Genes: Definition and Structure

Introductory article

Millard Susman, *University of Wisconsin, Madison, Wisconsin, USA*

The word 'gene' has two meanings: (1) the determinant of an observable trait or characteristic of an organism, or (2) the DNA sequence that determines the chemical structure of a specific polypeptide molecule or RNA molecule. Since the observable characteristics of organisms include the chemical structures of their constituent molecules, these two definitions merge at the molecular level.

Article Contents

Historical Definition of a Gene

The observable characteristics of an organism are referred to collectively as its phenotype. Humans have known for thousands of years that organisms within species – cattle, grain crops, dogs, date palms, horses and, of course, humans – differ from one another phenotypically and that phenotypic variations are, to a large extent, heritable. Long before the mechanisms of heredity were understood in any depth, humans were selectively breeding plants and domestic animals to enhance desirable qualities and to eliminate undesirable ones.

Early theories of heredity proposed that offspring were a concoction of fluids derived from one or both parents and that inherited characteristics were somehow determined by the properties of these fluids. Biologists up to the time of Charles Darwin, including Darwin himself, believed that inherited characteristics were literally dissolved like sugar in water. This notion led to a difficult problem: Characteristics conveyed from one generation to the next in the form of liquids would become more and more dilute with the passage of time. How could natural selection produce evolutionary change if favourable variations in phenotype were as evanescent as a drop of juice?

Genes as 'particles'

Modern genetics began with the realization that the gene is not a fluid, but a 'particle', which can maintain its integrity over many generations. Thus, favourable variants of a given gene can become more numerous in a population as the forces of natural selection work over time.

The word 'gene' was coined by W. Johannsen in 1909, but the modern concept of the gene originated with Gregor Mendel, who in the 1860s studied the inheritance of characteristics that differed sharply and unambiguously among true-breeding varieties of garden peas. Mendel found that a hybrid between two phenotypically distinct varieties resembled one of the two parents – the dominant parent (**Figure 1**). When these hybrids were allowed to self-pollinate, however, one-quarter of the offspring resembled the other parent, the so-called recessive parent. For example, a cross between a tall variety and a short one produced only tall offspring, whereas one would expect offspring of intermediate height if the determinants of tallness mixed liked fluids. When these tall hybrid plants were allowed to self-fertilize, they produced a mixed progeny, 3/4 tall and 1/4 short. 'Shortness' had not been dissolved or lost in the hybrid plants; it had somehow been preserved in a form that reappeared in the next generation. Mendel correctly interpreted these observations to mean that height was governed by paired 'factors' that we now call genes.

Mendel guessed that a true-breeding tall plant contained a pair of *T* genes, one inherited from its male parent and the other from its female parent. It could be symbolized *TT*. Such a plant produced gametes (germ cells) containing a single *T* gene. Similarly, a true-breeding short plant was *tt* and produced only *t* gametes. A cross between a tall plant and a short one would yield a *Tt* hybrid. In *Tt* hybrids, the *T* gene happened to be dominant so that the hybrid plants were tall. (The term dominant was coined by Mendel to describe this situation.) Unlike its parents, the hybrid would produce two kinds of gametes, *T* and *t*, and, when these joined randomly during self-fertilization, the offspring would be *TT*, *Tt* or *tt* in proportions of 1:2:1. Because *T* was dominant to *t*, the phenotypes of the *TT* and *Tt* offspring would be identical; hence the 3:1 phenotypic ratio among the progeny. The presence of short *tt* individuals in this second generation demonstrated that the recessive 'short' form of the gene was neither dissolved nor altered during its passage through the *Tt* heterozygote. The *t* form of the gene (called the *t* allele) retained its identity in the *Tt* heterozygote and reappeared unchanged in *tt* (homozygous recessive) offspring when *Tt* hybrids were allowed to self-fertilize. In short, the various alternative forms (alleles) of a given gene acted like particles, not like droplets.

Experience has now taught us that dominance is not always complete, as it was in the crosses that Mendel did with his peas. Sometimes a cross of an *AA* homozygote by an *aa* homozygote produces an *Aa* heterozygote with a

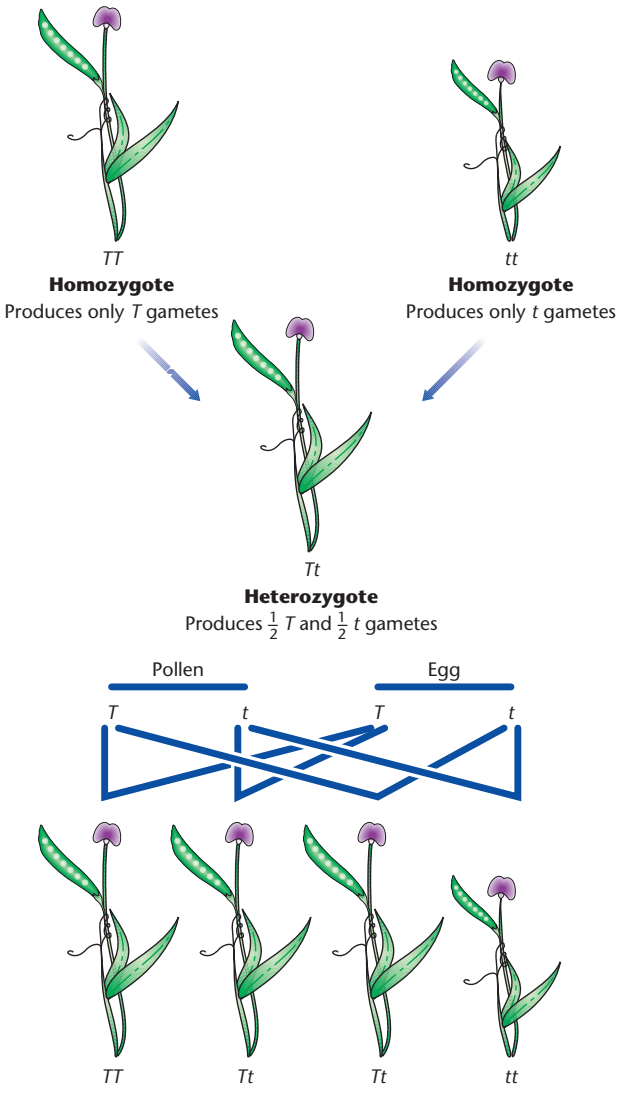

Figure 1 Mendel's demonstration of the particulate nature of the gene. Two true-breeding strains of peas, one tall and the other short, were crossed to one another. All of their offspring were tall. At the level of their genes, the tall plants were *TT* homozygotes, and the short plants were *tt*. When a tall plant is crossed to another tall plant, the gametes are all *T*, and the offspring produced are all *TT* homozygotes. Like the parents, they are tall. Similarly, *tt* homozygotes, when crossed to one another, produce only short, *tt* homozygotes. A cross of a tall plant with a short one produces *Tt* heterozygotes, which are tall because the *T* allele of the tallness gene is dominant to the *t* allele. When these *Tt* heterozygotes make gametes, half of the gametes are *T* and half are *t*. In a cross between two *Tt* heterozygotes (or a self-cross, which is genetically the same) *T* and *t* gametes join at random to form zygotes. A *T* egg is equally likely to be fertilized by a *T* pollen nucleus or a *t* pollen nucleus. Similarly, a *t* egg is equally likely to be fertilized by a *T* pollen nucleus or a *t* pollen nucleus. Thus, the progeny of a cross between two heterozygotes is 1/4 *TT*, 1/2 *Tt* and 1/4 *tt*. The homozygous *TT* offspring are tall, and so are the heterozygous *Tt* offspring (because *T* is the dominant allele). However, the *tt* homozygotes are short. This experiment shows that the *t* allele remains unchanged in the *Tt* heterozygote and reappears in the offspring of the *Tt* heterozygote when an appropriate cross is done.

phenotype intermediate between the phenotypes of the parents. This is called incomplete dominance. In such a situation, a cross between two *Aa* individuals produces offspring in which both the original *AA* phenotype and the original *aa* phenotype reappear. The ratio will be 1 *AA* : 2 *Aa* : 1 *aa*. Again, the important lesson of this experiment is that the original *A* and *a* forms of the gene do not intermingle in the *Aa* hybrid. They retain their integrity and emerge unchanged in the next generation.

Genes make enzymes that control cellular chemistry

The idea that genes are responsible for the manufacture of proteins was first proposed in 1902 by a British physician, Sir Archibald Garrod, who realized that alkaptonuria was an inherited metabolic condition in humans and hypothesized that it was due to the absence of an enzyme (a catalytic protein) required for the breakdown of homogentisic acid.

Systematic investigation of the relationship between genes and enzymes did not occur until 1941, when George W. Beadle and Edward L. Tatum isolated many nutritional mutants of the fungus, *Neurospora crassa*, each of which required a specific dietary supplement for survival. Beadle and Tatum in 1941 demonstrated that each mutant was deficient in an enzyme that catalysed a specific step in the biosynthesis of the required nutrient. This led to the conclusion that the major role of genes was to carry the information for manufacturing the enzymes that catalysed the chemical reactions of the cell. This idea, originally called the 'one gene-one enzyme hypothesis', now goes by the name 'one gene-one polypeptide' because it soon became apparent that some enzymes consist of complexes of several polypeptides, each produced by a different gene. Furthermore, many polypeptide products of genes are not enzymes at all, but serve as structural elements, sensory components, transport agents or regulators. Furthermore, as we shall discuss below, some of the end products of genes are not polypeptides at all, they are RNA molecules.

The Genetic Code

Genes are composed of DNA (or, in the case of some viruses, RNA). DNA is a polymeric molecule made up of small subunits – monomers – strung together to make very long linear chains. There are four different monomers in DNA: the purine bases, adenine (A) and guanine (G), and the pyrimidine bases, thymine (T) and cytosine (C). Most DNA molecules are double-stranded and consist of two complementary strands in which A in one strand is always paired with T in the second and G in one strand is always paired with C in the second. This complementarity of the two strands – that is, the strict adherence to the A:T and G:C pairing rules – is the basis for replication of the genes.

The two strands separate during replication, and each serves as a template for the synthesis of a new complementary strand.

Proteins are polymers made up of chains of amino acids. They constitute most of the chemical and structural machinery of the cell. All together, there are 20 different amino acids in the proteins, and a typical protein is a string of several hundred amino acids. Genes contain the instructions for stringing together the correct sequence of amino acid for each of the thousands of proteins needed by the cell. The message contained in each gene consists of code-words for the amino acids in the protein product of the gene, written in precisely the order in which the amino acids must be connected to one another to make the protein. This message must be written in a language that uses the four 'letters' – A, T, G and C – to form words that can be read by the protein-synthesizing machinery of the cell.

The structure of the genetic language is very simple. All of the words in the dictionary, which is called the genetic code, are three letters long. These three-letter words are called codons. With four different letters in the alphabet, it is possible to make $4 \times 4 \times 4 = 64$ different codons, far more than enough to specify the names of all 20 amino acids. When the code, shown in **Figure 2**, was finally cracked, it turned out that every one of the 64 possible codons had a meaning; 61 specified the name of an amino acid and three meant 'stop'. The three stop codons tell the protein-synthesizing machinery of the cell when the end of a protein has been reached. The 61 codons that correspond to the names of 20 amino acids must include some synonyms, and they do. The existence of synonyms in the code was called, by Francis Crick, 'degeneracy'. Most synonymous codons have the first two letters in common and differ only in the third base. There are only two amino acids, methionine and tryptophan, that have unique codons. There is one amino acid, isoleucine, that has three codons. There are three – leucine, serine and arginine – that have six.

Second base

First base	U		C		A		G		Third Base
U	UUU	*phe*	UCU	*ser*	UAU	*tyr*	UGU	*cys*	U
	UUC	*phe*	UCC	*ser*	UAC	*tyr*	UGC	*cys*	C
	UUA	*leu*	UCA	*ser*	UAA	*stop*	UGA	*stop*	A
	UUG	*leu*	UCG	*ser*	UAG	*stop*	UGG	*trp*	G
C	CUU	*leu*	CCU	*pro*	CAU	*his*	CGU	*arg*	U
	CUC	*leu*	CCC	*pro*	CAC	*his*	CGC	*arg*	C
	CUA	*leu*	CCA	*pro*	CAA	*gln*	CGA	*arg*	A
	CUG	*leu*	CCG	*pro*	CAG	*gln*	CGG	*arg*	G
A	AUU	*ile*	ACU	*thr*	AAU	*asn*	AGU	*ser*	U
	AUC	*ile*	ACC	*thr*	AAC	*asn*	AGC	*ser*	C
	AUA	*ile*	ACA	*thr*	AAA	*lys*	AGA	*arg*	A
	AUG	*met*	ACG	*thr*	AAG	*lys*	AGG	*arg*	G
G	GUU	*val*	GCU	*ala*	GAU	*asp*	GGU	*gly*	U
	GUC	*val*	GCC	*ala*	GAC	*asp*	GGC	*gly*	C
	GUA	*val*	GCA	*ala*	GAA	*glu*	GGA	*gly*	A
	GUG	*val*	GCG	*ala*	GAG	*glu*	GGG	*gly*	G

Figure 2 The genetic code. RNA contains four bases, adenine (A), guanine (G), cytosine (C) and uracil (U). One difference between DNA and RNA is that RNA contains U instead of T. These four bases can be assembled into 64 possible triplet codons, every one of which has a meaning in the genetic code. The codons are traditionally written in their mRNA form. The amino acid translations of the codons are shown in italics. The abbreviations stand for phenylalanine, leucine, isoleucine, methionine, valine, serine, proline, threonine, alanine, tyrosine, histidine, glutamine, asparagine, lysine, aspartate, glutamate, cysteine, tryptophan, arginine, and glycine. The three 'stop' codons, UAA, UAG and UGA are given the whimsical names ochre, amber and opal, respectively.

Reading Frames

In 1962, Francis Crick *et al*. showed that the rules for reading the genetic code are simple: The protein-synthesizing machinery of the cell reads the code from a fixed starting point and moves along the genetic message in steps that are three bases long, stopping at each step to read the triplet of bases and insert the corresponding amino acid into the protein product. The starting point determines the reading frame of the message; once started, the protein-synthesizing machinery treats each successive triplet of bases as a codon.

Crick *et al*. showed that the insertion or removal of a single base in the DNA of a gene drastically altered the genetic message but that adding or removing three bases often resulted in a genetic message almost as good as the original message. The three added (or deleted) bases were not adjacent to one another, but could, in some cases, be scattered over a length corresponding to ten codons or more. Crick and his collaborators pointed out the only reasonable explanation of this remarkable result: If a single base were added or removed from the message, the reading frame would be shifted one notch, and the entire message would be misread from the site of the added (deleted) base to the end. An added or deleted base is therefore referred to as a frameshift mutation. If the gene contained three such frameshift mutations within a fairly short segment, the net result would be that the message between the first and last of the three mutations would be out of step with the 'real' message, but beyond the third mutation, the correct reading frame would be restored. This readily explains the functionality of the protein product of a gene containing three frameshift mutations of the same sign.

A corollary conclusion from this experiment is that most of the possible triplets of bases must correspond to some amino acid. If only 20 of the 64 possible triplets had any

meaning in the genetic code, then the segment between the first and the last of three successive frameshift mutations would almost certainly contain some nonsense. In reality, however, the frameshifting of a short segment of the message produced a meaningful message most of the time. Clearly the more-or-less random set of codons produced by frameshifting a short segment of the genetic message must usually encode some amino acids, although not the amino acids found in the original protein. Of course, frameshifts of this sort can be tolerated only if they occur in codons that specify relatively unimportant amino acids – that is, amino acids that can be altered without inactivating the protein product of the gene.

Initiation and Termination Signals

The expression of a gene – that is, the production of its protein product – occurs in two steps. First, the DNA in the gene serves as a template to produce a corresponding messenger RNA (mRNA), a process called transcription. Then, the protein-synthesizing machinery – consisting of ribosomes, transfer RNAs (tRNAs), and a variety of protein enzymes and 'factors' – uses the mRNA template to direct the synthesis of a protein, a process called translation. The DNA of the chromosome contains many genes lined up one after another, but mRNAs generally contain the message for only one gene (in eukaryotes) or several adjacent genes (in prokaryotes). Thus, there must be 'start' and 'stop' signals for transcription that identify the proper locations for the ends of an mRNA. The mRNAs themselves have untranslated regions at both ends – segments of base sequence that are not used to encode any amino acids. Thus, there must be 'start' and 'stop' signals for translation that identify the proper locations for the ends of the protein product.

Transcriptional starts and stops

The synthesis of mRNA is catalysed by the enzyme RNA polymerase. The starting point for a given mRNA is determined by nucleotide sequences on the DNA itself. These sequences, called promoters, serve as binding sites for RNA polymerase and various accessory proteins, called transcription factors. The binding of RNA polymerase to the promoter determines the point at which transcription of the genetic message will begin and the amount of mRNA that will be produced. The binding of RNA polymerase to the promoter often involves interactions between proteins that bind at the promoter site and proteins that bind at secondary regulatory sites. The initiation of transcription in eukaryotes is especially complex, involving the participation of many transcription factors, some of which may bind to DNA sites thousands of base pairs distant from the promoter itself.

Signals for stopping transcription are called terminators. These base sequences at the ends of genetic messages are recognized by RNA polymerase, often with the help of accessory protein factors. Termination in both prokaryotes and eukaryotes involves the recognition of specific base sequences on the mRNA copy of the genetic message. Termination is more complex in eukaryotes than in prokaryotes and involves processing of the end of the message, which usually includes the addition of a long string of 100 or more adenines, the so-called poly-A tail.

In eukaryotes, virtually all mRNAs are decorated at their ends with molecules that are attached after the basic message has been transcribed from the DNA of the gene. A methyl guanine is added at one end. This is called the cap and the end of the mRNA carrying this added guanine is called the capped end. The poly-A tail is added at the other end. The beginning of the message in eukaryotic mRNA is always close to the capped end.

Translational starts and stops

The sequence of bases in a messenger RNA contains signals that are recognized by the protein-synthesizing machinery as starts and stops for translation. All protein messages begin with a codon for methionine. In prokaryotes, a specific short sequence of bases on the mRNA is recognized by the ribosome as a ribosome-binding site (also called the Shine–Dalgarno sequence). The AUG methionine codon closest to the ribosome-binding site is used as the first codon in the polypeptide product of the mRNA. In eukaryotes, ribosomes simply bind to the capped end of the mRNA and drift along the molecule until they reach the first AUG, which they use as the 'start' codon.

The end of a protein is signalled by one of the three stop codons: UAA, UAG and UGA. The recognition of these codons by the translation apparatus requires the participation of specific protein release factors. A mutation that converts a codon within a gene to a 'stop' codon will result in the premature termination of polypeptide synthesis. Such mutations are called nonsense mutations. If the inserted 'stop' codon is UAA, the mutant is an ochre mutation. UAG and UGA are, respectively, amber and opal nonsense codons.

Genes with related functions often lie close to one another on the chromosome. This is true both for prokaryotes and eukaryotes, but the production of mRNAs in prokaryotes and eukaryotes differs in an important way. The mRNAs of prokaryotes are often polygenic, which means that they carry messages corresponding to two or more genes that lie side-by-side on the chromosome. In contrast, almost all eukaryotic mRNAs are monogenic.

The existence of polygenic mRNAs in prokaryotes raises an interesting question. Translation of the first gene on

such a message involves the binding of a ribosome to a ribosome-binding site, as described above. But how do the downstream genes get translated? Do they, too, have ribosome-binding sites? In most cases, the answer seems to be that genes downstream of the first gene do not require a ribosome-binding sequence as part of their 'start' signal. Evidently, ribosomes that begin at the first gene on an mRNA simply continue along the mRNA, releasing a protein when they encounter a 'stop' codon at the end of one gene and then 'scanning' the nearby bases for an AUG start codon, at which the ribosome begins the synthesis of the next protein.

Definition of a Cistron

The word 'cistron' was coined in 1956 by Seymour Benzer at a time when molecular genetics was still in its infancy. He used the term to identify a segment of a genome that is responsible for a single genetic 'function', as determined by the *cis-trans* complementation test (**Figure 3**). The underlying idea of the test is that two mutations might produce similar phenotypic effects in several different ways. First, the two mutations might alter the same gene and, consequently, affect the same enzymatically catalysed step in a biochemical pathway. Alternatively, the two mutations might affect genes that encode enzymes for different steps in a single biochemical pathway. A third possibility is that two mutations might block steps in two different biochemical pathways that converge.

It is important to realize that a gene is not a point on a chromosome, but a segment of a chromosome. Mutations within a single gene may occupy different sites (that is, different DNA bases) within the gene. Any two mutations within a single gene are said to be 'alleles' in the sense that they affect the same genetic function. Mutations at exactly the same site are called homoalleles; mutations at different sites within a gene are heteroalleles.

The *cis–trans* test determines whether two mutations, m_1 and m_2 affect the same function by comparing the phenotypes of two kinds of double heterozygote, shown in **Figure 3**. Each heterozygote carries two mutations. In the *cis* configuration, the two mutations are both on the same chromosome, and the wild-type alleles of the mutations are on the second chromosome. In the *trans* configuration, one mutation is on one chromosome and the second mutation is on the other. The *cis*-test is simply a control to guarantee that the wild-type (+ +) is dominant to both m_1 and m_2. We expect the phenotype of the *cis* heterozygote to be wild-type. In the *trans*-test, we expect the phenotype to be mutant if the mutations m_1 and m_2 affect the same genetic function and wild-type if they affect different genetic functions. Noncomplementary mutations, which affect the same function, are said to reside in the same cistron.

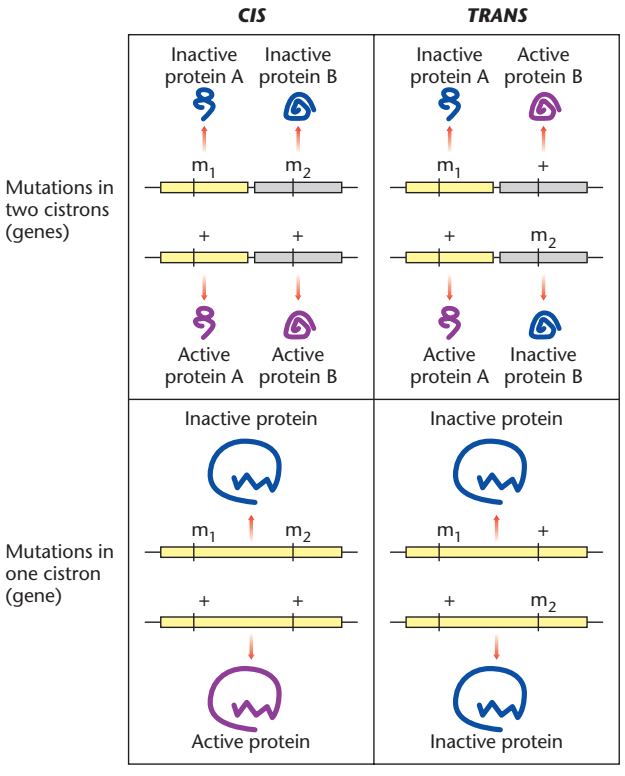

Figure 3 The *cis-trans* complementation test. A wild-type gene (+) makes a functional protein, which is shown in the figure as a purple line. A mutant gene (*m*) makes a nonfunctional protein, which is shown in the figure as a blue line. If the two mutations, m_1 and m_2, affect different genes (top frames), a heterozygous cell that contains both mutations and their wild-type (+) counterparts will contain some functional protein product from each gene. This will be true whether the mutations are in the *cis* configuration or the *trans* configuration. If the two mutations affect the same gene (bottom frames), then the *cis* heterozygote will make some functional protein and will have the wild-type phenotype, but the *trans* heterozygote will make only nonfunctional protein and will have the mutant phenotype.

Complementary mutations, which affect different functions, are said to reside in different cistrons.

When Benzer coined the term cistron in 1956, its use was restricted to describing the results of *cis–trans* complementation tests. Over time, the word has become almost synonymous with 'gene'.

Genes that do not make proteins

It should be noted in passing that the 'protein-synthesizing machinery' includes components composed of RNA. These components are the ribosomes, which are a complex of ribosomal RNA (rRNA) molecules and many small protein molecules, and transfer RNAs (tRNAs), which deliver amino acids to the nascent polypeptide and match the codons with appropriate amino acids through complementary base pairing. The rRNA and tRNA molecules

are transcribed from the DNA of the chromosome, but they do not contain a message that is translated into protein. They are, in fact, the final products of the genes that make them. Other untranslated RNA molecules are found, for example, in the small nuclear ribonucleoproteins (snRNPs) that are involved in the splicing of eukaryotic mRNAs and in the enzyme, telomerase, that maintains the linear ends of eukaryotic chromosomes. The base sequences of these RNA molecules are critical for their function, and mutations in the genes that encode these RNAs have phenotypic effects. As with any other mutations, recessive mutations in genes that encode RNA final products can be assigned to cistrons on the basis of *cis-trans* complementation tests.

ORFs and the Identification of Genes in DNA Sequences

A vast multinational effort to sequence the genomes of many organisms has produced a huge database containing the complete DNA sequences of a number of bacteria (both eubacteria and archaebacteria) and several simple eukaryotes (the yeast, *Saccharomyces cerevisiae*; the roundworm, *Caenorhabditis elegans*; and, soon, the fruit-fly, *Drosophila melanogaster*). In addition, we have the DNA sequences of the mitochondria of several eukaryotic species and large portions of the genomes of several complex eukaryotes, including mice and humans.

With the advent of DNA sequencing, the discovery of new genes no longer requires the identification of mutants with modified phenotypes. Genes can now be identified by computers that examine base sequences of DNA looking for a meaningful genetic message. Using the genetic code, the computer reads segments of DNA in all six possible reading frames (three reading frames on each of the two strands of DNA) and looks for segments that encode sequences of amino acids uninterrupted by 'stop' codons. On average, a random sequence of triplets will include a stop codon approximately once in every 21 codons. Therefore, if the computer finds a sequence of, say, 240 bases (80 triplets) that contains no stop codons, that sequence is likely to contain genuine genetic message. (A random DNA sequence of 240 bases containing equal amounts of A, T, C and G will lack a stop codon only 2.1% of the time.) A long sequence containing no 'stops' is called an open reading frame (ORF) and is considered likely to be a gene or a part of a gene.

Our confidence that a given ORF is actually a gene, or a part of a gene, can often be bolstered by comparing it to DNA sequences from other organisms. DNA sequences of genes tend to be conserved in evolution, whereas noncoding segments of DNA are free to accumulate mutations over time. Natural selection prevents rapid accumulation of mutations within genes, and homologies between genes are often recognizable in the DNA sequences of organisms whose ancestors diverged billions of years ago. Thus, comparative studies using the rapidly expanding database of DNA sequences from many different organisms are making it easy to identify the regions that correspond to genetic message.

Comparative studies of genetic messages have revealed another important fact about the evolution of proteins. Proteins often have multiple functions, and separate domains of the protein molecule carry out these functions. For example, one domain of a protein might recognize a binding site on DNA and a second might recognize a hormone that stimulates DNA binding. Comparative studies show that many multifunctional proteins have evolved by cobbling together old domains in new combinations and by modifying the specificities of the domains to serve new purposes. Consequently, ORFs corresponding to DNA-binding domains or steroid hormone-binding domains may appear in several different genes within a single organism and may show homology to ORFs in genes encoding proteins with widely varying functions in distantly related organisms.

Complexities raised by introns

Eukaryotic genes often contain introns – 'intervening sequences' of bases that are not part of the genetic message of the gene. The introns must be removed from the RNA transcript in order to make functional mRNA. The active mRNA is made up of the exons – the message proper of the gene – that are spliced together in the cell nucleus during the processing of primary transcript. Processed mRNA is then exported to the cytoplasm, where protein synthesis occurs in eukaryotic cells.

The existence of introns complicates the identification of genes. An ORF may be only part of a gene. The DNA sequence for the beginning of the gene, which will contain the start codon, the promoter, and, perhaps, other 'upstream regulatory elements', might be separated from a given ORF by several introns and exons. Similarly, the exon containing the end of the gene might be separated from a given ORF by several introns and exons. Since introns are not used to direct polypeptide synthesis, they can contain triplets that would be read as 'stop' codons in the reading frame of the surrounding gene. Furthermore, most of the sequence contained in an intron is free to accumulate random mutations over evolutionary time and therefore may bear little resemblance to the DNA of introns in homologous genes in distantly related organisms. The identification of the several exons belonging to a single gene requires the identification of the introns that separate those exons.

Recurrent themes help to identify both the genes themselves and nearby DNA sequences that play a role in regulating the activity of genes. The exons of homo-

logous genes in different organisms tend to be conserved (similar), whereas the introns tend to show rapid sequence divergence over evolutionary time. Thus, comparison of similar sequences from different organisms can often help to distinguish introns from exons. The beginnings and ends of introns tend always to be the same – the bases GU at one end and AG at the other – which helps to confirm the identity of introns.

Different protein encoding units within the same ORF

A single segment of DNA in the genome may encode amino acids in two or more distinguishable proteins. This can happen in several different ways. One mechanism, an example of which is found in the immune system, is a recombinational assortment of exons by breakage and rejoining of the DNA itself. The immune system generates a diversity of immunoglobulin proteins by assembling genes from a large collection of exons. The region that encodes the kappa light chain of human immunoglobulin, for example, contains about 300 'V' exons corresponding to one end of the molecule, the so-called variable region. There are five alternative J, or joining region, exons and one constant region exon. During the differentiation of the antibody-producing cells, the chromosomes undergo internal rearrangements so that one of the V exons becomes joined to one of the J exons. The result is a cell line in which the mRNA for kappa light chain always corresponds to a specific one of the 300 possible V regions, a specific one of the five possible J regions, and the C region. Thus, all kappa light chains have the same amino acid sequence in the constant region, but very different amino acid sequences in the variable region.

A second mechanism that can generate different proteins from a single ORF is alternative splicing. The RNA that serves as precursor to mRNA can be spliced differently in different cell lines so that a segment that is treated as an intron in one cell line (that is, it is removed from the final messenger RNA) is treated as an exon in another (that is, it becomes part of the final message). An interesting example of alternative splicing is found in a group of genes involved in determining the sex of the fruitfly, *Drosophila melanogaster*. The mRNA for the sex lethal (*Sxl*) gene is spliced differently in embryos destined to become females or males. The resultant protein product of the mRNA in females is functional, but the corresponding protein made in males is not. The difference between the two sexes turns on this one splicing difference.

Other mechanisms can lead to the production of different proteins from a single ORF. For example, alternative transcriptional starts can generate mRNAs that start at different places and lead to differences at the corresponding end of the protein product. Alternative termination can lead to variations at the other end of the protein. Several strange cases have been described in which bases in the mRNA itself can be 'edited' (replaced with other bases) so that the mRNA message does not match the base sequence in the DNA. Finally, posttranslational processing of the protein itself can lead to the production of different proteins from a single genetic message.

Overlapping reading frames

In rare cases, a single genetic message can be read in two different reading frames to produce two different functional gene products. The very small bacteriophage, ϕX174, which was the first completely sequenced genome, provided the first example of overlapping genes. Two of the proteins produced by ϕX174 come from DNA sequences that are completely contained within genes that produce other proteins. Because each of these nested sequences is translated in a different reading frame from that of the surrounding gene, each produces a protein whose amino acid sequence is totally different from the product of the surrounding gene. A third gene in the phage begins in one gene and ends in another and is read in a different reading frame from either of the genes that it overlaps. Other examples of overlapping genes have been found in viruses, including a mammalian virus, *Simian virus 40* (*SV40*). Although a few examples of gene overlaps have been found in the genes of eukaryotes, such arrangements appear to be very infrequent.

Further Reading

Alberts B, Bray D, Lewis J *et al*. (1994) *Molecular Biology of the Cell*, 3rd edn. New York: Garland.

Benzer S (1956) The elementary units of heredity. In: McElroy WD and Glass B (eds) *A Symposium on the Chemical Basis of Heredity*. Baltimore: The Johns Hopkins Press.

Cairns J, Stent GS, and Watson, JD (eds) (1992) *Phage and the Origins of Molecular Biology*. Plainview, NY: Cold Spring Harbor Laboratory Press.

Corwin HO and Jenkins JB (eds) (1976) *Conceptual Foundations of Genetics*: *Selected Readings*. Boston: Houghton Mifflin.

Griffiths AJF, Gelbart WM, Miller JH and Lewontin RC (1999) *Modern Genetic Analysis*. New York: WH Freeman and Company.

Judson HF (1996) *The Eighth Day of Creation*: *Makers of the Revolution in Biology*. Plainview, NY: Cold Spring Harbor Laboratory Press.

Lodish H, Baltimore D, Berk A *et al*. (1995) *Molecular Cell Biology*, 3rd edn. New York: Scientific American Books.

Snustad DP and Simmons MJ (2000) *Principles of Genetics*, 2nd edn. New York: John Wiley and Sons.

Stern C and Sherwood ER (eds) (1966) *The Origin of Genetics: A Mendel Source Book*. San Francisco: WH Freeman.

Sturtevant AH (1965) *A History of Genetics*. New York: Harper and Row.