From DNA to Protein

The Transfer of Genetic Information

Maria Szekely

Department of Biochemistry,
Imperial College, London, UK

M

Acknowledgements

I am greatly indebted to scientists, colleagues and friends from all over the world for the kind help that they gave me throughout the task of writing this book.

I wish to express my gratitude to Dr F. Sanger and to colleagues in his Division at the MRC Laboratory of Molecular Biology, Cambridge, in particular to Dr G. G. Brownlee, Dr B. G. Barrell, Dr C. Milstein, Dr P. H. Hamlyn and Dr F. E. Baralle; also to Professor H. G. Wittmann and to colleagues in his Division at the Max-Planck-Institut für Molekulare Genetik, Berlin, in particular to Dr R. Brimacombe, Dr R. A. Garrett, Dr A. R. Subramanian, Dr V. A. Erdmann, Dr K. H. Nierhaus, Dr K. Isono and Dr S. Isono. They all not only provided me with valuable material for this book in the form of reprints, preprints, photographs, etc. but also gave me the opportunity of having very fruitful discussions on recent progress in their field of research.

It is a pleasure to thank all those who kindly helped by sending me photographs for illustrations, preprints and other miscellaneous information on their research work. In this connection I wish to thank Dr G. Stöffler, Dr P. Leder, Dr M. Wu, Dr R. H. Cohn, Dr P. Sloof, Dr B. E. Griffin, Dr A. Jacobson, Dr J. A. Lake, Dr S. L. McKnight, Dr K. E. Koths, Dr D. Dressler, Professor O. L. Miller Jr, Dr D. F. Klessig, Dr B. Johnson, Mr M. J. F. Fowler, Dr A. Klug, Dr R. A. Cox, Dr A. J. Shatkin and Professor A. R. Fersht.

I am grateful for the valuable comments and advice of those who read some chapters of the manuscript, in particular to Dr F. Sanger, Dr H. Gould, Dr B. Johnson and Mrs T. Anderton.

I am also indebted to Miss M. Gordon for her help in compiling and correcting reference lists.

Abbreviations used throughout the text

MW, molecular weight

ss and ds (in relation to DNA or RNA), single stranded and double stranded, respectively.

A, C, G, T and U stand for nucleotides in RNA or DNA, N for an unspecified nucleotide. Only where the text leaves ambiguities in this respect are d and r used to distinguish between deoxyribonucleotides and ribonucleotides. Similarly, the position of the phosphate group (pN for 5′ phosphorylated nucleotides and Np for 3′ phosphorylated nucleotides) is indicated only if there is special emphasis on this structure or if the text would otherwise allow ambiguous interpretation.

bp, base pairs

kb, kilobases

cDNA, complementary DNA

r-protein, ribosomal protein

rRNA; ribosomal RNA

In the lists of references:

PNAS, *Proc. natl. Acad. Sci. U.S.A.*

Nature NB, *Nature New Biology*

Contents

PART 1: THE GENETIC MATERIAL

PART 2: THE TRANSCRIPT

PART 3: THE SYNTHESIS OF PROTEINS

Introduction

The last twenty-five years saw the birth and rapid development of a new discipline, molecular biology, which unites modern trends of biochemistry, biophysics and genetics. We can place its origin at the discovery of the double-helical structure of DNA, the molecular structure which gave us the first insight into the exact nature of genetic information. Development followed along the lines of research into the conservation and transfer of genetic information. Our present knowledge in this field has evolved through distinct stages landmarked by the introduction of new concepts, those of the double helix, the messenger, the Central Dogma and the genetic code, and with them new attitudes in approaching these problems. It may be added that recent results on the structure of overlapping genes and split genes have revealed new possibilities of how genetic information can be built into the DNA molecule. These results, which may well lead to a new concept of gene structure, will be discussed later, in Chapters 1 and 3.

THE DNA DOUBLE HELIX

At the time of Watson and Crick's discovery[1], DNA had already been recognised as the genetic material which comprises, encoded in its structure, the information for all genetically determined characteristics of any living organism. This also implied that such information is passed on from one generation to the next by producing exact replicas of these molecules. The Watson–Crick model of DNA revealed the way in which the genetic information can be built into the molecular structure and already pointed to the biological mechanism by which faithful copying of the structure can be achieved, ensuring that this information will be conserved over the generations. All the required information can be encoded into the sequence of nucleotides in the two strands of DNA. We can calculate in how many different ways the four bases can be arranged in a DNA stretch of known length; the number of different nucleotide sequences which in theory may form a 1000 nucleotide-long stretch (this is a reasonable size for one gene) is $4^{1000} = 10^{600}$. This enormous variability of nucleotide sequences in the two polynucleotide chains is more than sufficient to account for the number of different genes in different chromosomes. It follows that exact reproduction of the nucleotide sequence is the key to the conservation of genetic information.

The structure of the double helix was based on the recognition of the rules of base pairing: hydrogen bonding occurs between complementary bases, between A and T and between G and C. The strict complementarity of the two strands in the double-helical molecule means that the nucleotide sequence of one strand is unambiguously determined by that of the other. It also follows from the rules of base pairing that each DNA strand can serve as a template for the synthesis of a new strand of exactly defined nucleotide sequence. The mechanism of DNA replication that emerges, based on the rules of complementarity, explains how the two strands of DNA direct the synthesis of complementary strands, resulting in the production of two daughter molecules identical to the parental DNA (*figure 0.1*).

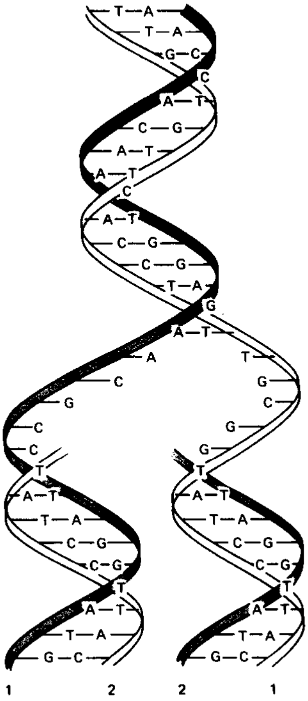

Figure 0.1 DNA double helix in the process of replication. 1: Parental strands; 2: newly synthesised strands. Shaded strands are complementary to white strands. Note that the base sequences of shaded strands 1 and 2 are identical, as are the base sequences of white strands 1 and 2.

The formation of pairs between complementary bases also proved to be the basis of the transfer of information from one molecular species to another. The synthesis of RNA on a DNA template follows the above mechanism as far as the copying of nucleotide sequences is concerned. In the synthesis of proteins the translation of nucleotide sequences into amino acid sequences is achieved via interaction of complementary trinucleotides. Base pairing has even more widespread significance: it is probably the basis of recognition of different control signals and plays a role in every nucleic acid-nucleic acid interaction.

THE MESSENGER CONCEPT

The messenger concept evolved from the necessity of finding a mediator between the biochemically stable, invariant DNA and the much more flexible protein pattern of the cell which, in prokaryotes, adjusts rapidly to the altered environment and to changes in the needs of the cell for different proteins. In eukaryotes the compartmentalisation of the cell which separates the site of protein synthesis from the site of genetic material localisation emphasised the need for a more mobile mediator which could establish direct contact with both the DNA which carries the information and the protein in which the information is expressed. In their classical paper, Jacob and Monod[2] describe the molecule which fulfils this function as a polynucleotide with a high rate of turnover which can temporarily associate with ribosomes and which is synthesised on, and reflects the base composition of, the DNA of the structural genes.

The obvious candidate for the function of messenger was RNA, a smaller molecule, biochemically less stable, present in the nucleus as well as in the cytoplasm, which can carry, encoded in its nucleotide sequence, the same messages as those present in the genetic material. Experimental evidence that RNAs indeed act as messengers was obtained soon after the launching of the messenger hypothesis, but it was years before the first mRNA was in fact isolated from animal cells. The difficulties which caused this delay were partly inherent in the problem–the huge number of different mRNA species in the cell, the low amount in which any individual species is present, the lability of bacterial mRNAs–and were partly due to technical shortcomings–the inadequacy of current techniques for separation and identification of mRNAs. Improvement in the methods of fractionation by gel electrophoresis and development of efficient cell-free protein-synthesising systems led eventually to the isolation of a pure RNA fraction from reticulocytes which directed the synthesis of globin in a reticulocyte lysate[3]. Isolation of a number of other mRNAs followed when more information had been collected on some structural characteristics of these molecules which facilitated their isolation, and when, in the knowledge of the genetic code and in possession of new methods for nucleic acid sequence analysis, identification of mRNAs by their nucleotide sequences became possible. Progress in the physical mapping of genomes and in the isolation of single genes also opened the way to the isolation of the corresponding mRNAs from bacterial cells. Today, quite a few cellular and viral mRNAs can be obtained in pure form; their number is increasing rapidly, as is the information collected on the structure of these molecules.

THE CENTRAL DOGMA

The Central Dogma was formulated by Crick to fit into a clear pattern the different observations available in the late 'fifties on the flow of information from one molecular species to another[4]. The rules laid down in the Central Dogma govern the transfer of genetic information in all living organisms. In their original, simpler form these rules described the ways generally used for information transfer in cellular processes:

$$\text{DNA} \xrightarrow[\text{transcription}]{} \text{RNA} \xrightarrow[\text{translation}]{} \text{protein}$$

which means that information can be transferred from DNA to RNA and from RNA to protein, but information cannot leave the protein and be passed to another molecule. This is strictly true for the mechanisms operating within the cell's own system of information transfer: RNA is synthesised on a DNA template, a process called transcription; and protein is made on an mRNA template by the process called translation. Transcription produces an RNA copy of one or a few messages encoded in the genetic material: the mRNA molecule. Translation leads to the synthesis of a protein the amino acid sequence of which corresponds to the same information content as that encoded into the nucleotide sequence of mRNA. The position of RNA in the diagram corresponds to the function of the messenger discussed above: it constitutes the link between the genetic material and the protein. It is the aim of this book to describe the mechanisms by which faithful transfer of information occurs between these molecules and to explore the manifold interactions between nucleic acids and proteins which ensure the high accuracy achieved in these processes.

The above diagram applies to the general mechanisms functioning in the cells. It does not account for the special processes observed in certain viruses which also make use of other ways to store, transfer and express genetic information. In RNA viruses the genetic information is originally encoded into RNA and not DNA. Some viruses of this group can replicate and produce messenger RNAs by copying their RNA molecules, thus producing new RNA strands. Others, the RNA tumour viruses, follow a different mechanism. They contain an enzyme, reverse transcriptase, which copies the viral RNA into a DNA molecule. The discovery of this enzyme[5] caused considerable excitement, as it seemed to bring about a flow of information in a direction opposite to that defined in the Central Dogma. In 1971, Crick[6] described the complete range of 'permitted' routes of information transfer, incorporating these special viral processes (*figure 0.2*).

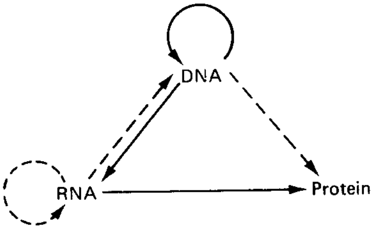

Figure 0.2 The routes of information transfer permitted by the Central Dogma.

The diagram in *figure 0.2* includes both the usual (solid lines) and special (broken lines) routes. It also shows (circular arrows) that DNA is used as a template for self-replication and that a similar self-replication process of RNA may exist in some viruses.

THE GENETIC CODE

If we only consider the processes of transcription and translation with respect to the ways in which genetic information is transferred from one molecule to another and disregard for the present the actual enzymic mechanisms by which these molecules are synthesised, we find that the first step involves simple copying of a nucleotide sequence, while the second requires a specific code system to relate the nucleotide sequences in RNA to the amino acid sequences in protein. Establishing the exact nature of this code and disclosing the actual code alphabet were the problems of molecular biology which probably attracted the greatest interest in the early 'sixties.

Some characteristics of the genetic code could be predicted on theoretical considerations, before the code alphabet had been determined. The discrepancy between the number of nucleotides, 4, and the number of amino acids, 20, suggested the necessity of a triplet code: only a combination of 3 nucleotides could yield more than 20 different codewords. However, the total number of trinucleotides, 64, is higher than that required to provide one codeword for each amino acid. This indicated either a degenerate code, in which several codons are used for one amino acid, or the existence of a number of 'nonsense codons', codewords which are not used in the message because they do not correspond to any amino acid. The nature of the code is a decisive factor with regard to the possible structures of mRNAs: a great number of nonsense codons would cause strong restrictions on the primary structure of any mRNA, while a degenerate code would allow a degree of flexibility: different primary structures could still comprise the same information. With the determination of the code alphabet and later with the nucleotide sequence analysis of different messages, it has been clearly established that the code is degenerate, and the corresponding flexibility in mRNA structure has also become apparent. Only three nonsense codons have been detected, all of which serve as termination signals of protein synthesis. Further characteristics, such as the lack of punctuation, i.e. the lack of any additional signals which separate the codon for one amino acid from that for the next, and the one ambiguity in the code, involving the initiation codon, were revealed in the course of investigations which led to the deciphering of the code alphabet and were confirmed when nucleotide sequences of natural mRNAs had been established. The lack of punctuation has the consequence that any nucleotide sequence can be interpreted in three different 'reading frames' (*figure 0.3*).

This is not only a theoretical possibility, for it has been found recently that the same nucleic acid sequence can comprise two messages: it can direct the synthesis of two different proteins which are encoded in the same sequence but in

```
          A U G C G C G C U U C G A U A A A A A U G A

(a)     | met | arg | ala | ser | ile | lys | met |
(b)         | ala | arg | phe | asp | lys | asn |
(c)       | cys | ala | leu | arg |
```

Figure 0.3 Decoding of a message in three different reading frames. The amino acid sequences (a) and (b) actually exist in proteins of ΦX174 phage. In reading frame (c) termination should occur after the fourth amino acid, as a nonsense codon, UAA, is present in the next position.

different reading frames[7]. Two of the amino acid sequences shown in *figure 0.3* actually exist in two viral proteins. It follows that faithful decoding of any message depends on recognition of the proper reading frame, which requires recognition of the specific site where decoding starts. There are two codons which can define initiation sites: AUG, which is generally used as initiation codon, and GUG, which occurs only infrequently at the initiation sites of mRNAs. Considering the important role of initiation codons in ensuring the accuracy of translation, the discovery that both triplets AUG and GUG are ambiguous in their interpretation was rather unexpected: in addition to functioning as initiation codons they can also code for amino acids at internal positions, AUG for methionine, GUG for valine. These represent the only ambiguity of the genetic code.

The code alphabet, shown in *figure 0.4*, has been worked out over a period of 4 years, by the combined efforts of several research groups using different approaches. The first codon, UUU, specifying phenylalanine, was determined by Nirenberg and Matthaei[8], who discovered that in a cell-free system the synthetic polynucleotide poly(U) directed the synthesis of polyphenylalanine. By similar experiments, some more codewords were revealed, but the method was restricted to nucleotide triplets present in synthetic polynucleotides that could function as artificial messengers. It had a further shortcoming in that it provided only the composition of the triplet and not the actual sequence of the three nucleotides. Khorana's group applied a more direct approach: they synthesised oligonucleotides of known structure and translated them into oligopeptides[9]. These experiments yielded the exact structure of a number of codewords and confirmed in a direct way the triplet nature of the code. Nirenberg and Leder[10] eventually developed a new assay technique which led to the elucidation of all trinucleotides coding for the 20 amino acids. This was based on the property of aminoacyl-tRNAs (see below) that they can bind to the ribosome in the presence of their cognate codon. Phenylalanyl-tRNA, for example, binds to the ribosome in the presence of the trinucleotide UUU. By testing the binding of all available aminoacyl-tRNAs with all possible trinucleotides, Nirenberg and coworkers were able to assign each codon to an amino acid, with the exception of the three nonsense codons.

Concomitantly with the investigation into the nature of the genetic code, studies have been carried out to reveal the mechanism by which the coded message can be 'read' by the protein-synthesising system. Considering the very different chemical structures involved, direct and specific interaction between trinucleotides and amino acids could be ruled out. The participation of an adaptor molecule was indicated in the decoding process. This proved to be an RNA molecule, and by a lucky coincidence, it was an RNA species which could be isolated in pure form and was therefore accessible to functional studies at that time impossible with other RNAs. The specific adaptors were found to be transfer RNAs (tRNAs), small RNA molecules which can fulfil this function because they possess a specific site which recognises the codon (by complementary base pairing) and another specific site to which the corresponding amino acid can be bound. These molecules thus provide the link between the codon and the amino acid specified by it. Another function of tRNAs is that they carry the amino acids to the site of protein synthesis. The aminoacylated tRNA can bind to the ribosome if its cognate codon is present. (This property has been

Second letter

First letter	U		C		A		G		Third letter
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Ter*	UGA	Ter*	A
	UUG		UCG		UAG		UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

Figure 0.4 The genetic code.* The three triplets UAA, UAG, UGA with no amino acid allocated to them, are nonsense codons which lead to the termination of the polypeptide chain.

exploited to decipher the genetic code.) In the process of decoding the message, transfer RNAs thus play a central role.

The mechanism of these processes, as well as of other steps in translation, will be described in later chapters of this book. It may be useful, however, to present here a brief outline of the gross events in translation, as we know them today, so as to enable the reader to fit the different components of the system and their functions into a proper pattern.

THE MAIN EVENTS IN TRANSLATION

The ribosomal particle provides the basic machinery for the synthesis of polypeptide chains, for all components of the translation system attach to it: the mRNA which carries the genetic message, the aminoacyl-tRNA which reads the message and delivers the corresponding amino acid to the ribosome, and the growing peptide chain, also bound to a tRNA molecule, in the form of peptidyl-tRNA.

The first step in translation is the formation of an *initiation complex*, which consists of the mRNA, the ribosome which binds to the specific initiation site in this mRNA and the initiator tRNA which interacts with the initiation codon (AUG or GUG) and carries a free or formylated methionine residue. This methionine is required for initiation; it is removed from most proteins before or shortly after synthesis is terminated. Some protein factors, the initiation factors, also participate in the formation of the initiation complex; they are released after completion of this first stage. At the next stage, elongation, the ribosome moves along the mRNA in the direction from the 5′ to the 3′ end. To each codon which subsequently comes into contact with the ribosome, a specific aminoacyl-tRNA binds and then attaches also to the ribosome, in a position which places the amino acid at the functional centre where the peptide bond is formed. Protein factors, the elongation factors, also take part in these processes. The actual chemical reaction which leads to peptide bond formation is carried out by components of the ribosomal particle itself. It involves a transfer of the existing peptidyl group from tRNA to the α-amino group of the incoming aminoacyl-tRNA.

In this way, growth of the peptide chain takes place from the N-terminus to the C-terminus. At each step of the elongation of the polypeptide chain, there are two tRNA molecules attached to two specific sites on the ribosome: one tRNA carries the existing peptide chain and the other carries the new amino acid. When a termination (nonsense) codon is reached which has no cognate tRNA, only the peptidyl-tRNA remains on the ribosome, and it undergoes hydrolysis instead of peptide transfer, yielding a free complete polypeptide chain. With the aid of protein factors, the release factors, all components are released from the ribosome and the whole process can start anew.

In the following chapters, the different mechanisms will be described by which the genetic information is transferred from one molecule to another: from the parental DNA to the daughter molecules in replication, from DNA to RNA in transcription and from RNA to protein in translation. The structure of DNA and RNA will be discussed in relation to the ways in which information is encoded into these molecules. The structural features which function as control signals, safeguarding the accuracy or ensuring the flexibility of these processes, will be emphasised. As the main subject of this book is the mechanism of information transfer, less emphasis will be placed on the purely enzymological aspects of the synthesis of nucleic acids and proteins. The many rather sophisticated techniques which are applied today to study these procedures will be described only as far as the principle of the methods is concerned, with the exception of the determination of nucleotide sequences. As the nucleotide sequence of a DNA or mRNA molecule constitutes the genetic information, it is of special interest to describe the current techniques which enable us to establish these sequences and to learn in which form this information is built into the genome or into the messenger molecule.

REFERENCES

1 Watson, J. D. and Crick, F. H. C. *Nature,* **171**, 737, 964 (1953).
2 Jacob, F. and Monod, J. *J. Mol. Biol.,* **3**, 381 (1961).
3 Laycock, D. G. and Hunt, J. A. *Nature,* **221**, 1118 (1969).
Lockard, R. E. and Lingrel, J. B. *Biochem. Biophys. Res. Comm.*, **37**, 204 (1969).
4 Crick, F. H. C. *Symp. Soc. Exptl. Biol.,* **12**, 138 (1958).
5 Temin, H. M. and Mizutani, S. *Nature,* **226**, 1211 (1970).
Baltimore, D. *Nature,* **226**, 1209 (1970).
6 Crick, F. *Nature,* **227**, 561 (1970).
7 Barrell, B. G., Air, G. M. and Hutchison, C. A. III *Nature,* **264**, 34 (1976).
Smith, M., Brown, N. L., Air, G. M., Barrell, B. G., Coulson, A. R., Hutchison, C. A. III and Sanger, F. *Nature,* **265**, 702 (1977).
Shaw, D. C., Walker, J. E., Northrop, F. D., Barrell, B. G., Godson, G. N. and Fiddes, J. C. *Nature,* **272**, 510 (1978).
8 Nirenberg, M. W. and Matthaei, J. H. *PNAS,* **47**, 1588 (1961).
9 Khorana, H. G. *et al. Cold Spring Harbor Symp.,* **31**, 39 (1966).
10 Nirenberg, M. and Leder, P. *Science,* **145**, 1399 (1964).
Nirenberg, M. *et al. Cold Spring Harbor Symp.,* **31**, 39 (1966).

PART 1

The Genetic Material