

✓ 1/3/14 (168)

PREFACE

GENES V

Benjamin Lewin

OXFORD UNIVERSITY PRESS
Oxford New York Tokyo
1994

Personalis EX2172

Oxford University Press, Walton Street, Oxford OX2 6DP

*Oxford New York Toronto
Delhi Bombay Calcutta Madras Karachi
Kuala Lumpur Singapore Hong Kong Tokyo
Nairobi Dar es Salaam Cape Town
Melbourne Auckland Madrid
and associated companies in
Berlin Ibadan*

Oxford is a trade mark of Oxford University Press

*Published in the United States
by Oxford University Press Inc., New York*

© Oxford University Press and Cell Press, 1994

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, without the prior permission in writing of Oxford University Press. Within the UK, exceptions are allowed in respect of any fair dealing for the purpose of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act, 1988, or in the case of reprographic reproduction in accordance with the terms of licences issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms and in other countries should be sent to the Rights Department, Oxford University Press, at the address above.

This book is sold subject to the condition that it shall not, by way of trade or otherwise, be lent, re-sold, hired out, or otherwise circulated without the publisher's prior consent in any form of binding or cover other than that in which it is published and without a similar condition including this condition being imposed on the purchaser.

A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data

*Typeset by BPCC Electronic Composition Limited
Printed in U.S.A.*

(as we see in detail in Part 2). Messenger RNA is synthesized by the same process of complementary base pairing used to replicate DNA, with the important difference that it corresponds to only one strand of the DNA double helix. **Figure 6.16** shows that the sequence of messenger RNA is complementary with the sequence of one strand of DNA and is identical (apart from the replacement of T with U) with the other strand of DNA. The convention for writing DNA sequences is that the top strand runs 5'-3', with the sequence that is the same as RNA.

A messenger RNA includes a sequence of nucleotides that corresponds with the sequence of amino acids in the protein. This part of the nucleic

acid is called the **coding region**. But the messenger RNA includes additional sequences on either end; these sequences do not directly represent protein. The gene is considered to include the entire sequence represented in messenger RNA. Sometimes mutations impeding gene function are found in the additional, noncoding regions, confirming the view that these comprise a legitimate part of the genetic unit.

Figure 6.17 illustrates this situation, in which the gene is considered to comprise a continuous stretch of DNA, needed to produce a particular protein. It includes the sequence coding for that protein, but also includes sequences on either side of the coding region.

Eukaryotic genes can be interrupted

The simple view of the gene was upset in 1977 by the discovery of **interrupted genes**. The primary evidence for the existence of these interruptions was a comparison between the structure of DNA and the corresponding messenger RNA. The messenger RNA always includes a nucleotide sequence that corresponds exactly with the protein product according to the rules of the genetic code. *But the gene includes additional sequences that lie within the coding region, interrupting the sequence that represents the protein.*

This discrepancy between the structure of the DNA and messenger RNA is common in eukaryotes, and has been found in an archebacterium (representing another evolutionary branch of the bacteria). It has not been found in the eubacteria (which provide the typical prokaryotes), although it is present in a bacteriophage of *E. coli*.

The sequences of DNA comprising an interrupted gene are divided into the two categories depicted in **Figure 6.18**:

◆ **Exons** are the regions that are represented in the messenger RNA.

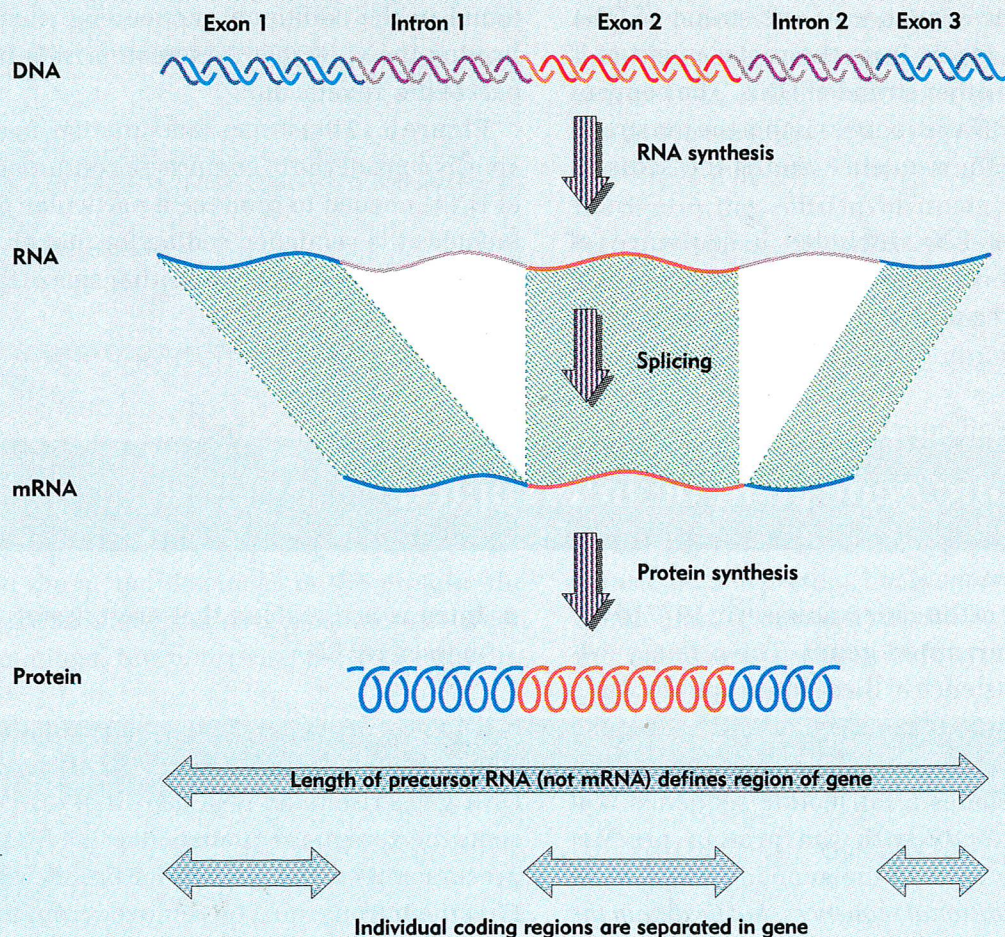
◆ **Introns** are regions that are missing from the messenger RNA.

The process of gene expression requires a new step, one that does not occur in eubacteria. The DNA gives rise to an RNA copy that exactly represents the genome sequence. But this RNA is only a precursor; it cannot be used for producing protein. First the introns must be removed from the RNA to give a messenger RNA that consists only of the series of exons. This process is called **RNA splicing**.

How does this change our view of the gene? Following splicing, the exons are always joined together in the same order in which they lie in DNA. Thus the colinearity of gene and protein is maintained between the individual exons and the corresponding parts of the protein chain. The *order* of mutations in the gene remains the same as the order of amino acid replacements in the protein. But the *distances* in the gene do not correspond at all with the distances in the protein. The length of the gene is defined by the length of the initial (precursor) RNA instead of by the length of the messenger RNA.

Figure 6.18

Interrupted genes are expressed via a precursor RNA, from which the introns are removed when the exons are spliced together to give the messenger RNA.



All the exons are represented on the same molecule of RNA, and their splicing together occurs only as an *intramolecular* reaction. There is usually no joining of exons carried by *different* RNA molecules, so that the mechanism excludes any splicing together of sequences representing different alleles. So mutations located in different exons of a gene cannot complement one another; thus they continue to be defined as members of the same complementation group.

If mutations that affect the sequence of a higher eukaryotic protein could be mapped with sufficient

resolution, they should be found in clusters. Each cluster should correspond to an exon, and should be separated from the next cluster on the genetic map by a distance corresponding to the relative length of the intron. Thus recombination frequencies cannot be taken as a guide to relative distance within an interrupted gene.

What are the effects of mutations in the introns? Since the introns are not part of the messenger RNA, mutations in them cannot directly affect protein structure. However, they can prevent the production of the messenger RNA—for example, by

inhibiting the splicing together of exons. A mutation of this sort acts only on the allele that carries it, therefore fails to complement any other mutation in that allele, and so constitutes part of the same complementation group as the exons.

In the interrupted genes of eukaryotes, most introns appear to serve no function other than to be removed during gene expression. However, there are some exceptions, most notably in the yeast mitochondrion, in which an intron itself codes for the production of a protein that functions independently from the protein coded by the exons. In this case, mutations in the intron fall into a different

complementation group from that represented by mutations in the exons. Thus one complementation group may lie between clusters of mutations that comprise another group.

Eukaryotic genes are not necessarily interrupted. Some correspond directly with the protein product in the same manner as eubacterial genes. In yeast, the majority of genes in fact is uninterrupted. In higher eukaryotes, most genes are interrupted; and in fact the introns are usually much longer than exons, creating genes that are very much larger than their coding regions. The structure of such genes is the subject of Chapter 23.

Some DNA sequences code for more than one protein

Most genes consist of a sequence of DNA that is devoted solely to the purpose of coding for one protein (although the gene may include noncoding

regions at either end and introns within the coding region). However, there are some cases in which a sequence of DNA does not have a unique function in

Figure 6.19

Two proteins can be generated from a single gene by starting (or terminating) expression at different points.

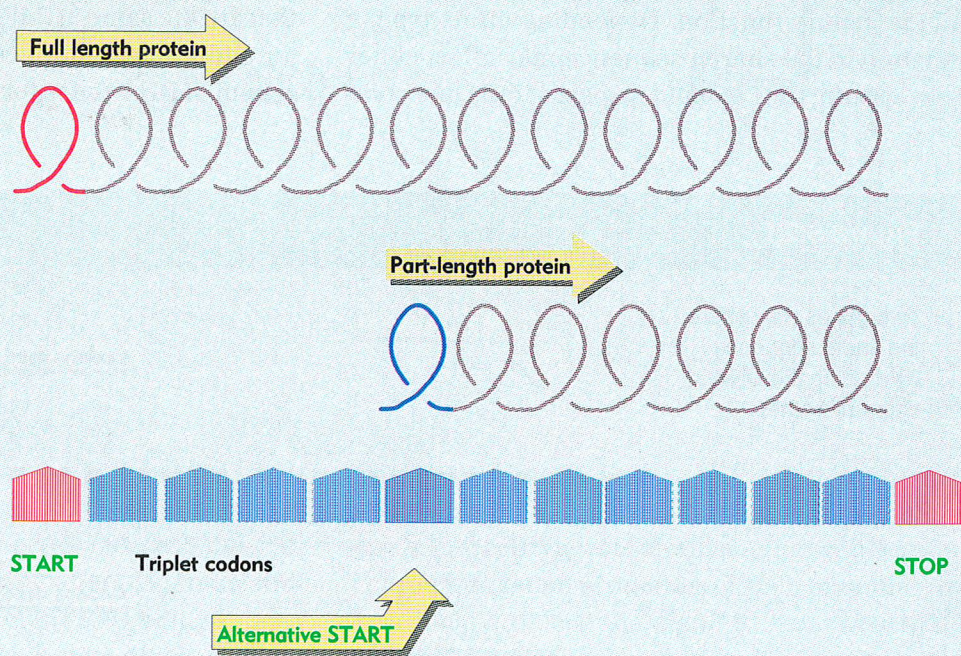


Table 7.1

Considering *E. coli* in terms of its macromolecular components.

Component	Proportion of Dry Cell Mass	Molecules/Cell	Number of different kinds	Copies of each kind
Wall	10%	1	1	1
Membrane	10%	2	2	1
DNA	1.5%	1	1	1
mRNA	1%	1,500	600	2-3
tRNA	3%	200,000	60	>3000
rRNA	16%	38,000	2	19,000
Ribosomal protein	9%	10^6	52	19,000
Soluble protein	46%	2.0×10^6	1850	>1000
Small molecules	3%	7.5×10^6	800	

An overall view of the attention devoted to protein synthesis in the intact bacterium is given in Table 7.1. The 20,000 or so ribosomes account for a quarter of the cell mass. There are >3000 copies of each tRNA, and altogether, the tRNA molecules outnumber the ribosomes by almost tenfold; most of them are present as aminoacyl-tRNAs, that is, ready to be used at once in protein synthesis. Because of their instability, it is difficult to calculate the

number of mRNA molecules, but a reasonable guess would be 1500, in varying states of synthesis and decomposition. There are ~600 different types of mRNA in a bacterium. This suggests that there are usually only 2-3 copies of each mRNA per bacterium. On average, each probably codes for ~3 proteins. If there are 1850 different soluble proteins, there must on average be >1000 copies of each protein in a bacterium.

The meaning of the genetic code

Any one of four possible nucleotides can occupy each of the three positions of the codon, so that there are $4^3 = 64$ possible combinations of trinucleotide sequences. Thus there are more codons than the 20 amino acids from which proteins are synthesized.

Soon after the discovery of the triplet nature of the code, two methods were developed to allow codons to be assigned systematically to amino acids. Both used *in vitro* systems consisting of components of the apparatus that synthesizes proteins:

- ◆ A protein-synthesizing system from *E. coli* was used to translate *synthetic polynucleotides*. The first report of success with such a system was Nirenberg's demonstration in 1961 that polynucleic acid [poly(U)] can act as an mRNA to direct the assembly of phenylalanine into polyphenylalanine. This result means that UUU must be a codon for phenylalanine. Subsequently, many other synthetic polynucleotides, consisting of known sequences of

different bases, were used by Khorana; they allowed the meaning of about half of the 64 codons to be assigned.

- ◆ The **ribosome-binding assay** for making codon assignments was developed by **Nirenberg** and **Leder** in 1964. A trinucleotide can be used to mimic a codon, by causing the corresponding aminoacyl-tRNA to bind to a ribosome. A **triple complex** of trinucleotide•aminoacyl-tRNA•ribosome can be isolated by taking advantage of the ability of ribosomes to bind to nitrocellulose filters. The aminoacyl-tRNA itself does not bind to such filters, but is retained as part of the triple complex. Its retention is detected by means of a radioactive label in the amino acid component. Then the meaning of each trinucleotide

can be determined by testing which one of 20 labeled aminoacyl-tRNA preparations is retained on the filter.

The two techniques together assigned meaning to all of the codons that represent amino acids. Since then, the sequencing of DNA has made it possible to compare corresponding nucleotide and amino acid sequences directly. *The sequence of the coding strand of DNA, read in the direction from 5' to 3', consists of triplets corresponding to the amino acid sequence of the protein read from N-terminus to C-terminus.* The entire genetic code has been confirmed in overwhelming detail from such analysis.

The code is summarized in **Figure 7.9**. A striking feature is its **degeneracy**: almost every amino acid

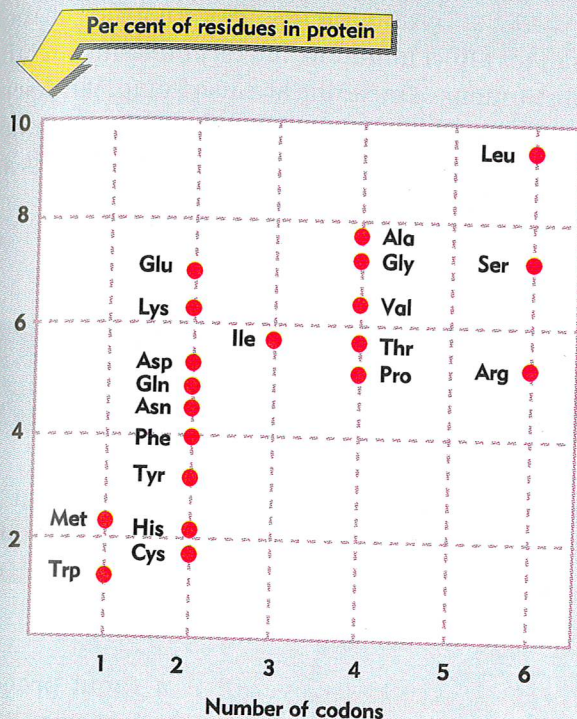
Figure 7.9

All the triplet codons have meaning: 61 represent amino acids, and 3 cause termination (TERM).

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

Figure 7.10

The number of codons for each amino acid does not correlate with its frequency of use in proteins.



is represented by several codons. The only exceptions are methionine and tryptophan. Codons that have the same meaning are called **synonyms**. Figure 7.10 plots the number of codons representing each amino acid against the frequency with which the amino acid is used in proteins (in *E. coli*). Although there is a tendency for amino acids that are more common to be represented by more codons, this is slight, and therefore it does not seem that the genetic code has been optimized with regard to the utilization of amino acids.

Codons representing the same or related amino acids tend to be similar in sequence. Often the base in the third position of a codon is not significant, because the four codons differing only in the third base represent the same amino acid. Sometimes a distinction is made only between a purine versus a

pyrimidine in this position. The reduced specificity at the last position is known as **third-base degeneracy**.

The tendency for similar amino acids to be represented by related codons minimizes the effects of mutations. It increases the probability that a single random base change will result in no amino acid substitution or in one involving amino acids of similar character. For example, a mutation of CUC to CUG has no effect, since both codons represent leucine; a mutation of CUU to AUU results in replacement of leucine with isoleucine, a closely related amino acid.

Three codons (UAA, UAG and UGA) do not represent amino acids. They are used specifically to terminate protein synthesis; one of these **stop codons** marks the end of every gene.

Is the genetic code the same in all living organisms?

Comparisons of DNA sequences with the corresponding protein sequences reveal that the *identical set of codon assignments is used in bacteria and in eukaryotic cytoplasm*. As a result, mRNA from one species usually can be translated correctly *in vitro* or *in vivo* by the protein synthetic apparatus of another species. Thus the codons used in the mRNA of one species have the same meaning for the ribosomes and tRNAs of other species.

The universality of the code argues that it must have been established very early in evolution. Originally there may have been a stereochemical relationship between amino acids and the codons representing them. Then the system now used for protein synthesis evolved by selection for features such as greater efficiency or accuracy.

Perhaps the code started in a primitive form in which a small number of codons were used to represent comparatively few amino acids, possibly even with one codon corresponding to any member of a group of amino acids. More precise codon meanings and additional amino acids could have been introduced later. One possibility is that at first only two of the three bases in each codon were used; discrimination at the third position could have evolved later.

Evolution of the code could have become 'frozen'

at a point at which the system had become so complex that any changes in codon meaning would disrupt existing proteins by substituting amino acids. Its universality implies that this must have happened at such an early stage that all living organisms are descended from a single pool of primitive cells in which this occurred.

Exceptions to the universal genetic code are rare. Changes in meaning in the principal genome of a species usually concern the termination codons. For example, in a mycoplasma, UGA codes for tryptophan; and in certain species of the ciliates

Tetrahymena and *Paramecium*, UAA and UAG code for glutamine.

Systematic alterations of the code have occurred only in mitochondrial DNA. Such changes may have been possible because the mitochondrion is a relatively closed system, concerned with the synthesis of a few specific proteins. The majority of these changes affect initiation and termination, but other substitutions of meaning are also found. We discuss changes in the genetic code in Chapter 8, in the context of the interaction between aminoacyl-tRNA and messenger RNA.

The ribosomal sites of action

Synthesis of proteins involves an assembly line in which the ribosomes proceed inexorably along the messenger, bringing in the aminoacyl-tRNAs that provide the actual building blocks of the protein product. The ribosome itself constitutes a small mobile factory, in which a compact package of proteins and rRNAs forms several active centers able to undertake various catalytic activities. Different sets of accessory factors assist the ribosome in each of the three stages of protein synthesis: initiation, elongation, and termination. Energy for ribosome movement is provided by hydrolysis of GTP.

◆ **Initiation** involves the reactions that precede formation of the peptide bond between the first two amino acids of the protein. It requires the ribosome to bind to the mRNA, forming an initiation complex that contains the first aminoacyl-tRNA. This is a relatively slow step in protein synthesis, and usually determines the rate at which an mRNA is translated.

◆ **Elongation** includes all the reactions from synthesis of the first peptide bond to addition of the last amino acid. Amino acids are added to the chain one at a time; the addition of an amino acid is the most rapid step in protein synthesis.

◆ **Termination** encompasses the steps that are needed to release the completed polypeptide chain; at the same time, the ribosome dissociates from the mRNA.

Protein synthesis overall is a rapid process, although the rate depends strongly on temperature. In bacteria at 37°C, ~15 amino acids are added to a growing polypeptide chain every second. So it takes only ~20 seconds to synthesize an average protein of 300 amino acids. In eukaryotes, the rate of protein synthesis is slower; in red blood cells at 37°C, elongation typically sees ~2 amino acids added to the chain per second.

Most of the experiments to define the stages of protein synthesis have been performed with *in vitro* systems, consisting of ribosomes, aminoacyl-tRNAs, other enzymatic factors, and an energy source. In these systems, the rate of protein synthesis is slower by an order of magnitude than the rate *in vivo*.

Messenger RNA is associated with the small subunit, ~30 bases of the mRNA being bound at any time. But only two molecules of tRNA are involved in peptide bond synthesis at any moment. So polypeptide elongation involves reactions taking place at just two of the (roughly) ten codons covered by the ribosome.