Molecular Genetics of Bacteria

Jeremy Dale

Department of Microbiology, University of Surrey, UK

A Wiley–Interscience Publication

JOHN WILEY & SONS

Chichester · New York · Brisbane · Toronto · Singapore

Library of Congress Cataloging in Publication Data

Dale, Jeremy.
Molecular genetics of bacteria.
'A Wiley-Interscience publication.'
Bibliography: p.
Includes index.
1. Microbial genetics 2. Molecular genetics.
3. Genetic engineering. I. Title.
QH434.D35 1989 589.9'015 88-33823
ISBN 0 471 92209 9

British Library Cataloguing in Publication Data

Dale, Jeremy
Molecular genetics of bacteria.
1. Bacteria. Genetics
I. Title
589.9'015

ISBN 0 471 92209 9

Typeset by Inforum Typesetting, Portsmouth
Printed and bound in Great Britain
by Biddles Ltd, Guildford, Surrey

Preface

This book is addressed primarily to the multitude of students for whom genetics is just one of many subjects that has to be undergone in order to achieve, ultimately, some qualification or other. You have probably heard something from time to time of the exciting, and perhaps frightening, advances that are being made by genetic engineers, but it all seems too far off and too complicated to entertain ambitions of understanding what is going on.

Unfortunately, the impression that genetics is difficult and complicated can be fostered for some students by the over-generous provision found in most of the existing text-books. My intention has been to select, from the vast amount of information and concepts available, that material that I feel is useful and interesting (practically and/or scientifically). Any such selection is inevitably a personal one, and involves leaving out much that others would consider just as important. I hope however that it establishes a working platform from which I can lead you in the direction of some of the most exciting discoveries that are being made in any branch of science at the moment. In this way, perhaps you will come to realise that genetics can be comprehensible and at least some of you will develop a genuine interest in the field that will inspire you to enquire further.

Contents

CHAPTER ONE

Nucleic Acid Structure and Function

his book it is assumed that you will already have a working knowledge of essentials of molecular biology, especially the structure and synthesis of nucleic acids and proteins. The purpose of this chapter therefore is merely to serve as a reminder of some of the most relevant points, and to highlight those features that are particularly essential for an understanding of later chapters.

In bacteria, the genetic material (i.e. the cell component that carries information from one generation to the next) consists of double-stranded DNA. The same is true of most organisms but viruses provide a notable exception: although many do have double-stranded DNA, in some (such as the bacterial viruses ϕX174 and M13) the genetic material is single-stranded DNA. Still other viruses have RNA as the genetic material, either single-stranded (as in the bacterial virus MS2 and many mammalian viruses, such as influenza and polio) or (as in some other animal and plant viruses) double-stranded RNA. Some of these viruses are considered in more detail in Chapter 6.

Leaving these viruses on one side, and reverting to the typical bacterial situation of double-stranded DNA, there are a few features of this DNA that are especially noteworthy. The basic components of DNA (Figure 1–1) are the sugar 2′-deoxyribose, phosphate residues and four heterocyclic bases: the purines adenine and guanine, and the pyrimidines thymine and cytosine. Each DNA strand is composed of a 'backbone' of 2′-deoxyribose residues linked by phosphate groups, which join the 3′ position of one sugar to the 5′ position of the next (Figure 1–2). One of the four bases is linked to the 1′ position of each deoxyribose. It is the sequence of these four bases that carries the genetic information. The two strands are twisted around each

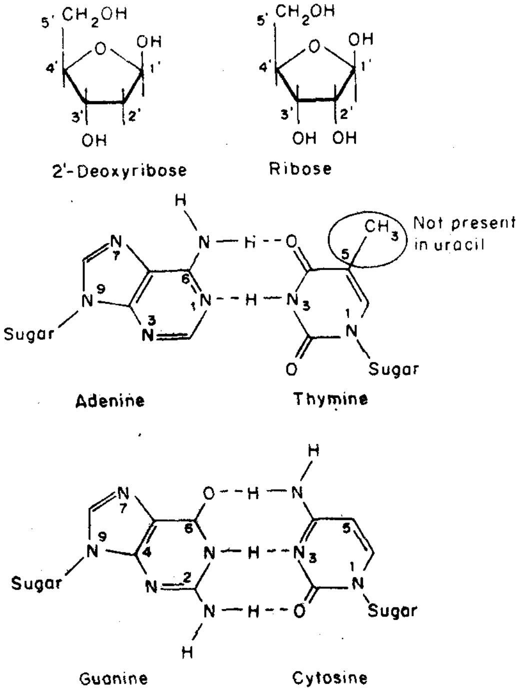
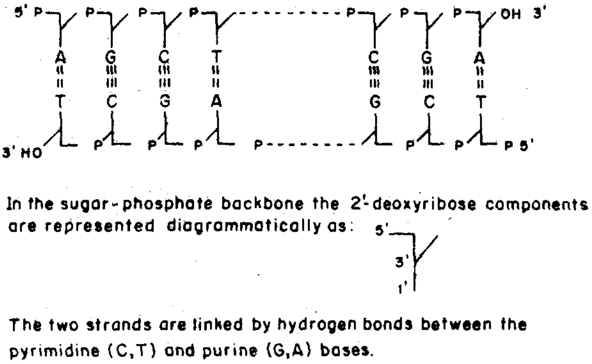


Figure 1–1 Structure of the basic elements of DNA and RNA



In the sugar-phosphate backbone the 2'-deoxyribose components are represented diagrammatically as:

The two strands are linked by hydrogen bonds between the pyrimidine (C,T) and purine (G,A) bases.

Figure 1–2 Diagrammatic structure of DNA

other in the now familiar 'double helix', with the bases in the centre and the sugar–phosphate backbone on the outside. The two strands are linked by means of hydrogen bonds between the purine bases (adenine, guanine) and the pyrimidines (thymine, cytosine). The only arrangement of these bases that is consistent with maintaining the helix in its 'correct' conformation is when adenine is paired with thymine and guanine with cytosine. Any other pairing will result in distortion of the helix. Note that the purines are larger than the pyrimidines, and that this arrangement involves one purine opposite a pyrimidine at each position, so the distance separating the strands can remain constant. The structure of RNA differs from that of DNA in that it is usually single-stranded (with the exceptions noted above), contains the sugar ribose instead of deoxyribose, and uracil instead of thymine.

Since the two strands of DNA are only linked by non-covalent hydrogen bonds, they can be easily separated in the laboratory by any changes in physical conditions, such as increased temperature or high pH, that tend to disrupt hydrogen bonds. Separation of the two DNA strands, a process known as denaturation, is readily reversible. Reducing the temperature, or the pH, will allow hydrogen bonds between complementary DNA sequences to re-form; this is referred to as re-annealing. If DNA molecules from different sources are denatured, then mixed and allowed to re-anneal, it is possible to form hydrogen bonds between regions of DNA that are similar but not identical. *Hybridisation* between single-stranded DNA molecules forms the basis of the use of *DNA probes* to detect specific sequences in samples of DNA. This technique, which is described more fully in subsequent chapters, forms an important part of modern molecular biology.

Temporary separation of localised regions of the two DNA strands also occurs under physiological conditions inside the cell as an essential part of processes such as replication and transcription. Note that there are three hydrogen bonds linking guanine and cytosine, while the adenine–thymine pairing has only two hydrogen bonds. The two DNA strands are therefore more strongly attached in those regions with a high G+C content. Because of this, such regions are more resistant to denaturation and conversely re-anneal more readily. The influence of base composition on the ease of separation of two nucleic acid strands may play an important role in the control of processes such as the initiation of RNA synthesis where an A–T-rich region may facilitate the initial separation of the DNA strands (Chapter 5).

A further noteworthy feature of the helix is that each strand can be said to have a direction, based on the orientation of the linkages in the sugar–phosphate backbone. Each phosphate group joins the 5′ position of one sugar residue to the 3′ position of the next deoxyribose. In Figure 1–2, the upper strand has a free 5′ group at the left-hand end and a 3′ OH group at the right-hand end. It is therefore said to run (from left to right) in the 5′ to 3′ direction. Conversely, the lower strand, reading from left to right, runs in the 3′ to 5′ direction. As we will see further on, all nucleic acids are synthesised in the 5′ to 3′ direction; that is, they start at the 5′ end and grow by the addition of nucleotides to the free 3′ OH group. The phosphate to make the link is provided by the substrate, which is the nucleoside 5′-triphosphate; the energy required is provided by the release of the two terminal phosphate groups.

REPLICATION OF DNA

The pairing of the bases in the centre of the helix pays a key role in the transmission of genetic information, both from one cell to its successors, and also in the synthesis of components of the cell. A DNA strand can act as a *template* for the synthesis of a new nucleic acid strand in which each base forms a hydrogen-bonded pair with one on the template strand (G with C, A with T, or A with U for RNA molecules). The new sequence is thus *complementary* to that on the template strand (and is of course identical to the second strand of the original DNA molecule). The copying of DNA molecules to produce more DNA is known as *replication*; the synthesis of RNA on a DNA template is called *transcription*.

Replication is of course a much more complicated process than implied by the above statement, and for a full treatment you will need to refer to a molecular biology textbook. Some of the main features are summarised in Figure 1–3. The opposite polarity of the DNA strands is a complicating factor. One of the new strands (the 'leading' strand) can be synthesised continuously in the 5′ to 3′ direction. The enzyme responsible for this synthesis is known as DNA polymerase III. With the other new strand, however, the overall effect is of growth in the 3′ to 5′ direction. Since nucleic acids can only be synthesised in the 5′ to 3′ direction, the new 3′ to 5′ strand (the 'lagging' strand) has to be made in short fragments which are subsequently joined together by the action of another enzyme, DNA ligase. Furthermore, DNA polymerase III (as with other DNA polymerases) is incapable of starting a new DNA strand, but can only extend a previously existing molecule. This restriction does not apply to RNA polymerases, which are able to initiate the synthesis of new nucleic acids. Each fragment is therefore started with a short piece of RNA, produced by the action of a special RNA polymerase known as primase. This RNA primer can then be extended by DNA polymerase III. The primer is subsequently removed, and the gap filled in, by a different DNA polymerase (DNA polymerase I); this enzyme can carry out both of these actions since it has exonuclease as well as

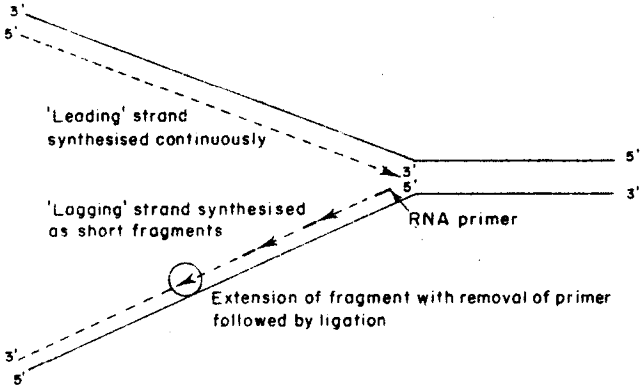


Figure 1–3 Simplified view of the main features of DNA replication

polymerase activity. After the gap has been filled, the fragments are joined together by DNA ligase.

A further complication arises from the twisting of the two DNA strands around each other. DNA molecules within the cell cannot normally rotate freely. In bacterial cells, for example, the DNA is usually circular. Therefore it is not possible to produce a pair of daughter molecules by just separating the two strands and synthesising the complementary strands, as is implied by the simplified representation in Figure 1–3. The strands have to be unwound to be separated If they are not free to rotate, separating the strands at one point will cause overwinding further along. Unless this problem is overcome, the molecules would quickly get in a hopeless tangle. (If you don't understand this, try it for yourself with some bits of string!) The problem is resolved by the activity of enzymes known as *DNA topoisomerases*, of which the best known is *DNA gyrase*. By breaking and rejoining DNA strands, these enzymes are capable of altering the extent to which the two DNA strands are wound around one another. One such enzyme, operating ahead of the replication fork, ensures that the DNA is in a suitable conformation so that the two strands can be separated.

The role of topoisomerases does not end there. Within the cell, DNA does not normally exist as a simple helix, but the helix itself is wound up into coils (*supercoiled*). This occurs through the activity of DNA gyrase, which reduces the winding of the helix after replication, producing 'underwound' DNA. Since a DNA helix has a natural tendency to exist with a specific degree of winding (one turn per 10.4 base pairs), an underwound DNA molecule will be stressed; the stress can, however, be relieved by adopting a negatively supercoiled configuration. (Conversely, an overwound molecule would tend to form supercoils in the opposite direction.) A discussion of the mathematical relationship between winding of the helix and supercoiling is not necessary here; the relationship can fortunately be demonstrated quite easily with the same piece of string. Fix one end and twist the other, while holding the string taut. Then, without letting go of the string, bring the two ends together. The stress introduced by twisting the string is relieved by the string twisting round itself, i.e. by forming supercoils.

Bacterial DNA is normally negatively supercoiled, to the extent of about one twist per 200 base pairs. A supercoiled molecule is much more compact, which is useful in the laboratory as it helps in the isolation of bacterial plasmids (Chapter 7). These are small circular pieces of DNA that exist independently of the chromsome, and play a key role in gene cloning (Chapter 10). The small size of the plasmid means it can be isolated in the intact supercoiled form, whereas the chromosomal DNA will be broken into linear fragments during lysis of the cell.

The compact supercoiled structure of the DNA is also useful to the bacterium, since the *E. coli* chromosome, in its expanded state, would be several hundred times longer than the cell itself. Supercoiling enables all this information to be packaged within the cell in an organised manner. It is also thought likely that the supercoiling (and other structural features) of the DNA play a role in the regulation of gene expression.

GENE EXPRESSION

The expression of the genetic material occurs for the most part through the production of proteins, involving two consecutive steps in which the information is converted from one form to another: *transcription* and *translation* (Figure 1–4).

The first step is the conversion of the information into the form of an RNA molecule known as messenger RNA (mRNA). The process by which this occurs (transcription) is carried out by the enzyme RNA polymerase. The concept is very similar to that of DNA replication, in that the DNA acts as a template, and the sequence of bases in the newly synthesised nucleic acid reflects that in the template strand, owing to the hydrogen bonding between the bases and the specificity of the enzyme concerned. There are two major differences between transcription and replication (beyond the differences between RNA and DNA). Firstly, only a comparatively short molecule is produced, and secondly, only one of the DNA strands is transcribed. (Some genes use one strand, and some use the other, but in general any specific region of DNA is only transcribed from one strand.) Transcription is therefore considerably simpler than replication: the mRNA can be made continuously, using a single enzyme, and there are fewer topological problems.

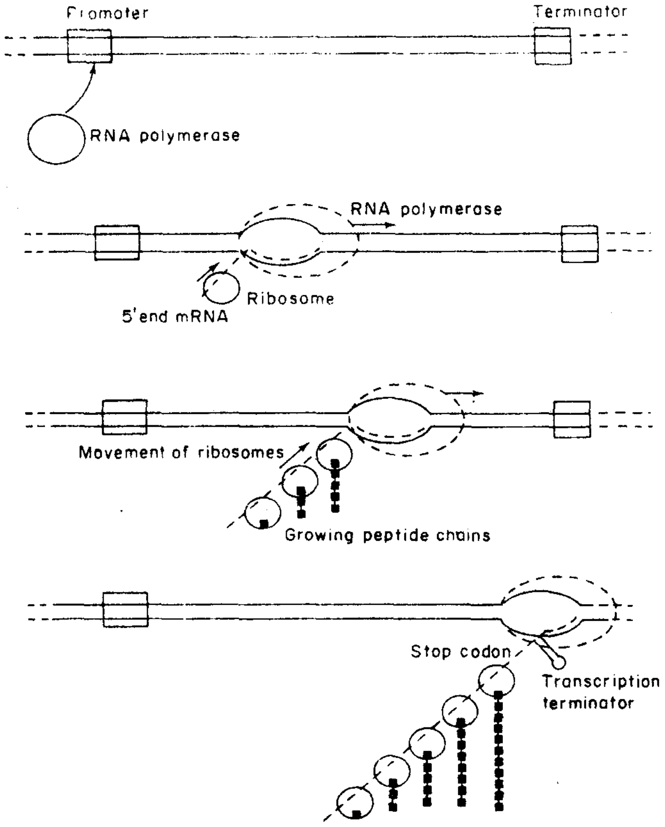


Figure 1–4 Maii features of gene expression

Since transcription results in the synthesis of comparatively short mRNA molecules (often just a few kilobases long, corresponding to a defined block of several genes) there must be a large number of signals around the chromosome that direct the RNA polymerase to start transcription at the required place, and to stop when the block of genes has been transcribed. The start signals (promoters) also convey the information as to the direction in which transcription should proceed, or which strand to work from — which is another way of saying the same thing.

After the RNA polymerase binds to the promoter, transcription starts from an adjacent site, and the mRNA is synthesised in the 5′ to 3′ direction, using a single strand of the DNA as the template. Transcription then proceeds through the gene, or group of genes, until a termination signal is

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

Some examples of non-standard code:

In yeast mitochondria:
UGA codes for Trp (instead of stop)
CUN codes for Thr (instead of Leu) (N = U,C,G or A)
AUA codes for Met (instead of Ile)

In human mitochondria:
UGA, AUA code for Trp, Met, respectively
AGA, AGG = stop codons

Mycoplasma capricolum:
UGA = Trp

Tetrahymena (a ciliated protozoan):
UAA = Gln

Figure 1–5 Standard genetic code

reached, when the mRNA and the RNA polymerase are released. The ways in which the promoter and terminator signals operate will be dealt with in Chapter 5.

The mRNA carries the information for the sequence of amino acids in a protein in the form of the genetic code (Figure 1–5), in which each occurrence of one of the 64 groups of three nucleotides (triplets or *codons*) conveys the information for a specific amino acid (or in some cases a stop signal to indicate the end of the protein). The code shown is almost universal.

However, there are some exceptions, principally in mitochondria. Some examples of these differences are shown in Figure 1–5. More exceptions are likely to be found yet, but the basic features of the code are constant in all known cases.

Synthesis of the protein itself is mediated by ribosomes, which in bacteria attach to a specific sequence on the mRNA (the *ribosome binding site*, also known as the *Shine–Dalgarno sequence* after the workers who first recognised its significance). This sequence is partly complementary to a region at the 3′ end of the 16 S rRNA, so that binding of the ribosomes can be mediated by hydrogen bonding between the complementary base sequences. This will normally occur as soon as the binding site is available, so the mRNA will start to be translated while it is still being formed (Figure 1–4). Protein synthesis (translation) starts at a specific triplet: usually AUG but sometimes GUG. This *initiation codon* is found between four and ten bases from the ribosome binding site. The position of the ribosome binding site and the initiation codon determines the *reading frame*. From this point, the sequence is read in consecutive groups of three bases; as we will see later, the addition or deletion of a single base will change the reading frame, with the consequence that the coding property of the subsequent message is totally different.

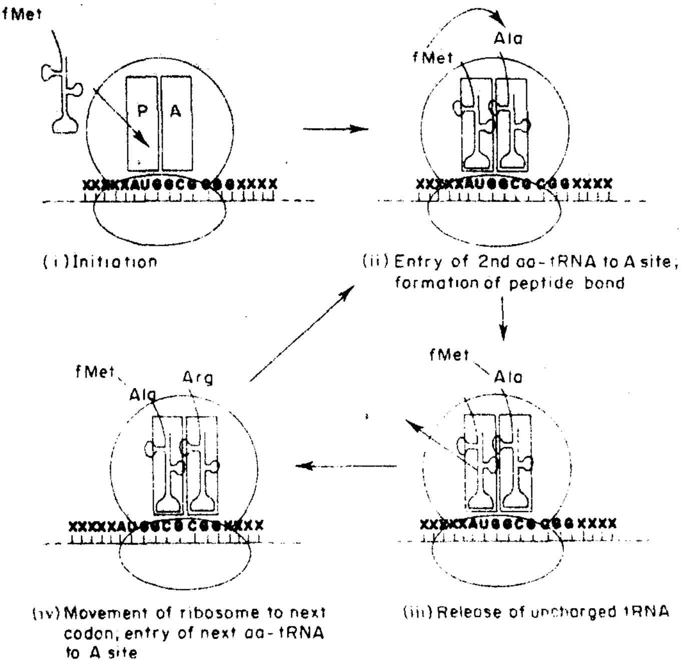
Recognition of each triplet *codon* is mediated by small RNA molecules known as transfer RNA (tRNA). One part of the tRNA molecule contains a sequence (the *anticodon*) that forms hydrogen bonds with the codon. There is at least one tRNA species specific for each amino acid. The appropriate amino acid is added to the tRNA by a specific enzyme (one of a number of aminoacyl tRNA synthetases) which has a crucial dual specificity: it is capable of recognising a single tRNA species and also the correct amino acid with which that tRNA should be charged. Thus, for example, the codon UGG which codes for tryptophan will be recognised by a specific tryptophan tRNA; this tRNA will be recognised by the tryptophanyl tRNA synthetase which is also specific for tryptophan. This therefore ensures that the tRNA is charged with the appropriate amino acid.

It is important to realise that there are three separate elements to the specificity of this process: codon–anticodon interaction, recognition of the specific tRNA by the aminoacyl tRNA synthetase (which involves features of the tRNA other than the anticodon), and recognition by the enzyme of the appropriate amino acid. Since tRNA molecules are all basically quite similar, and some amino acids (such as isoleucine and valine) are also similar to one another, it would not be surprising if mistakes were made occasionally. The low frequency of such errors (it has been estimated that one protein in a thousand contains one incorrect amino acid) is due to the existence of an editing mechanism whereby the synthetase is able to cleave the amino acid from an incorrectly charged tRNA molecule.

If all three elements of specificity were absolute, then there would have to be at least 61 different tRNA species: one for each of the 64 codons less the three stop codons, for which there is no corresponding tRNA molecule. For many of the amino acids there are indeed multiple tRNA species with different codon specificities. Some of these tRNA molecules are present at comparatively low levels in the cell, which would indicate that there could be a difficulty in translating that particular codon. This can be correlated to some extent with the frequency of occurrence of particular codons (*codon usage*), in that those codons that require a rare tRNA species tend also to occur less commonly, at least in highly expressed genes.

However, this is not the complete story. For many tRNA molecules, the codon–anticodon recognition is not absolutely precise; in particular, there is some latitude allowed in the matching of the third base of the codon. A rather complex set of rules (the *wobble* hypothesis) has been developed to account for the extent of allowable mismatching. The consequence of this is that some tRNA molecules are able to recognise more than one codon. The number of tRNA species required for recognition of the complete set of codons is thus considerably less than 61.

In bacteria, the AUG initiation codon is recognised by a specific tRNA molecule, tRNA$^{\text{fMet}}$. After this tRNA molecule is charged with methionine, the amino acid is modified by another enzyme to form *N*-formylmethionine. Aminoacylated tRNA molecules normally bind to a site on the ribosome known as the A site (acceptor), while their anticodon region pairs with the mRNA. Only after peptide bond formation is the tRNA able to move to a second site on the ribosome, the P (peptide) site. The fMet–tRNA$^{\text{fMet}}$ (i.e. the tRNA$^{\text{fMet}}$ charged with formylmethionine) is unique in being able to enter the P site directly. The tRNA$^{\text{fMet}}$ anticodon recognises (forms base pairs with) the AUG codon (Figure 1–6). The charged tRNA corresponding to the second codon then enters the A site on the ribosome and peptide bond formation occurs by transfer of the fMet residue to the second amino acid. The tRNA$^{\text{fMet}}$, now uncharged, is released, and the ribosome moves one codon along the mRNA, which is accompanied by movement of the second tRNA molecule (now charged with a dipeptide) from the A site to the P site. The A site is thus free to accept the charged tRNA corresponding to the third codon. When the ribosome has moved far enough, the ribosome binding site is exposed again and another ribosome can attach to it. A single mRNA molecule will therefore carry a number of ribosomes actively translating the sequence. Each ribosome moves along the mRNA until a *stop codon* is reached. The absence of a corresponding tRNA species capable of recognising this codon causes translation to stop at this point. The polypeptide chain is then released and the ribosome dissociates from the mRNA. (This description of the process is of necessity highly simplified. More detailed accounts are available in the molecular biology texts listed in the 'further reading' section.)



(i) Initiation

(ii) Entry of 2nd aa-tRNA to A site, formation of peptide bond

(iii) Release of uncharged tRNA

(iv) Movement of ribosome to next codon, entry of next aa-tRNA to A site

Figure 1–6 Outline of protein synthesis

GENE ORGANISATION

In bacteria, genes with related functions are often located together in a group known as an operon (Figure 1–7). This group of genes has a single promoter site and is transcribed into a single *polycistronic* mRNA molecule, which carries the information for several proteins. At the 5′ end of the mRNA there may be a length of sequence known as the leader, which is not translated into any of the identifiable proteins coded for by that gene. This sequence can play a major role in the regulation of the operon (see Chapter 5). Translation does not start at the first AUG encountered, but at a specific position determined by the ribosome binding site and the adjacent AUG initiation codon. This combination also determines which of the three possible reading frames is used for the first structural gene.

After the ribosome has translated the first *cistron*, it may dissociate as normal, in which case translation of the next cistron will require the attachment of ribosomes to another binding site adjacent to the initiation