ACCESS Science (http://accessscience.com/)
The Science Authority

Gene

Article by:

Sturtevant, Alfred H. Formerly, Biology Division, California Institute of Technology, Pasadena, California.

Winkler, Malcolm E. Department of Microbiology and Molecular Genetics, University of Texas, Houston, Texas.

Marmur, Julius Department of Biochemistry, Albert Einstein College of Medicine of Yeshiva University, New York, New York.

Publication year: 2014

DOI: http://dx.doi.org/10.1036/1097-8542.284400 (http://dx.doi.org/10.1036/1097-8542.284400)

Content

The basic unit in inheritance. There is no general agreement as to the exact usage of the term “gene” since several criteria that have been used for its definition have been shown not to be equivalent.

The nature of this difficulty will be indicated after a description of the earlier position. The facts of mendelian inheritance indicate the presence of discrete hereditary units that replicate at each cell division, producing remarkably exact copies of themselves, and that in some highly specific way determine the characteristics of the individuals that bear them. The evidence also shows that each of these units may at times mutate to give a new equally stable unit (called an allele), which has more or less similar but not identical effects on the characters of its bearers. *See also:* **Allele (/content/allele/024000)**; **Mendelism (/content/mendelism/414500)**

These hereditary units are the genes, and the criteria for the recognition that certain genes are alleles have been that they (1) arise from one another by a single mutation, (2) have similar effects on the characters of the organism, and (3) occupy the same locus in the chromosome. It has long been known that there were a few cases where these criteria did not give consistent results, but these were explained by special hypotheses in the individual cases. However, such cases have been found to be so numerous that they appear to be the rule rather than the exception. *See also:* **Chromosome (/content/chromosome/134900)**; **Mutation (/content/mutation/441200)**; **Recombination (genetics) (/content/recombination-genetics/575500)**

The term gene, or cistron, may be used to indicate a unit of function. The term is used to designate an area in a chromosome made up of subunits present in an unbroken unit to give their characteristic effect. It is probable that with increasing knowledge of the nature and properties of deoxyribonucleic acid (DNA) it will become possible to reach a more generally acceptable solution to the problems of terminology. *See also:* **Deoxyribonucleic acid (DNA) (/content/deoxyribonucleic-acid-dna/186500)**; **Nucleic acid (/content/nucleic-acid/460600)**

Alfred H. Sturtevant

Molecular biology

Every gene consists of a linear sequence of bases in a nucleic acid molecule. Genes are specified by the sequence of bases in DNA in prokaryotic, archaeal, and eukaryotic cells, and in DNA or ribonucleic acid (RNA) in prokaryotic or eukaryotic

viruses. The flow of genetic information from DNA to messenger RNA (mRNA) to protein is historically referred to as the central dogma of molecular biology; however, this view required modification with the discovery of retroviruses, whose genetic flow goes from RNA to DNA by reverse transcription and then to mRNA and proteins. The ultimate expressions of gene function are the formation of structural and regulatory RNA molecules and proteins. These macromolecules carry out the biochemical reactions and provide the structural elements that make up cells. *See also:* **Retrovirus (/content/retrovirus/584850)**

Flow of genetic information

The goal of molecular biology as it applies to genes is to understand the function, expression, and regulation of a gene in terms of its DNA or RNA sequence. The genetic information in genes that encode proteins is first transcribed from one strand of DNA into a complementary mRNA molecule by the action of the RNA polymerase enzyme. The genetic code is nearly universal for all prokaryotic, archaeal, and eukaryotic organisms. Many kinds of eukaryotic and a limited number of prokaryotic mRNA molecules are further processed by splicing, which removes intervening sequences called introns. In some eukaryotic mRNA molecules, certain bases are also changed posttranscriptionally by a process called RNA editing. The genetic code in the resulting mRNA molecules is translated into proteins with specific amino acid sequences by the action of the translation apparatus, consisting of transfer RNA (tRNA) molecules, ribosomes, and many other proteins. The genetic code in an mRNA molecule is the correspondence of three contiguous (triplet) bases, called a codon, to the common amino acids and translation stop signals (see **table**); the bases are adenine (A), uracil (U), guanine (G), and cytosine (C). There are 61 codons that specify the 20 common amino acids, and 3 codons that lead to translation stopping; hence the genetic code is degenerate, because certain amino acids are specified by more than one codon. *See also:* **Intron (/content/intron/350450)**

Genetic code, showing the correspondence between triplet codons in mRNA and the 20 common (L-isomer) amino acids inserted into polypeptides by the translation apparatus

First position in codon	Second position in codon				Third position in codon
	U	C	A	G	
U	Phenylalanine	Serine	Tyrosine	Cysteine	U
	Phenylalanine	Serine	Tyrosine	Cysteine	C
	Leucine	Serine	Translation stop	Translation stop	A
	Leucine	Serine	Translation stop	Tryptophan	G
C	Leucine	Proline	Histidine	Arginine	U
	Leucine	Proline	Histidine	Arginine	C
	Leucine	Proline	Glutamine	Arginine	A
	Leucine	Proline	Glutamine	Arginine	G
A	Isoleucine	Threonine	Asparagine	Serine	U
	Isoleucine	Threonine	Asparagine	Serine	C
	Isoleucine	Threonine	Lysine	Arginine	A
	Methionine	Threonine	Lysine	Arginine	G
G	Valine	Alanine	Aspartate	Glycine	U
	Valine	Alanine	Aspartate	Glycine	C
	Valine	Alanine	Glutamate	Glycine	A
	Valine	Alanine	Glutamate	Glycine	G

The sequence of amino acids in a protein is determined by the series of codons starting from a fixed translation initiation codon. AUG and GUG are the major translation start codons of prokaryotic genes. AUG is almost always the translation start codon of eukaryotic genes. The bacterial start AUG and GUG codons specify a modified form of methionine, whereas AUG or GUG codons internal to reading frames specify methionine or valine, respectively. *See also:* **Genetic code (/content/genetic-code/284900)**

The translation apparatus reads the next codon in the mRNA and attaches the specified amino acid onto methionine through a peptide bond. In most cases, this linear process of moving to the next codon and attaching the corresponding amino acid continues until one of the translation stop codons is encountered. Meanwhile, the nascent polypeptide chain folds by itself, or with the assistance of proteins called chaperones, into the functional protein. The biochemical rules that govern protein folding are often referred to as the second genetic code. In addition, some eukaryotic and prokaryotic proteins undergo protein splicing, which removes internal polypeptide segments called inteins. Some RNA transcripts are not translated; instead, they are cut and processed to form structural RNA molecules, such as tRNA and the three large RNA molecules associated with proteins in the ribosomes. *See also:* **Protein (/content/protein/550200)**; **Ribonucleic acid (RNA) (/content/ribonucleic-acid-rna/589000)**

Isolating genes

In many cases, only genes that mediate a specific cellular or viral function are isolated. The recombinant DNA methods used to isolate a gene vary widely depending on the experimental system, and genes from RNA genomes must be converted into a corresponding DNA molecule by biochemical manipulation using the enzyme reverse transcriptase. The isolation of the gene is referred to as cloning, and allows large quantities of DNA corresponding to a gene of interest to be isolated and manipulated.

After the gene is isolated, the sequence of the nucleotide bases can be determined. The goal of the large-scale Human Genome Project is to sequence all the genes of several model organisms and humans. The sequence of the region containing the gene can reveal numerous features. If a gene is thought to encode a protein molecule, the genetic code can be applied to the sequence of bases determined from the cloned DNA. The application of the genetic code is done automatically by computer programs, which can identify the sequence of contiguous amino acids of the protein molecule encoded by the gene. If the function of a gene is unknown, comparisons of its nucleic acid or predicted amino acid sequence with the contents of huge international databases can often identify genes or proteins with analogous or related functions. These databases contain all the known sequences from many prokaryotic, archaeal, and eukaryotic organisms. Putative regulatory and transcript-processing sites can also be identified by computer. These putative sites, called consensus sequences, have been shown to play roles in the regulation and expression of groups of prokaryotic, archaeal, or eukaryotic genes. However, computer predictions are just a guide and not a substitute for analyzing expression and regulation by direct experimentation. *See also:* **Human Genome (/content/human-genome/757575)**; **Molecular biology (/content/molecular-biology/430300)**

Prokaryotic gene

In eubacteria, cyanobacteria (blue-green algae), and many bacteriophages, the genetic material is double-stranded DNA. However, in some bacteriophages, the genetic material is single-stranded DNA or even RNA. As in other organisms, bacterial genes specify structural and regulatory RNA molecules, which do not encode proteins, or mRNA molecules, which do encode proteins. Some sites that play important cellular roles but are not copied into RNA molecules are also considered genes, such as the origin for bacterial chromosome replication. *See also:* **Bacteriophage (/content/bacteriophage/069900)**; **Cyanobacteria (/content/cyanobacteria/174950)**

Structure and expression

The arrangement of prokaryotic genes varies from simple to complex. Bacterial genes are delineated by sites in the DNA of the bacterial chromosome and in the RNA of transcripts. Transcribed genes start with promoters, which are binding sites of RNA polymerase. RNA polymerase melts the DNA near one end of each promoter region and then begins copying one of the DNA strands into an RNA molecule with a complementary sequence of bases. The starting end of bacterial RNA molecules initially contains a triphosphate group, which is not modified or capped as in eukaryotic cells. The first base is usually a purine

(adenine or guanine). As RNA polymerase adds nucleotides to a growing RNA chain, it moves down the DNA molecule in one direction only until it encounters a signal to terminate the transcription process. Promoters face in both directions around the bacterial chromosome, and the orientation within promoters directs RNA polymerase one way or the other. Bacterial genes do not usually overlap extensively, and only one DNA strand is transcribed in most regions of bacterial chromosomes.

Genes that specify structural RNA molecules, such as tRNA or ribosomal RNA (rRNA) required for protein synthesis, are initially transcribed into long precursor molecules. Mature structural RNA molecules are cut from these precursor transcripts by the concerted activities of specific ribonucleases, which are enzymes that break the chemical bonds in the phosphodiester backbone of RNA. For genes that specify proteins, the nascent mRNA is translated as soon as a ribosome-binding site clears the transcribing RNA polymerase. Simultaneous synthesis and translation of mRNA molecules is a fundamental property of prokaryotic cells, because they lack the nuclear boundary that separates transcription from translation in eukaryotic cells. Segments of mRNA between the transcript beginning and the first ribosome-binding site are called leader regions and are often shorter than 100 nucleotides; however, longer leaders are found that play roles in gene regulation.

The ribosome binding site consists of two parts. The translation start codon of bacterial genes is usually AUG or GUG, which specifies a modified form of the amino acid methionine that is often removed from the final protein product. However, not every AUG or GUG in a transcript directs ribosome binding. AUG and GUG start codons in mRNA are preceded by another short segment of nucleotides, called the Shine-Delgarno sequence. The bases in the Shine-Delgarno sequence pair with complementary bases in 16S rRNA molecules in the ribosome and properly position the start codon for translation. *See also:* **Ribosomes (/content/ribosomes/589200)**

Following initiation, the ribosome usually moves down the mRNA, reading one triplet (three-base) codon at a time. The amino acid corresponding to each codon in the genetic code is attached to the preceding amino acid by formation of a peptide bond. The ribosome continues synthesis of the polypeptide chain until one of the three translation stop codons is encountered, and the polypeptide and mRNA are released from the ribosome. The polypeptide chain is folded and sometimes binds to other folded polypeptides to form an enzymatically or structurally active protein. Some time after synthesizing the translation stop codon, RNA polymerase encounters a signal to stop transcription and to release the mRNA and DNA from the enzyme. Two kinds of transcription stop signals are used in bacteria. Factor-independent termination involves formation of a folded structure preceding a run of uracil residues in the nascent RNA chain. Factor-dependent termination involves interaction between a protein called Rho and RNA polymerase.

Expression level and regulation

Bacteria are ideally suited for survival as single cells in many environmental conditions. Part of this survivability involves controlling the expression of genes to optimize bacterial metabolism in response to environmental changes. For example, when a bacterium is presented with the amino acid tryptophan, a series of regulatory events are set off that turn off the genes encoding the enzymes that synthesize tryptophan. However, lack of tryptophan triggers the synthesis of these biosynthetic enzymes.

The expression level and regulation of a bacterial gene is influenced by five different processes. First, expression level depends on how frequently RNA polymerase transcribes a bacterial gene. The rate of transcription initiation depends on the relative intrinsic strength of a gene's promoter and whether transcription initiation is activated or repressed by additional protein regulatory factors that bind to DNA at or near a promoter in response to environmental factors. In addition, bacterial cells contain several kinds of RNA polymerase molecules that recognize different DNA consensus sequences as promoters in response to changing environmental conditions. Second, gene expression level depends on whether a transcribing RNA polymerase encounters a transcription termination signal called an attenuator that precedes the translated regions of certain genes. The frequency of termination at attenuator sites is sometimes controlled by environmental factors. Third, gene

expression level depends on the stability of an mRNA molecule, because the longer its lifetime, the more frequently an mRNA species can be translated. Generally, bacterial mRNA molecules have chemical half-lives of only about 2 min, which is short compared to most eukaryotic mRNA. Fourth, gene expression level depends on the efficiency of translation of a given mRNA molecule. This efficiency depends on the relative intrinsic strength of a ribosome binding site and whether access of ribosomes to a binding site is regulated by protein factors or folded structures in the mRNA transcript. Finally, gene expression level depends on the relative stability of the gene product RNA or protein molecules. Stable RNA and protein molecules will accumulate in cells, compared to ones that are rapidly degraded. *See also:* **Bacteria (/content/bacteria/068100)**; **Bacterial genetics (/content/bacterial-genetics/068700)**; **Prokaryotae (/content/prokaryotae/547750)**

Malcolm E. Winkler

Eukaryotic gene

Eukaryotic genes are arranged in a linear array on one set of chromosomes (haploid germ cells) or two sets of chromosomes (diploid somatic cells). There are about 100,000 genes in the mammalian genome, located within chromatin at specific sites in the nucleus. Eukaryotic organelles (such as mitochondria or chloroplasts) also contain genomes that encode a much smaller number of proteins.

Transcription

In eukaryotes, transcription and translation are compartmentalized, respectively, in the nucleus and in the cytoplasm. Eukaryotic genes are transcribed by three different RNA polymerases that use the ribonucleotide triphosphates as substrates and the DNA as the template: polymerase I transcribes the larger rRNA genes; polymerase II generates mRNA by transcribing genes with open reading frames that encode proteins (the enzyme also generates certain small, nuclear RNAs that complex with nuclear proteins and play important roles in splicing); polymerase III transcribes the genes for tRNA and small rRNA as well as those for other small nuclear RNAs of mostly unknown functions. The initiation of transcription of mRNA takes place in eukaryotic genes at multiple sites, usually 30–100 base pairs downstream from a short adenine/thymine–rich sequence referred to as the TATA box.

Eukaryotic genes are usually mosaics of coding (exons) and noncoding (introns) sequences. Since the pattern of removal of the introns (splicing) from the transcripts of a single gene may vary, a gene can encode several related proteins. When this happens, it is at variance with the long-standing one-gene–one-protein hypothesis. Mature, spliced mRNA (and in rare cases proteins) generates encoded proteins that lack a colinearity between the sequence of the DNA and its encoded protein. *See also:* **Exon (/content/exon/248350)**; **Transposable elements (/content/transposable-elements/706750)**

The polymerase II transcript is posttranscriptionally modified by capping of its 5′ end, by cleavage and polyadenylation of the 3′ end, and by splicing before the mature mRNA is transported to the cytoplasm to be translated.

Translation

Translation of eukaryotic mRNAs almost always initiates at the first ATG codon, probably controlled by ribosomes and associated initiation factors that scan the mRNA from its 5′ end. Prokaryotic protein initiation takes place at internal ATGs, depending on the location of the Shine-Delgarno sequence complementary to a rRNA sequence.

Operons have not been detected in eukaryotes. Genes that are not contiguous but coordinately regulated are members of regulons. Both eukaryotes and prokaryotes (for example, *GAL* in yeast and *arg* in *Escherichia coli*) have some of their genes organized in this manner. The expression of some genes is inducible, responding to specific changes in cell environment; others are constitutively expressed, at a uniform level. Genes that carry out specific functions may be expressed in a tissue-

specific manner and regulated developmentally.

The regulation of eukaryotic genes with open reading frames is more complex than that of prokaryotic genes. Several cis-acting upstream elements (called promoters) are involved; these serve as binding sites for multiple positive and negative trans-acting gene-specific and general transcription factors. These factors, together with cofactors, interact with polymerase II to regulate its activity. Also, the regulation of some eukaryotic genes may be controlled by chromatin structure and by DNA upstream or downstream enhancer sequences that act independently of their distance (some located up to several thousand base pairs away) and/or orientation. *See also:* **Eukaryotae (/content/eukaryotae/245250)**; **Genetic engineering (/content/genetic-engineering/285000)**; **Genetics (/content/genetics/285300)**

Julios Marmur

Bibliography

E. A. Birge, *Bacterial and Bacteriophage Genetics*, 5th ed., Springer, New York, 2006

M. Karin (ed.), *Gene Expression: General and Cell-Type-Specific*, Birkhauser, Boston, 1993

J. E. Krebs, E. S. Goldstein, and S. T. Kilpatrick, *Lewin's Genes X*, Jones and Bartlett, Sudbury, MA, 2011

S. R. Maloy, J. E. Cronan, Jr., and D. Freifelder, *Microbial Genetics*, 2d ed., Jones and Bartlett, Sudbury, MA, 1994

J. D. Watson et al., *Molecular Biology of the Gene*, 6th ed., Benjamin Cummings, Menlo Park, CA, 2007

Additional Readings

D. Hyatt et al., Prodigal: Prokaryotic gene recognition and translation initiation site identification, *BMC Bioinformatics*, 11(1):119, 2010 DOI: 10.1186/1471-2105-11-119 (http://dx.doi.org/10.1186/1471-2105-11-119)

J. Krebs, E. Goldstein, and S. T. Kilpatrick, *Lewin's Genes X*, Jones & Bartlett Publishers, Sudbury, MA, 2011

J. E. Richards and R. S. Hawley, *The Human Genome: A User's Guide*, 3d ed., Academic Press, London, UK, 2011