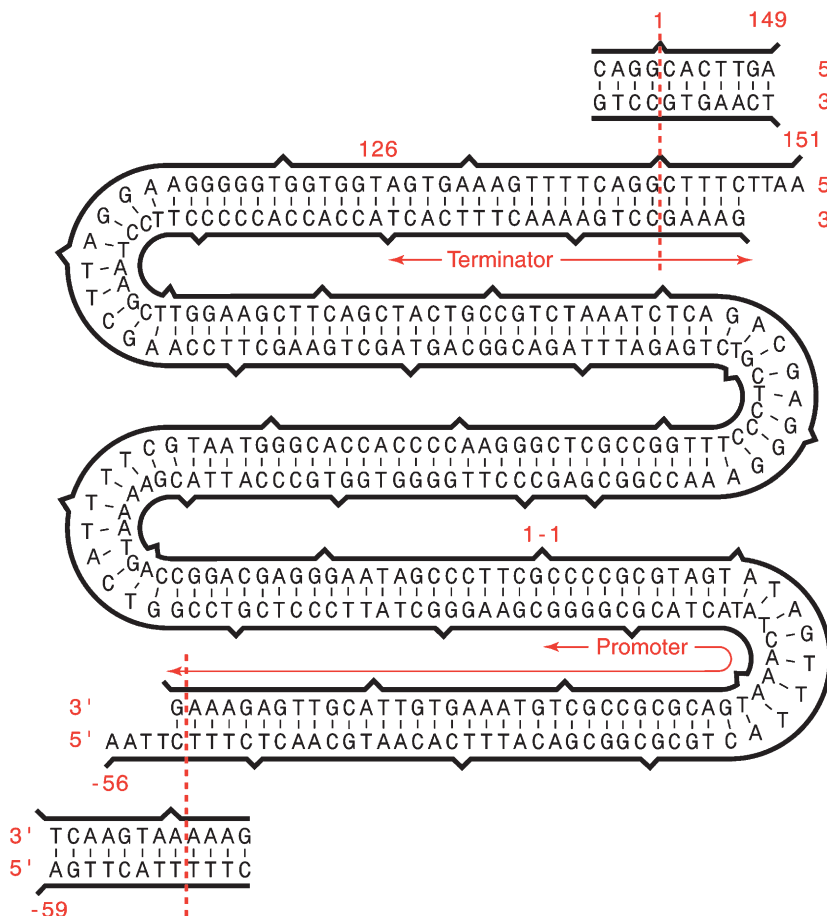


Figure 1.6 Khorana's totally synthetic DNA corresponding to the tyrosine suppressor transfer RNA gene (reprinted from Belagaje, R. *et al.*, in *Chemistry and Biology of Nucleosides and Nucleotides*, R.E. Harmon, R.K. Robins and L.B. Townsend (eds), Academic Press, New York. © (1978), with permission from Elsevier)

industry and synthetic genes became routinely used in the production of proteins. Further, **oligodeoxyribonucleotides**, the short pieces of single-stranded DNA for which Khorana developed the first chemical syntheses, became invaluable general tools in the manipulation of DNA, for example, as primers in DNA sequencing, as probes in gene detection and isolation, and as mutagenic agents to alter the sequence of DNA. From the late 1980s, research accelerated into synthetic **oligonucleotide analogues** as **antisense** modulators of gene expression in cells, as therapeutic agents (see Section 5.7) and for the construction of microarray chips for gene expression analysis.

The availability of synthetic DNA also provided new impetus in the study of DNA structure. In the early 1970s, new X-ray crystallographic techniques had been developed and applied to solve the structure of the dinucleoside phosphate, ApU, by Rich and co-workers in Cambridge, USA. This was followed by the complete structure of yeast phenylalanine tRNA, determined independently by Rich and by Klug and colleagues in Cambridge, England. For the first time, the complementary base pairing between two strands could be seen in greater detail than was previously possible from studies of DNA and RNA fibres. ApU formed a double helix by end-to-end packing of molecules, with Watson–Crick pairing clearly in evidence between each strand. The tRNA showed not only Watson–Crick pairs, but also a variety of alternative base pairs and base triples, many of which were entirely novel (see Sections 2.3.3 and 7.1.2).

Then in 1978, the structure of synthetic d(pATAT) was solved by Kennard and her group in Cambridge. This tetramer also formed an extended double helix, but excitingly revealed that there was a substantial sequence-dependence in its conformation. The angles between neighbouring dA and dT residues were quite different between the A–T sequence and the T–A sequence elements. Soon after, Wang and colleagues discovered that synthetic d(CGCGCG) adopted a totally unpredicted, left-handed Z-conformation. This was soon followed by the demonstration of both a **B-DNA** helix in a synthetic dodecamer by Dickerson in California and an A-DNA helix in an octamer by Kennard, and finally put paid to the concept that DNA had a rigid, rod-like structure. Clearly, DNA could adopt different conformations dependent on sequence and also on its external environment (see Section 2.3). More importantly, an immediate inference could be drawn that conformational differences in DNA (or the potential for their formation) might be recognized by other molecules. Thus, it was not long before synthetic DNA was also being used in the study of DNA binding to carcinogens and drugs (see Chapters 8 and 9) and to proteins (see Chapter 10).

These spectacular advances were only possible because of the equally dramatic improvements in methods of oligonucleotide synthesis that took place in the late 1970s and early 1980s. The laborious manual work of the early gene synthesis days was replaced by reliable automated DNA synthesis machines, which, within hours, could assemble sequences well in excess of 100 residues (see Section 4.1.4). Khorana's vision of the importance of synthetic DNA has been fully realized.

1.7 FRONTIERS IN NUCLEIC ACIDS RESEARCH

The last decade of the twentieth century was characterized by the quest to determine the complete DNA sequence of the **human genome**. Efforts by a publicly funded international consortium gathered considerable pace in the late 1990s in response to a challenge from a private company and the resultant concerns over the availability of sequencing data to the research community. The completion of the human genome sequence was duly announced by the consortium in April 2003, 50 years after papers on the discovery of the structure of the DNA double helix had been published and only 25 years since the first simple bacteriophage genome sequences were obtained. Genome sequences of many other organisms have also been completed, for example, mouse, nematode, zebrafish, yeast and parasites such as *Plasmodium falciparum* (see Section 6.5). The vast quantity of DNA sequence information generated has led to the founding of the new discipline of **Bioinformatics** in order to analyse and compare sequence data. One big surprise was that the human genome contains far fewer genes than expected, only about 24,500. We now know that production of the considerably larger number of human proteins and their regulation during cell division and biological development involves control of gene expression at many different stages (*e.g.* transcription, alternative splicing, RNA editing, translation, see Chapters 6 and 7), a full understanding of which is likely

to occupy biologists well into the twenty-first century. A recent technical advance here is the development of **microarrays** of synthetic oligonucleotides or **cdNAs** as hybridisation probes of DNA or RNA sequences both for mutational and gene expression analysis (see Section 5.5.4). This has led to the science of ‘-omics’, such as **genomics** and *ribonomics*, where DNA sequence variations can be studied and global effects of particular pathological states or external stimuli can be gauged on a whole genome basis.

A number of other advances have also been made in nucleic acids chemistry. First, a strong revival in the synthesis of nucleoside analogues has led to a number of therapeutic agents being approved for clinical use in treatment of AIDS and HIV infection as well as herpes and hepatitis viruses (see Section 3.7.2). Further, synthetic oligonucleotide analogues have become clinical agents for the treatment of viral infections and some cancers, although few have passed full regulatory approval as yet. The exploitation of the ‘antisense’ technology as a principle of therapeutic **gene modulation** has led to the investigation of a large number of nucleic acid analogues to enhance activity (see Section 5.7.1). As the twenty-first century arrived, gene modulation technology was finding increasing use to validate gene targets in cell lines and animals. At the same time, there was increasing recognition that other mechanisms of action can contribute to therapeutic effects of oligonucleotides in humans, such as stimulation of the immune system by ‘CpG’ domains (see Section 5.7.1), which may be harnessed perhaps for use as vaccine adjuvants.

The provision of synthetic RNA has also become routine (see Section 4.2) resulting in major advances in our understanding of catalytic RNA (**ribozymes**, see Sections 5.7.3 and 7.6.2) and protein-RNA interactions (see Section 10.9). New techniques of *in vitro* selection of RNA sequences have extended the potential of ribozymes and **aptamers** to carry out artificial reactions or bind unusual substrates, for example to act as ‘riboswitches’ responsive to certain analytes (see Section 5.7.3). A considerable upsurge of research in RNA biology has paralleled the availability of synthetic RNA. New ways have been elucidated for specific RNA sequences and structures to play important roles in gene regulation (*e.g.* **microRNA**, see Section 5.7.2). The exciting discovery of ‘**RNA interference**’ as a natural cell mechanism has led to the development of short synthetic RNA duplexes (siRNA and shRNA) as new gene control reagents that now rival, and may well surpass, antisense oligonucleotides for therapeutic and diagnostic use (see Section 5.7.2).

Dramatic advances have also been made in high-resolution structural determination of DNA and RNA sequences and their complexes with proteins (see Chapter 10), which are providing useful insights into molecular recognition and suggesting new approaches for drug design. In addition, the study of DNA recognition by small molecules in the minor groove has taken a major leap forward with the development of hairpin polyamides as a novel class of DNA-specific reagents with potential as drugs (see Section 9.7.4). Targeting of unusual DNA telomeric G-tetraplex structures is also an active area of current drug design (see Section 9.10).

The heady days of the discovery of the double helix and the elucidation of the genetic code are long gone, but in their place have come even more exciting times when many more of us now have the opportunity to answer fundamental questions about genetic structure and function and can utilise the insights and tools now available in the nucleic acids. ‘*You ain’t heard nothin’ yet folks*’ (Al Jolson, *The Jazz Singer*; July 1927).

REFERENCES

1. J.S. Fruton, *Molecules and life*. Wiley Interscience, New York, 1972, 180–224.
2. F.H. Portugal and J.S. Cohen, *A century of DNA*. MIT Press, Cambridge, MA, 1977.
3. O.T. Avery, C.M. MacLeod and M. McCarty, Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *J. Exp. Med.*, 1944, **79**, 137–158.
4. F. Miescher, *Die histochemischen und physiologischen arbeiten*. Vogel, Leipzig, 1897.
5. J.G. Buchanan and Lord Todd. *Adv. Carbohydr. Chem.*, 2000, **55**, 1–13.
6. D.H. Hayes, A.M. Michelson and A.R. Todd, Mononucleotides derived from deoxyadenosine and deoxyguanosine. *J. Chem. Soc.*, 1955, 808–815.
7. E. Chargaff, Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia*, 1950, **6**, 201–209.
8. J.D. Watson, *The Double Helix*. Atheneum Press, New York, 1968.

9. R. Olby, *The Path to the Double Helix*. Macmillan, London, 1973.
10. J.D. Watson and F.H.C. Crick, A structure for deoxyribose nucleic acid. *Nature*, 1953, **171**, 737–738.
11. H.F. Judson, *The Eighth Day of Creation*. Jonathan Cape, London, 1979.
12. M.W. Nirenberg, J.H. Matthei, O.W. Jones, R.G. Martin and S.H. Barondes, Approximation of genetic code via cell-free protein synthesis directed by template RNA. *Fed. Proc.*, 1963, **22**, 55–61.
13. S. Ochoa, Synthetic polynucleotides and the genetic code. *Fed. Proc.*, 1963, **22**, 62–74.
14. H.G. Khorana, Polynucleotide synthesis and the genetic code. *Fed. Proc.*, 1965, **24**, 1473–1487.
15. S.N. Cohen, The manipulation of genes. *Sci. Am.*, 1975, **233**, 24–33.
16. F. Sanger, Sequences, sequences and sequences. *Ann. Rev. Biochem.*, 1988, **57**, 1–28.
17. H.G. Khorana, Total synthesis of a gene. *Science*, 1979, **203**, 614–625.
18. K. Itakura, T. Hirose, R. Crea, A.D. Riggs, H.L. Heyneker, F. Bolivar and H.W. Boyer, Expression in *Escherichia coli* of a chemically synthesized gene for the hormone somatostatin. *Science*, 1977, **198**, 1056–1063.