10

Viral Genes—Structure and Controls

Because of their minuteness and relative simplicity, the viruses afford insights into structural arrangements and activities that might long be overlooked in higher, more complex organisms. In many cases, however, these parasites have become degenerate in part by replacement of their original genes and translational products by those of the host cells. The resulting degree of dependency varies with the species, for many, including bacteriophage T4 among numerous others, expend considerable energy synthesizing macromolecules that duplicate or augment metabolic activities already present in its cellular habitat (Schmidt, 1985). By way of illustration, the genome of that T-even phage contains genes that encode one enzyme that cleaves the bacterial tRNAs near the anticodon and another pair that together repair the resulting damage, an altogether fruitless cycle. While useless to the virus, such relict genes are important from a biological point of view in suggesting that these organisms once were more completely supplied with genes and thus were less dependent upon living cellular types for existence. Hence, degeneration and evolutionary conservation have played antagonistic roles in molding the multitudinous diversity of extant viruses from their forebears of billions of years in the past.

10.1. VIRAL GENOMES

Nothing sheds light more brightly on the extent of diversification that has occurred among the viruses than the structure of their genomes, but aside from that, knowledge of the nucleic acid content and nature is a necessary preliminary to an understanding of the viral genes and their transcriptional controls. Viral genomes fall into four great classes, double- and single-stranded DNA and the same two classes of RNA, but variation on each of these themes is extensive. Since beginning with the more familiar macromolecular organizational type offers greater clarity, that procedure is followed here. Because of their involvement in cancer production, retroviruses, including the AIDS agent among others, are largely omitted from the present chapter, being reserved for the next, which is concerned with oncogenesis.

L. S. Dillon, *The Gene*

10.1.1. Bacterial Double-Stranded-DNA Viral Genomes

As the result of their position at the tip of the viral phylogenetic tree, the double-stranded-DNA viruses should be expected to contain highly complex genomes, and such is quickly found to be the actual case. Even within the bacteria, in which viruses are referred to as bacteriophages, quite a variety of major types occur. Although far from simple, these as a whole have not attained the complexity of structure that characterizes the vaccinia (cowpox) virus of metazoans, for example, so analysis of them first is logical. Additionally, and equally importantly, these organisms have been far more thoroughly investigated than any from eukaryotic cells.

The T-Even Bacteriophages. Two series of bacteriophages have designations beginning with T, one set having odd numbers, such as T1, T3, and T7, the other having even numbers, those referred to as T2 and T4 being especially important. Since beginning discussion with the latter, better known group affords distinct advantages, this frequent practice is adhered to here also. Like the bacterial genome, that of the T-evens is circularized, the ends of the molecule in the present instance having "terminal redundancy," repetitious sequences that by their complementarity enables them to adhere in catemeric fashion. These redundant ends vary in length from several hundreds to a few thousands of base pairs (Dillon, 1978, pp. 350–362). The DNA is highly modified by the conversion of its deoxycytidines to hydroxymethylcytidines, variously substituted with α- and β-glucose and gentibiose, a disaccharide; the extent and type of glucosylation are strongly correlated to the species of bacteriophage. As a group, these organisms have relatively immense genomes, being ~166,000 base pairs long in T4 (Gerald and Karam, 1984), encoding more than 60 genes.

Among the viruses as a unit, the genomic organization is based in great measure on the timing of transcription of several sectors. At the simplest level, there are two subdivisions, "early" and "late," the former referring to the region transcribed directly after the viral genome has entered the host cell, and the latter to that part transcribed after a viral RNA polymerase has been produced or replication of the genome has begun. Sometimes the first of these stages is modified by addition of other categories, such as "immediate early," which refers to genes transcribed immediately upon entry, followed by a second stage also involving host polymerase. At times the immediate early is called "early early," and remainder, "late early." Especially in types with large genomes, such as that of T4, a phase called the "middle" may exist between the early and late portions (Pulitzer *et al.*, 1985). In such cases, the late phase is postponed until replication of the nucleic acid structure commences, but these points and others related to them become clearer when the transcriptive processes are examined later in this chapter.

The T-Odd Bacteriophage Genomes. Because the full genome of one representative (T7) has been completely sequenced, the basic features of the T-odd phages can be discussed more satisfactorily than the T-evens. In that species the genome consists of 39,936 base pairs, which embrace 53 genes (Stahl and Zinn, 1981; Dunn and Studier, 1983; Moffatt *et al.*, 1984). Unlike that of the T-evens, their double-stranded DNA molecule is linear, for the ends are not complementary, a further difference being the absence of modified bases of any sort. This structure exemplifies clearly the rule that most viruses carry genetic information very efficiently, packing a maximum number of genes into a DNA molecule restricted in size by the capsid into which it fits. Very short spacers

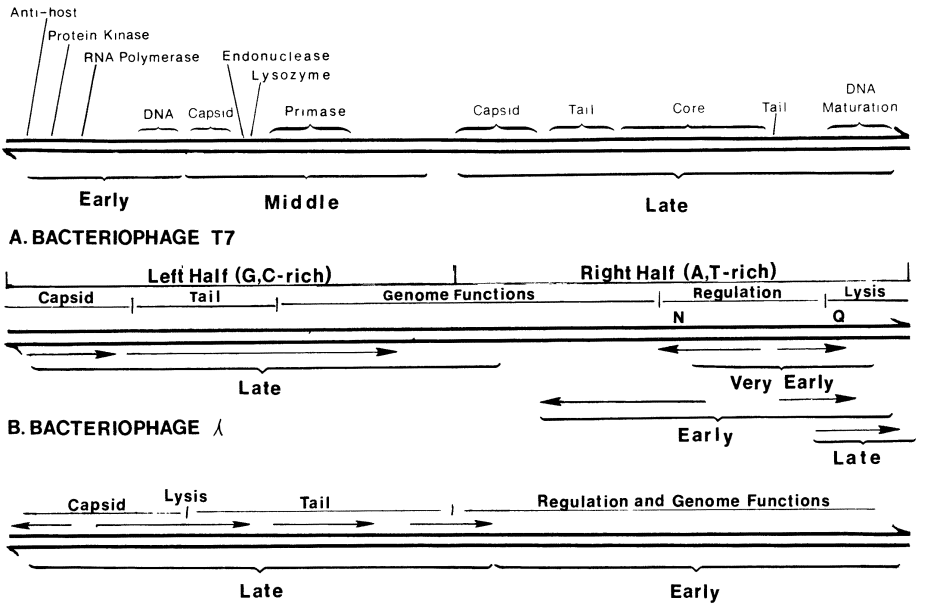


Figure 10.1. Structural arrangements of various double-stranded DNA bacteriophage genomes. (Part A is based on Studier and Dunn, 1983, and part C on Christie and Calender, 1985.)

intervene between the cistrons, as shown in detail in the discussion on transcription, so that 92% of the genome is used for coding purposes (Studier and Dunn, 1983). Although the DNA of T5 is about three times as long as that of T7, one of the smallest of the group, that of such T-evens as T4 is double that length, so the present organisms are decidedly the simpler in organization. This relative simplicity is reflected in morphological traits, for the tail of T-odd types lacks the contractility of those of the T-evens (Dillon, 1978, pp. 363–367).

The genome is divided into three major sectors (Figure 10.1A), early, middle, and late. In the first of these regions are ten genes, whose products inactivate host restriction processes, serve as viral RNA polymerase and protein kinases, or are active in replication. The middle portion is slightly the largest, carrying 22 cistrons, encoding DNA-binding protein, endo- and exonucleases, lysozyme, primase, and DNA polymerase, along with numerous unidentified products. In the late section are 21 genes, largely specifying proteins needed in construction of the capsid (referred to as the head in these viruses) and tail structures.

10.1.2. Temperate Double-Stranded-DNA Bacteriophage Genomes

Both of the preceding groups are said to be virulent, because they induce the immediate production of progeny, the accumulation of which quickly leads to the lysis of the host

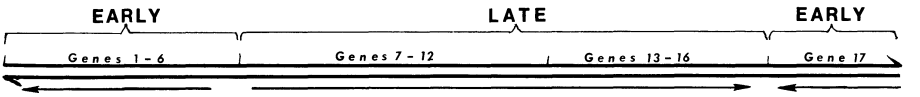
cell. There are numerous other types, however, that may be virulent as the foregoing ones, or they may, under certain largely unknown conditions, be "temperate," the viral genome becoming associated with that of the host and maintaining itself over long periods of time by replicating in unison with host DNA. Eventually, the phage genome, then known as a "prophage," commences to produce progeny, soon resulting in lysis as in virulent types. This delayed lysis is referred to as the "lysogenic response." It is this latter type of life history in which viral behavior resembles some of the transposons of the preceding chapter, a feature also of great significance in cancer-related forms.

Bacteriophage λ. Bacteriophage λ is one of the better known members of the temperate variety. As in the T-odds, it has a nonretractile tail, and the genome is linear and double-stranded, encoding 30–40 proteins. Among the most outstanding characteristics is that both strands are transcribed during the early stage, two subdivisions of which are recognized. Very early transcription, apparently from divergent promoters, results in the production of several enzymes, including protein N needed for the early substage. The transcription of the latter similarly proceeds bidirectionally (Figure 10.1B), generating a number of products needed for regulatory, lytic, and genomic functions, including protein Q, which is essential to late transcription. Whereas these two substages are confined to the right, A,T-rich half of the DNA, the late stage is confined to the G,C-enriched portion and is unidirectional (Figure 10.1B).

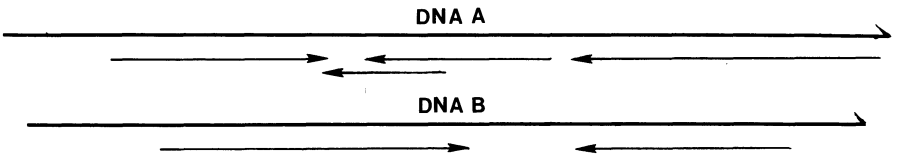
In part because of its possible relationship to comparable processes in oncogenic species, the mechanism of inserting the λ genome into that of the bacterium has been the center of much research. The processes of integrating the two genomes in the present instance, is, like many others, site-specific, taking place at definite points in each. The viral region of insertion, designated as *attP,* is 240 base pairs in length, while that of the host, called *attB,* consists of only 25 (Campbell, 1983; Griffith and Nash, 1985). In addition, two enzymes are involved in the reactions, an integrase (Int), encoded by the viral DNA, and IHF (Integration Host Factor), produced by the bacterium. Although precise data are lacking, the processes appear comparable to those of movable elements, excluding the replication of the element aspect, but including duplication of host DNA at the termini.

Other Temperate Double-Stranded DNA Phages. Although a number of additional types of temperate double-stranded DNA bacteriophages are known, most are not sufficiently investigated at the molecular level to merit space here. Among the better documented ones in this group are the several members of the P type, P1, P2, and P22 being especially important. The first of these tail-bearing forms contributes greatly to a later discussion; here the second, from *E. coli*, provides the insight into genomic organization. When illustrated in a reversed orientation as in Figure 10.1C, the DNA molecule of P2 somewhat resembles that of λ in arrangement, but fewer of the transcriptive activities take place from divergent promoters. Nor is a very early sector distinguished, there being only early and late subdivisions. The third in the series, P22, from *Salmonella,* has a genome still more similar to that of λ; it receives much attention in subsequent sections concerned with transcription.

Another type of bacteriophage is best represented by ϕ29, one of the better known phages infecting *B. subtilis*. Its DNA of ~18,000 base pairs is similarly linear and double-stranded, but it differs markedly from the preceding groups in having two sets of early genes, numbers 1–6 and 17 (Figure 10.2A). Moreover, all the early coding sectors are on



A. BACTERIOPHAGE ϕ29



B. TOMATO GOLDEN MOSAIC VIRUS

Figure 10.2. Genome organization of diverse DNA viruses. (A) Bacteriophage ϕ29 is one of the better known viral parasites of *Bacillus subtilis*. (After Holder and Whiteley, 1983.) (B) Like those of many plant viruses, the genome of the tomato golden mosaic virus consists of two separate parts, each single-stranded. (After Hamilton *et al.*, 1984.)

one strand and the 11 of the late section on the other, so that the transcriptive processes are strongly polarized. A close relative, M2, which also attacks this species of bacterium, receives considerable attention later.

10.2. MAMMALIAN DOUBLE-STRANDED DNA VIRUSES

Many of the viruses of mammals have genomes of double-stranded DNA, but the present discussion centers on five of the major families. These are the polyoma-, papova-, adeno-, herpes-, and poxviruses. The first three listed, being smaller and simpler, and therefore better known, receive the major portion of attention, not only here, but in subsequent sections, while the other two, despite their medical importance, sometimes are necessarily omitted altogether. In all of these groups, the DNA is in the form of a covalently closed, circular molecule twisted into a superhelix. The capsid never bears the tail universally present in the bacteriophages just considered, but is typically a skew icosahedral structure, consisting of 72 subdivisions called capsomers (Dillon, 1978, pp. 372–279).

10.2.1. Smaller DNA Viruses of Mammals

The papova- and polyomaviruses of primates rank among the most thoroughly understood forms from mammals, the simian virus 40 (SV40) representative of the second group being especially well documented. Accordingly, its genomic characteristics are given first to provide a basis for comparison.

The SV40 Genome. The genome of SV40 is relatively small, being only ~10% that of bacteriophage λ; its 5200 base pairs are sufficient to encode about ten proteins of molecular weight of 20,000 each. Apparently, the only modified nucleotide present is 5-

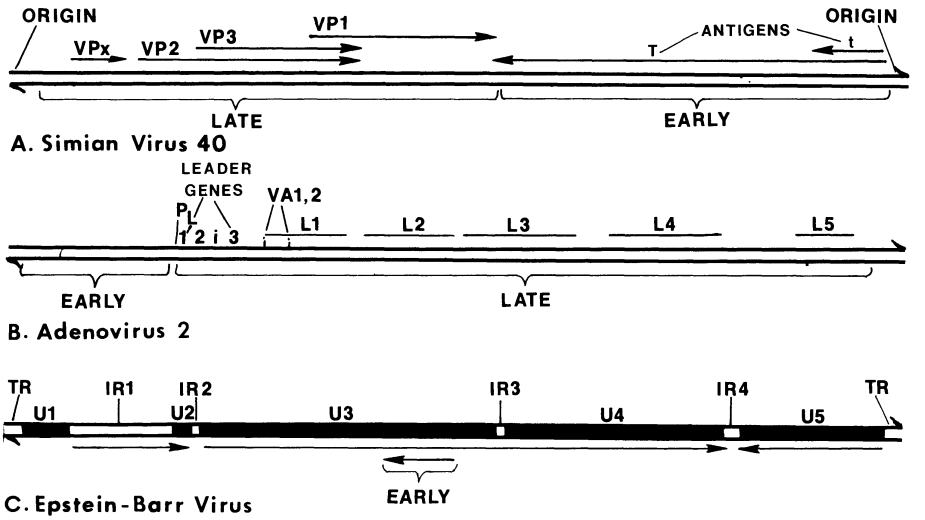


Figure 10.3. Genomes of double-stranded DNA viruses of mammals. (A) Early genes of SV40 are on the lower strand of DNA, the late ones on the upper. (Yang and Wu, 1979.) (B) Direction of transcription has not been clearly established for the genes of adenovirus 2 (P_L, major late promoter). (Based on Virtanen *et al.*, 1982.) (C) The genome of herpesvirus EBV is divided into five "unique" regions (U1–U5) and four internal repeated sequences (IR1–IR4). At the ends are terminal repeats (TR). (Based on Bodescot *et al.*, 1984; Gibson *et al.*, 1984; and Thomsen *et al.*, 1984.)

methylcytosine, and that only in small quantity. While within the virion, the DNA has a chromatinlike structure, since several species of protein are associated with that macromolecule. At least three of them resemble histones, but whether they are produced by the host or virus remains an unsettled issue.

Like most others from mammals, this virus is temperate, being able to insert its DNA into the host genome for an indefinite period of time. A similar nature enables a related species called the BK virus (BKV) of man to induce transformation of the inhabited cells, resulting in tumor formation. Both species have genomes of virtually identical size and are ~70% homologous, the sequence of each having been completely established (Fiers *et al.*, 1978; Reddy *et al.*, 1978; Seif *et al.*, 1979; Yang and Wu, 1979). The genes are arranged in two parts; those transcribed in the early stage are read in a counterclockwise fashion (right to left in Figure 10.3A), those of the late stage in the opposite direction. Since the two genes encoded in the early region are for antigens and the four of the late period are not known to code for any polymerases, these viruses must be dependent upon host products for transcription and DNA replication.

The Adenovirus Genome. The genome of adenoviruses, more than 50 types of which have been isolated from mammalian and avian sources, is about six times as large as that of SV40, the DNA molecule from type V (Ad5) consisting of 36,000 base pairs (Dekker and van Ormondt, 1984). Accordingly, the regular icosahedral capsid has far greater dimensions, being constructed of 252 capsomers of two types. The greater fraction

of these consists of unmodified simple structures called hexons, whereas the pentons, located one at each of the 12 angles, bear long, projecting fibers having a knob at their ends.

As shown in Figure 10.3B, the genome is subdivided into early and late regions, the latter occupying by far the greater portion. The products of the early phase have not been completely characterized, but five multicistronic mRNAs are known to be generated in the late period of type II adenovirus (Ad2), one of the more thoroughly analyzed members of the group (Virtanen *et al.*, 1982). Among the products encoded here is VA1, transcribed by host RNA polymerase III (Chapter 3, Section 3.4.2; Aleström *et al.*, 1982). That substance is essential for translation of viral late mRNAs, largely through its role in transport of those molecules from the nucleus into the cytoplasm (Katze *et al.*, 1984). Probably the most distinctive feature of this group of viruses is in the modifications that occur to the messengers prior to translation. About the most peculiar is the attachment of a leader sector to the 5′ ends of mRNAs, a segment generated from three exons by elimination of three introns (Keohavong *et al.*, 1982). However, sometimes additional parts are spliced onto the leader, the i-leader being one of the more frequent additives (Chow *et al.*, 1979).

10.2.2. Larger Double-Stranded DNA Viruses of Vertebrates

Size is not the sole characteristic that distinguishes the large double-stranded DNA viruses of vertebrates from the smaller ones of the preceding section. Outstanding among the distinctions is the presence of lipids in an envelope that encases the capsid, and, second, the genomic DNA is linear rather than circular. Among the better documented of the group are the herpes- and poxviruses, the former being a complex group.

The Herpesvirus Genome. In the herpesviruses, including herpes virus 1 and 2 (HSV1 and 2) and the Epstein–Barr (EBV) virus, the genomic DNA, although linear and double-stranded, is in the form of a toroid, surrounding a central core of protein (Dillon, 1978, pp. 376–379). That of EBV, the causative agent of infectious mononucleosis, is somewhat larger than those of HSV1 and HSV2, consisting of 172,000 base pairs, compared to the 150,000 of the other two (Gibson *et al.*, 1984). In broad terms, in each case the DNA is divided into five regions containing unique sequences (U1–U5) by four internal regions of repetitious sequences (IR1–IR4) and two terminal repeats (TR) (Figure 10.3C; Bodescot *et al.*, 1984). Insofar as is established, transcription during the late stage is by way of three polycistronic messengers, one of which is oriented in opposition to the others (Figure 10.3C). Some of the internal and terminal repeats are themselves comprised of short, reiterated segments (Jones and Griffin, 1983; Costa *et al.*, 1985), and certain ones have been compared with the switch region of immunoglobulin genes (Gomez-Marquez *et al.*, 1985).

That pattern is obviously simplistic, because it has been demonstrated that two early genes occupy at least a portion of the U3 region and therefore cannot be part of the polycistronic messenger shown (Gibson *et al.*, 1984). In an additional member of the class, human cytomegalovirus, which has a genome of 240,000 base pairs, an early region is placed similarly to that shown, confirming the above statement (Thomsen *et al.*, 1984). In that virus, transcription of this early sector is oriented in reverse direction, as indicated in the EBV genome (Figure 10.3C).

The Poxvirus Genome. The vaccinia virus, which is the prototype of the poxvirus class, has a genome slightly exceeding that of EBV in size, containing ~187,000 base pairs (Baroudy *et al.*, 1982). In part the two termini of each strand are complementary, so that the linear double-stranded DNA can fold into a single, continuous polynucleotide chain. These terminal repeated sectors, some 10,000 base pairs long, contain at least three genes that are transcribed early in infection (Blomquist *et al.*, 1984). Among the unique features of transcription is that the immediate early activities take place within the capsid, but the later ones occur within the cytoplasm of the host (Cochran *et al.*, 1985). No extensive mapping of the genome has been conducted, so the arrangement of the ~100 genes it includes awaits clarification. Sequencing of a region to the left of center indicates that some of the early and late genes are tightly clustered, with many overlaps (Plucienniczak *et al.*, 1985).

10.3. MISCELLANEOUS DNA VIRUSES

By far the majority of DNA viruses are members of the several double-stranded groups just described from bacteria and mammals, but there are several others that also merit attention. Two assemblies of species with this type of nucleic acid for their genomes are important plant pathogens, and another of much larger proportions has strong impact on the bacterial world.

10.3.1. Some Plant DNA-Viral Genomes

The pair of major DNA viral families infecting plants just mentioned are the caulimoviruses and geminiviruses, the former double- and the latter typically single-stranded; however, since it is known to have double-stranded varieties (Hamilton *et al.*, 1984), the latter family appears to be a transitional form.

The Caulimoviral Genome. The caulimoviruses are best represented by the prototype of the group, the cauliflower mosaic virus. Its circular double-stranded genome is 8024 base pairs in length, the sequences of three strains of which have been fully established (Franck *et al.*, 1980; Gardner *et al.*, 1981; Bâlazs *et al.*, 1982). All eight of its coding areas are contained in the minus strand, mostly closely spaced, but an intergenic spacer 700 base pairs long separates genes *VI* and *VII* (Dixon and Hohn, 1984). Depending upon the strain, either two or three breaks are present per genome, one in the minus, the other(s) in the plus; these represent regions of single-strand overlap and may have application in transcription. Replication appears to involve an RNA intermediate transcript followed by reverse transcription (Pfeiffer and Hohn, 1983).

The Geminiviral Genome. Among the better known of the geminivirus family is the tomato golden mosaic virus (TGMV), whose genomic sequence has now been established, as has that of the African cassava mosaic virus (Hamilton *et al.*, 1982, 1984). In each case the genome consists of two single strands of DNA, the B component of which is just slightly shorter than the A (Figure 10.2B). In TGMV the latter consists of 2588 and the former of 2508 nucleotide residues. Four open reading frames, including one of reversed polarity, have been identified in DNA A and two convergent ones in DNA B (Figure 10.2B).

10.3.2. Single-Stranded DNA Viruses

The single-stranded DNA viruses are a relatively small group, confined almost entirely to the prokaryotes, except a few from plants, one of whose genomes was just described above. Those of bacteria fall into two major classes based on the shape of the capsid. In the icosahedral (spherical) forms, the more important representatives include ΦX174, S13, and Φ1 (Dillon, 1978, pp. 379–384), whereas in the filamentous category are placed fd, f1, M13, AE2, IKe, and Pf. Basically, the genomic characteristics of the two subdivisions are quite similar, but each possesses distinctive features of its own.

The Icosahedral Class Genome. The single-stranded DNA of the icosahedral class is in the form of a circle, the nucleotides being unmodified except for a single 5-methylcytosine per strand. In stage I immediately following infection, the genome is converted to a double-stranded circle by synthesis of a complementary (minus) strand. This "replicative form" (RF) of the molecule is present in two structural varieties, one of which is supercoiled and has both strands covalently closed, the other relaxed and with at least one single-stranded break.

ΦX174 serves admirably as a representative of genomic organization in this family, for its complete sequence has been established (Sanger *et al.*, 1978). In its single strand of DNA of 5386 residues, ten genes are found, all for proteins, many of which overlap one another, as discussed later. Preceding the first coding sector, that for protein A, is a spacer of 63 nucleotides, another of 39 sites intervenes between genes *J* and *F*, a third, 110 residues long, separates *F* and *G*, and a final one of eight sites lies between *G* and *H*. All the other cistrons either overlap or are separated by only one or two bases.

The Genome of the Filamentous Class. In these long, slender, filamentous types, whose dimensions range between 8000 and 20,000 Å in length and 50 and 60 Å in width, the genome is similar in form to those of the icosahedral types. Upon entering the bacterium by way of the male sex-pili, the single (plus) strand at once is copied into a minus strand to produce the RF molecule. Since the genomes of at least four members have been completely sequenced, those of f1, M13, fd, and IKe, their structures and gene arrangements are well known (Beck *et al.*, 1978; van Wezenbeek *et al.*, 1980: Hill and Petersen, 1982; Peeters *et al.*, 1985). These show a remarkable degree of constancy, both in size and arrangement. The shortest, that of fd, is 5781 base pairs in length, while the largest, from IKe, is 6883; Figure 10.4A–D brings out the similarities of gene arrangement. There, the intergenic region (IR) is uniformly placed to the left on an arbitrary basis, its relative position in these straight-line diagrams, whereas the DNA is circular. In each case, the order of the genes and direction of transcription are consistently the same for the four.

10.4. SINGLE-STRANDED RNA VIRUSES

The RNA viruses, attacking as they do a far greater portion of the living world than their DNA relatives, greatly outnumber the latter in species and show a much broader spectrum of diversity. Like them, however, they fall into two major categories, single- and double-stranded. Beginning with the first of these offers many advantages, because it is not only the larger class, but also much the better investigated, which reason provides the basis for viewing the bacteriophages first.

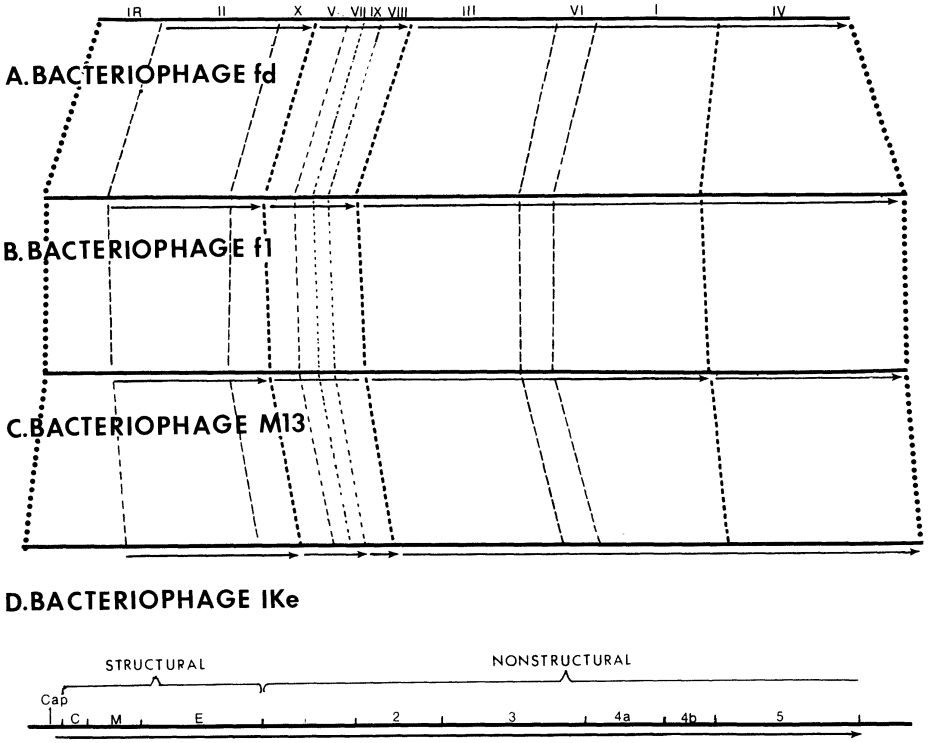


Figure 10.4. Genomic arrangements of four bacterial and one human double-stranded RNA viruses. (A–D) All four of these bacterial filamentous types are basically alike. (Part A is based on Beck *et al.*, 1978; Part B on Hill and Petersen, 1982; and parts C and D on Peeters *et al.*, 1985.) None of the transcriptional units have been fully established, being apparent only.

10.4.1. Genomes of Single-Stranded RNA Bacteriophages

Although four major groups have been erected for single-stranded RNA bacteriophages, only two have representatives sufficiently well characterized to justify their inclusion here. Group I contains MS2, R17, and f2 as principal members, while group III has Qβ as the sole important representative. In all cases the genome is small, consisting of between 3500 and 4500 nucleotides, and is enclosed in an icosahedral capsid constructed of 180 capsomers, plus one or more molecules of a second polypeptide, called either the A- or maturation-protein (Dillon, 1978, pp. 384–387). Before the more detailed genomic structures are discussed, it is of interest to note that comparisons of the 3′-terminal regions from 16 species representing all four groups (Inokuchi *et al.*, 1982) disclosed homology levels between members of groups I and II to be at 50–60% and of groups III and IV to be at ~50%. But only low degrees of kinships were displayed by representatives from either I or II with any from III or IV.

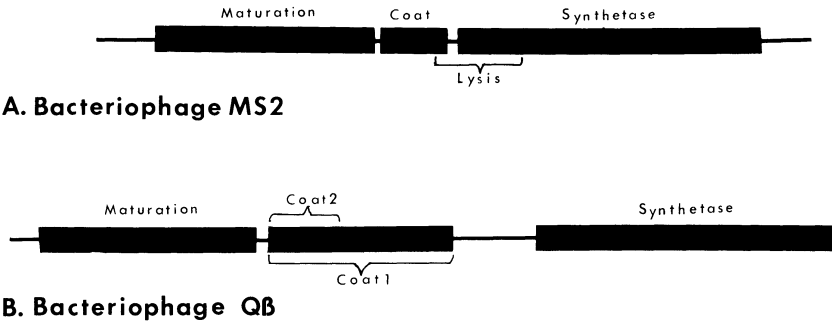


Figure 10.5. Genome structure in two important single-stranded RNA bacteriophages. (Based on Atkins *et al.*, 1979.)

Group I RNA Bacteriophage Genomes. Because the genomes of members of group I are all closely allied, differing at most by less than 4% (Min Jou and Fiers, 1976), attention can be confined entirely to the best known member, MS2. The sequence of the RNA of this form was among the very first messenger species to become completely established and the intricacies of its secondary structure revealed (Fiers *et al.*, 1975; Dillon, 1978, pp. 124–131; Atkins *et al.*, 1979). Within the 3569-nucleotide molecule are sequences of four genes, one each for the A-protein, the coat protein, an overlapping one for lysis, and the β subunit of the replicase (RNA-dependent RNA polymerase), the other three subunits of which are host products. Short intergenic spacers separate the cistrons, and there are untranslated sectors at each end (Figure 10.5A). Consequently, it is evident that this virus is highly dependent on the bacterium in which it lives for most of its requirements for transcription, translation, replication, and assembly.

Group III RNA Bacteriophage Genomes. Group III single-stranded RNA bacteriophages have not been such popular models for investigations at the molecular level as those of group I, and much of what research has been conducted has centered on the nature of the replicase (Mills *et al.*, 1978; Nishihara *et al.*, 1983). Consequently, relatively little has been established concerning the structural details of the genome. That of the chief representative, Qβ, has an RNA molecule 4220 nucleotides long, which is the minus strand, not the plus as in preceding types (Figure 10.5B).

The reason for the interest displayed in the replicase is its highly selective nature. It promotes the synthesis of the genomes of Qβ and other group III RNA bacteriophages, but not that of any other viral type. Nor can it replicate any RNA strands from *E. coli* (Nishihara *et al.*, 1983). Involved in this specificity are its strict requirements for template structure; only a single internal sequence of the Qβ RNA is recognized by the replicase and then only if the genomic 3′-terminal sequence is intact (Meyer *et al.*, 1981).

10.4.2. Single-Stranded RNA Viruses of Metazoans

A great diversity of types of single-stranded RNA viruses from metazoans has been explored at a satisfactory level, far too many to enumerate in detail here. Thus in the more important groups, each of several forms whose genomes have received adequate investi-

gation are described, the families being selected to give a view of the major variations of structure that exist.

The Picornavirus Genome. There can be little doubt that the picornavirus family deserves prominent position in this analysis, for its members include rhino- (cold-producing agents), cardio- (encephalomyocarditis and Mengo), aptho- (foot and mouth disease), and enteroviruses (polio and hepatitis A agents). In all these types, the RNA is the plus strand, serving as the polycistronic messenger for all viral structural and enzymatic proteins. The complete genome sequence of one of the first group, rhinovirus 14 (about 115 species have been identified), is 7212 nucleotides in length, of which ~6540 serve as the single reading frame (Stanway *et al.*, 1984; Callahan *et al.*, 1985). That of hepatitis A is a little longer, containing 7478 nucleotides, of which 6682 provide the single coding sector (Najarian *et al.*, 1985), and that of foot-and-mouth disease agent, with ~8500, is longer still (Carroll *et al.*, 1984). In each case three proteins are encoded, whose natures have not been determined.

The genomes of the three known serotypes of poliovirus have now had their sequences established. Their lengths of ~7500 nucleotides scarcely exceed that of rhinovirus, but their coding properties are more complex (Toyoda *et al.*, 1984). As there, much of the entire molecule serves as a single reading frame, but the resulting polyprotein is cleaved into nine polypeptides. Attached at the 5′ end of the genomic RNA is a covalently bonded protein of viral origin, while at the 3′ end is a poly(A) tail, which is apparently an encoded feature, not posttranscriptionally added (Nomoto *et al.*, 1982).

Despite the similarity of structure, little homology exists between the major representative types, the greatest kinship (44–65%) being displayed between the several proteins of rhino- and polioviruses (Callahan *et al.*, 1985), that is, between members of the same group.

The Flavivirus Genome. The flavivirus family includes more than 70 closely related pathogens of man and domestic animals, most of which are transmitted by blood-sucking insects or ticks. In it are included the agents for Japanese, St. Louis, and Murray Valley encephalitis, dengue, and yellow fever, the viral genome for the last of which has been completely sequenced (Rice *et al.*, 1985). Its single plus strand has proven to be considerably larger than any of the foregoing, containing 10,862 nucleotide residues and encoding nine identified genes. The first three of these are for structural proteins, the remainder for proteases and other nonstructural products (Figure 10.4E).

The Rhabdovirus Genome. The rhabdoviruses, too, have a single-stranded RNA genome, but here the nucleic acid molecule is the minus strand, needing to be transcribed to produce mRNAs. In vesicular stomatitis virus (VSV), the most familiar member of this group, the genome encodes six proteins, including a gene in the leader sequence that contains only 47 bases (Emerson, 1982). Because only a single promoter is present, located before the leader, transcription is polar and sequential, the more terminal cistrons being transcribed less frequently than those near the start site. Like all minus-stranded RNA viruses of mammals, the genome of VSV is encapsidated in ribonucleoprotein particles to form nucleocapsids, which serve as templates for both transcription and replication (Arnheiter *et al.*, 1985).

Larger RNA Genomes. In the larger RNA virus family known as the paramyxoviruses, the molecule is the minus strand as in the last group noted. It is best represented by the Sendai virus, whose 15,000-residue genome encodes seven structural proteins

employed in the capsid, plus an undetermined number of nonstructural ones (Giorgi *et al.*, 1983). Two of that number, P and C, have been demonstrated by sequencing to be largely overlapping, as detailed later, and that for the complex hemagglutinin-neuraminidase has been sequenced (Miura *et al.*, 1985). Here many of the cistrons are provided with individual promoters.

In many other larger RNA viral families of metazoans, including the coronaviruses, represented by the avian infectious bronchitis virus, the genome is similar, but larger, running to 20,000 residues in the species cited (Boursnell and Brown, 1984). Still others, including the influenza A virus, have a genome comprised of multiple segments of RNA, eight in the example mentioned (Lamb *et al.*, 1985).

10.4.3. RNA Viruses of Seed Plants

A rather large variety of viruses from seed plants has been recognized, all of which have single-stranded RNA as the genomic molecule. By far the greater part of these have the genome divided into multiple strands as in the influenza virus just mentioned, but the structure of the 3′ end provides a basis for recognition of three intergrading types, with a fourth group sharply distinguished from the rest. In type I, the RNA molecule (plus strand) terminates in a poly(A) tail, thus showing strong resemblance to the mRNAs of cellular organisms (Kozlov *et al.*, 1984), as well as to poliovirus. This type is represented by the como-, nepo-, poty-, and potexviruses. In type II, the RNA termini have a tRNA-like structure, as among the tymo-, tobamo-, bromo-, and cucumoviruses; in type III, the termini are devoid of both of those additions. Type IV, represented especially by the tobacco mosaic virus (TMV) family, has a unified circular single strand of RNA as the genome. Because the last of these groupings contains features of the others in a relatively simple form, discussion can expediently employ it as the model.

Type IV Viral Genomes. The type IV viral genome is adequately represented by that of the well-known TMV, whose sequence has been completely established (Goelet *et al.*, 1982; Nishiguchi *et al.*, 1985). It is a single-stranded circle of 6395 nucleotides, appearing to encode six proteins, at least four of which are known (Takamatsu *et al.*, 1983; Watanabe *et al.*, 1984). Some of the genes partly overlap others and the first cistron encodes two, the longer one resulting from readthrough of an amber stop (UAG) codon that terminates transcription of the shorter. In some strains at least, a tRNA is present at the 3′ terminus (Lamy *et al.*, 1975; Dillon, 1978, pp. 346–348), as in type II members. In this case, the structure accepts histidine.

Another type IV species, turnip yellow mosaic virus (TYMV), has a genome that similarly consists of a single molecule of single-stranded RNA, the plus strand, but in addition shows a feature characteristic of members of other groups. Here, as in several additional types, the 3′-terminal region of RNA encodes the coat protein, but never gives rise to that protein directly. Instead, an RNA molecule, referred to as "subgenomic RNA," is derived from that part by unknown processes. Thus, in TYMV there is a single genomic RNA and a subgenomic variety (Morch *et al.*, 1982).

Type I Viral Genomes. Among the better known viral genomes of type I is that of the cowpea mosaic virus (Stanley and Van Kammen, 1979), a member of the comovirus family. Its single-stranded RNA is subdivided into two parts, each of which has a poly(A) sector and serves as a messenger (J. W. Davies *et al.*, 1979). Moreover, both bear

proteins at the 5′ termini as in the picornaviruses of metazoans and in others of type I. To determine whether the poly(A) sector is preceded by an AAUAAA polyadenylation signal, as in cellular forms, the 3′ region has been sequenced, but no such signal proved to be present (J. W. Davies *et al.*, 1979). Nor should any have been suspected, for, as in polioviruses, the poly(A) here is a permanent part of the genomic structure, not a tail that is added by enzymes following transcription. Analyses and nucleotidyl structural comparisons have revealed homology to exist between comoviral RNA-dependent RNA polymerase and that of such picornaviruses as the poliovirus (Franssen *et al.*, 1984), which is likely kin to this group.

Type II Viral Genomes. As a whole the type II viruses have received relatively little attention at the molecular level; one of the few genomes from this source that has been adequately investigated is that of the barley stripe mosaic virus. In this species, the RNA is subdivided into two parts, each of which, as characteristic of this type, bears a tRNA at its 3′ end (Kozlov *et al.*, 1984) that accepts tyrosine. Unexpectedly, however, this virus intergrades with type I members in that the genomic RNAs bear poly(A) sequences. The A-rich sector is in the form of an oligo(A) tract located about 235 residues from the 3′ end, which is part of the $tRNA^{Tyr}$.

Type III Viral Genomes. Among the type III viral genomes, sequences have been established to a greater or lesser extent from at least three species, the alfalfa, brome, and cucumber mosaic viruses. In all of these the organization of the RNA is identical, there being three genomic strands, plus a subgenomic one. The sequences of all the parts have been established for the alfalfa and brome mosaic viruses and have proven to be quite similar, except in size. RNA1, the largest of the tripartite structures, is 3644 nucleotides in length in the former species, against 3234 in the second (Cornelissen *et al.*, 1983a; Ahlquist *et al.*, 1984), whereas RNA2 consists of 2593 and 2865, respectively (Cornelissen *et al.*, 1983b). The first structure has been sequenced from cucumber mosaic virus and was shown to be 3389 residues long (Rezaian *et al.*, 1985). In each case the nucleic acid encodes a single protein. The smallest, RNA3, however, is dicistronic, but the second (3′) portion of 881 residues is not translated, serving only as the source for the subgenomic RNA4, as in TYMV described earlier (Brederode *et al.*, 1980). As there, the latter codes for the coat protein. In another member of the group, brome mosaic virus, which infects grasses, the genome is similarly constructed, but the three genomic RNA molecules are larger, containing in order 3300, 3000, and 2100 bases, with RNA1 the largest (Ahlquist *et al.*, 1981).

10.5. TRANSCRIPTION SIGNALS IN VIRUSES

The complex diversity of viral genomes does not cease with their overall organization, infinitely varied though it is, but extends into the details of their gene structures. Whole books would be required to cover the characteristics of the cistrons of each viral type, despite the limitations of present knowledge. Hence, the purpose now is to provide samples through the viral world to give the flavor of the subject. First, details pertaining to transcriptional initiation and termination are provided, followed by such other peculiarities as the overlapping of genes.

10.5.1. Promoters of T-Series Bacteriophage Genes

As may well be expected, the bacteriophages are outstandingly more thoroughly explored in the matter of gene structure than is any other type of virus, so here, as in so many other aspects of viral morphology, function, and evolution, they provide the firmest basis for introducing the present topic. Also not unexpectedly, the T-series of these bacterial parasites lead the others in the thoroughness of coverage by investigators, and consequently provide the basic model.

Early Promoters of T-Series Bacteriophages. Since the early region of the genomes of the T-series bacteriophages is transcribed by the host (*E. coli*) polymerase, the translation signals should be expected to resemble those structures of bacteria. And so that of T7 proves to do—but not in the usual sense! What it does resemble of the *E. coli* TA-TA- box (the −10 sequence) is in its deviation from that "consensus" (Table 10.1), as earlier chapters disclosed in the bacterial genome. One of the three more active promoters, A1, and the less responsive farther upstream, A0, are quite similar to one another in structure, but their shared nucleotidyl sequence, AT-CT-A, shows little in common with the standard bacterial sequence (Dunn and Studier, 1983). In contrast, the other two active T7 early promoters, A2 and A3, along with that of gene *63* of T4, display somewhat greater resemblance to the accepted TA-TA- structure (Pribnow, 1975; Dunn and Studier, 1983; Rand and Gait, 1984). What is of particular interest, however, is that the first two promoters of T7 share so many homologous sites between themselves (Table 10.1), but not with the last two, the reverse of which statement is likewise true. Another point of kinship, previously unnoticed, is the standard translational stop signal TAA located upstream of each promoter, the significance of which is unknown.

Mid- and Late-Stage Promoters of T7. In six of the seven midstage sequences of T7 (Table 10.2) the combination CGACTCA is consistently found at the −10 region,

Table 10.1
Early Promoters of T-Series Bacteriophages[a]

				Promoter		Start Site	
T7 Early[b]							
A0	ACCTCC	TAA	CGTCC---	ATCCTAA	AGCCAA	C	ACC
A1	AAAGTC	TAA	CCTATAGG	ATACTTA	CAGCCA	T	CGA
A2	ATGAAG	TAA	CATGCAG-	TAAGATA	CAAATC	G	CTA
A3	ATGAAG	TAA	ACACGG--	TACGATG	TACCAC	A	TGA
T4 Early[c]							
Gene *63*	TCCCTC	GTG	TTGTGT	TATAGTA	GTCTTA	C	TGA

[a]Promoters and frequent start sites of transcription are underscored.
[b]Dunn and Studier (1983).
[c]Rand and Gait (1984).

Table 10.2
Mid- and Late-Phase Promoters of T-Series Bacteriophages[a]

		Promoter -10		Start Site	
T7 Mid[b]					
ϕ1.5	AGTTAA CTG	*GTAATA*	CGACTCA	CTAAAG	G AGG
ϕ1.6	TGGTCA CGC	*TTAATA*	CGACTCA	CTAAAG	G AGA
ϕ2.5	GCACCG AAG	*TAA-TA*	CGACTCA	CTATT-	A GGG
ϕ3.8	CGTGGA TAA	*TTAATT*	GAACTCA	CTAAAG	G GAG
ϕ4c	CCGACT GAG	*ACAATC*	CGACTCA	CTAAAG	A GAG
ϕ4.3	AGTCCC ATT	*CTAATA*	CGACTCA	CTAAAG	G AGA
ϕ4.7	TTCATG AAT	*ACTATT*	CGACTCA	CTATAG	G AGA
T4 Mid					
denV[c]	TACATC TCC	*TGTAGG*	TATGATA	CTATAG	A CCT
Gene 55[d]	(Incomplete)		TATGAAT	TGAGCT	A AGA
Orf D[e]	GCTCCT ATA	*TTGCTT*	TATAAAT	TTTT--	T GGT
T7 Late[b]					
ϕ6.5	GTCCCT AAA	*TTAATA*	CGACTCA	CTATAGG	G AGA
ϕ9	GCCGGG AAT	*TTAATA*	CGACTCA	CTATAGG	G AGA
ϕ10	ACTTCG AAA	*TTAATA*	CGACTCA	CTATAGG	G AGA
ϕ13	GGCTCG AAA	*TTAATA*	CGACTCA	CTATAGG	G AGA
ϕ17	GCGTAG GAA	*ATAATA*	CGACTCA	CTATAGG	G AGA
ϕOR	CACGAT AAA	*TTAATA*	CGACTCA	CTATAGG	G AGA
T3 Late					
pjB10[e]	AAACAC TGG	AAG*TAA*	TAACCCT	CACTAA	C AGG
HpaIN[e]	TCCAAC GTT	GTC*TAT*	TTACCCT	CACTAA	A GGG
pjB20[e]	GAAGTG AAA	GCC*TAA*	TTACCCT	CACTAA	A GGG
MboI-E[f]	TCAATG AGT	TTG*CAT*	TAACCCT	CACTAA	A GGG
T4 Late					
Gene 67[g]	TCGTTT CCA	AGA*CCC*	CGACCAA	GAACAA	G AGG
Orf[h]	(Incomplete)		-AAGCTT	GCTAAG	G AGA
P23[i]	CACTAT TAC	TGAGAG	TATAAATA	CTCCCT	G ATA
Gene 45[j]	TTTAAC +15	AAATTA	GTTATAA	AATTAA	A TCT

[a]Possible promoters and start sites are underscored; the regions adjoining the −10 sequences that may also be involved in promotion are italicized. Orf, open reading frame.

[b]Dunn and Studier (1983). [c]Valerie *et al.* (1984). [d]Gram and Rüger (1985). [e]Bailey *et al.* (1983). [f]Sarkar *et al.* (1985). [g]Völker *et al.* (1982). [h]Purohit and Mathews (1984). [i]Elliott and Geiduschek (1984). [j]Spicer *et al.* (1982).

where the promoters of prokaryotes are typically located, while that of the seventh (ϕ3.8) deviates only at the first two sites. Farther upstream is what could pass for the TA-TA- sequence (in italics) often considered a characteristic of the *E. coli* promoter, but it varies widely in structure from sequence to sequence. Furthermore, these runs lie more distant from the start site than in bacteria. In view of its location and conservation of structure, it is here proposed that the (CG)ACTCA just pointed out may serve as the transcriptional initiation signal in T7, either alone or in conjunction with nucleotides lying to either side. In the three sequences flanking the 5′ ends of T4 midphase genes, a structure closely akin to the bacterial standard is to be noted, properly placed around the −10 point. Downstream of the promoter in *denV* the CTATAG sequence is identical to the last of the T7 structures (ϕ4.7), but no uniformity of construction is to be observed in this sector, such as that marking those of T7.

Among the late genes of T7, similar constancy of construction is perceived (Table 10.2). In the six promoters from this region, precisely the same sequence is found in the −10 position as in the midsector, but this time there are no deviant forms. Indeed, the upstream hexanucleotide combination also is constant, being TTAATA in all except ϕ17, which deviates solely in having the initial base A. Moreover, all the downstream nucleotides are invariant to beyond the start site of transcription, the resemblance of the CTATAGG that begins this series to the CTAAAG of the midphase gene being self-evident. Again the constancy of structure and location of the CGACTCA in the −10 locality (underscored) argues for that sequence's serving as the promoter, possibly in conjunction with at least a part of the adjacent upstream and downstream nucleotides, as discussed in more detail just below.

The actual start sites of early transcription deviate widely (Table 10.1), since each of the four from T7 begins with a different nucleotide, but all are located relatively close to the translational start point, ATG. As far as can be detected, ancillary sites are absent from genes of all three phases (Elliott and Geiduschek, 1984).

A New Promoter Sequence? Because both mid- and late-phase genes are transcribed by RNA polymerase of viral origin, it is economical of space to treat them jointly in examining the foregoing proposal more extensively. First it needs to be noted that in comparison with the bacterial polymerase, the corresponding enzymes of these T-series bacteriophages are quite simple, consisting of a single polypeptide (Bailey *et al.*, 1983) that contains 884 amino acid residues (Moffatt *et al.*, 1984). Furthermore, other factors, still poorly known, also play important roles in the transcriptive processes, which are far more complex than implied by the simplicity of the polymerase (Kassavetis *et al.*, 1983; Pulitzer *et al.*, 1985). In phage T4, for instance, the products of at least four or five genes are essential.

As just seen, start and termination sites are well known in T7, and several studies have been made on promoters of T4 (Spicer *et al.*, 1982; Völker *et al.*, 1982; Rand and Gait, 1984; Valerie *et al.*, 1984; Gram and Rüger, 1985), but none seem to have been completed on other members of the two T-series. Nevertheless, the starting points of transcription of several genes of T3 have been indicated (Bailey *et al.*, 1983; Sarkar *et al.*, 1985), and through use of these, a further analysis of the promoter regions can be provided on a broader basis.

That the proposal made in connection with the initiation sectors of T7 may be justified is substantiated by those of three T3 late genes, whose promoter regions have

been sequenced (Table 10.2). There the combination T-ACCCT is found, while upstream from that point only three bases show relations to the TA-TA box. Combining the two yields the decanucleotide TAAT-ACCCT, obviously related to that standard promoter of *E. coli*. The downstream series is quite as invariant as that of T7, but its CACTAA shows no kinship in structure. However, when the proposed promoter and this portion (underscored) are properly aligned as follows, with unimportant bases in lowercase, an evolutionary relationship can be detected:

(T7)	TA–cgaCTCA	CTATAGG
(T3)	TAacc–CTCA	CTA-A

Thus, it may be that the promoter sequence of T3 has been derived by deletion of a portion of that of T7, with parts of the neighboring upstream series becoming added to it. In T4 all constancy of structure is lacking in the 5′ leaders of four late genes, including the sequences in the −10 region and adjoining parts, so that the sequence requirements for transcriptional initiation by the polymerase cannot be fully detected at present.

10.5.2. Promoters of Genes of Other Bacteriophages

As a whole, the processes of transcription in other bacteriophages, including initiation and termination, are still in their early stages of exploration, and details are intermittently available. Accordingly, what has been established in the remaining DNA types is combined in a single table with similar information regarding RNA species (Table 10.3). Consequently, different polymerases are involved in transcription of the latter than in the former. For ease of comparison, sequences from RNA varieties have the Us for uridine replaced by Ts for thymidine, as in DNAs. For the sake of continuity and clarity, several additional DNA phages are examined before any of the RNA type.

Promoters of Bacteriophage λ. Since the entire genomic sequence of bacteriophage λ has been established (Sanger *et al.*, 1982), a number of its promoter sites are known, five that have been studied experimentally being given in Table 10.3 (Schwarz *et al.*, 1978; Hoyt *et al.*, 1982; de Haseth *et al.*, 1983; Ho *et al.*, 1983; Shih and Gussin, 1983; Hoopes and McClure, 1985). Comparisons of the −10 regions of this quintet with the bacterial consensus sequence reveal only vague resemblances and no real homology. Moreover, very little similarity is found among the five promoters themselves. In the −35 zone, considerable constancy in structure is found to exist between P_I and P_{RE} and again between P_R and P_{RH}−, but this relationship is not reflected in the promoters of the respective pairs. The DNA-dependent RNA polymerase appears to favor A and T as the start sites for its activity.

Control of certain genes is a complex process, involving the products of several cistrons; this is especially the case in coding elements concerned with the establishment of lysogeny in this temperate virus. Among the genes and products (in parentheses) concerned with this activity are the *cI* (repressor), *int* (integrase), *xis* (excisionase), and *cII* (cII protein), expression being chiefly from the promoters P_I and P_L, along with P_{aQ}, the function of which remains hypothetical (Hoopes and McClure, 1985). The first two are associated with the integrase and excisionase genes, which overlap one another, *xis* extending farther upstream, so that the promoter P_I of *int* includes its translational start

Table 10.3
Promoters of Various DNA Bacteriophage Genes[a]

	Ancillary (-35)		Promoter (-10)		Start site	
E. coli consensus	*TGTT-GACANTTT*		TATAATC			
Phage λ						
P_R[b,c]	*GTGTTGACTATTTTA*	CCTCTGGCG--	GTGATAA	TGGTTGC	*A*	TGT
P_I[d]	*TTGC-GTGTAATTGC*	GGAGACTTTGC	GATGTAC	T------	*T*	G--
P_{RE}[e]	*TTGC-GTTTGTTTGC*	(Incomplete)	GTAAGTA	T------	*A*	G--
P_{aQ}[f]	*GCTC-GTGAACGTCA*	TGGAAACGGA-	ATCATAA	AGGAAGT	*T*	CGA
P_{RM}[e]	*GTGTTAGATATTTAT*	CCCTTGCGGTG	ATAGATT	TAA-CGT	*A*	TG
Phage SPO1[g]						
TF1	TTTGGAGAG*AAGTTT*	*CAAACA*CCC---	GATTTTT	TTATTA	*C*	GA
Phage Φ29[h]						
G3b (early)	GTGTTG*AAAATTGT*	*CGAACA*GGGTGA	TATAATA	AAAGAGC	*T*	AGA
Phage P2 (late)[i]						
F	*ATAGCCTGACATCTC*	CGGCGCAACT-	GAAAATA	-CCACT	*C*	ACC
O	*ATGGCGGAGGATGCG*	CATCGTCG---	GGAAACT	GATGCC	*G*	ACA
P	*TTAGCGATCGCGGGG*	CGCGACTCA--	GTAGCCT	TGCCGT	*G*	TAT
V	*ATAGCATAACTTTTA*	TATATTGT---	GCAATCT	CACATG	*C*	ATG
Phage Mu						
mom1 (late)[j]	TTAAGATAGTGGCGA	ATTGATGCAA-	AGGAGGTGA	GATGAA	*A*	TCA
mom2 (late)[k]	CACTCGACCCATGAT	GTTTTTTAAGA	TAGTGGCGA	ATTGAT	*G*	CAA
dam (late)[l]	GATCGAATCAATTAA	ATCGATCGG--	TAATACAG	ATCGAT	*T*	ATG
pC (early)[m]	GCTTTACATTAAGCT	TTTCAGTAA--	TTATCTT	TTTAGT	*A*	AGC
Phage ϕX174						
PG[n]	CTGTTGACAT(+11)	GTGGATTAC	TATCTGAG	TCCGAT	*G*	CTG
PA'[n]	AGCCTTGACCCTAAT	TTTGGTCG-	TCGCGTAC	GCAATC	*G*	CCG
PA1[n]	TAGCTTGCAAAA(+7)	CCTTATGGT	TACAGTATG	CCCATC	*G*	CAG
PA2[o]	TTGACACCCTCCCA-	ATTGTATGT	TTTCATG	CCTCC-	*A*	AAT
PD[o]	ACATTTTAAAAGAGC	GTGGATTAC	TATCTGA	GTCC--	*G*	ATG
PB1[o]	TAGCGTTGACCCTAA	TTTTGGTCG	TCGGGTA	CGCA--	*A*	TCG
PB2[o]	TTGCAAAATACGTGG	CCTTATGGT	TACAGTA	TGCCC-	*A*	TCG

(*continued*)

Table 10.3 (Continued)

	-35	Promoter (-10)		Start site	
Phage fd					
P(X)[n]	TTTGATGCAATT(+6)	GCTTCTGAC	TATAATA	GACAGG	*G* TAA
P(IV)[n]	ACTATTGACTCT(+7)	GTCTTAATC	TAAGCTA	TCGCT-	*A* TGT

[a]Experimentally confirmed sectors are underscored (promoters) or italicized (ancillary and start sites).
[b]de Haseth *et al.* (1983).
[c]Shih and Gussin (1983).
[d]R. W. Davies (1980).
[e]Ho *et al.* (1983).
[f]Hoopes and McClure (1985).
[g]Greene *et al.* (1984).
[h]Hattman and Ives (1984).
[i]Christie and Calendar (1985).
[j]Plasterk *et al.* (1984).
[k]Murray and Rabinowitz (1982).
[l]Plasterk *et al.* (1983).
[m]Krause *et al.* (1983).
[n]Otsuka and Kunisawa (1982).
[o]Sanger *et al.* (1977).

signal. However, the second promoter, P_L, is located still more strongly 5′ of that point, although its identity unfortunately has not been fully established (R. W. Davies, 1980). Thus, activation of P_L results in transcription of both integrase and excisionase, whereas transcription from P_I by the product of *cII* produces only mRNAs for integrase, the aborted coding sequence of *xis* becoming incorporated into its leader sequence (Campbell, 1983). The cII protein binds DNA, selectively interacting with a repeat sequence at the −35 location on the face of the DNA molecule opposite that employed by RNA polymerase (Ho *et al.*, 1983).

Promoters of Other DNA Phages. In the DNA of SPO1 of *Bacillus subtilis,* the typical thymidine is replaced enzymatically with 5-hydroxymethyluracil, a condition that is important in the transcriptive processes of the virus (Greene *et al.*, 1984). The gene *tf1* encodes a DNA-binding protein called transcription factor 1 (TF1), which reacts preferentially with sites containing the modified base. Thus, it negatively controls transcription. Transcription in SPO1 is carried out by three different polymerases. In transcribing early genes, the host enzyme is employed, in midphase that protein is modified by phage-encoded subunit σgp28, and in the late phase, modification involves two phage subunits, σgp33 and σgp34 (Lee and Pero, 1981; Pero, 1983). The promoter, which has been experimentally determined, shows few correspondences with any of bacteriophage λ, a statement equally true for the ancillary site at the −35 location.

Another phage from this same bacillus, that known as Φ29, provides some parallels of structure between the foregoing and the early gene cluster for the gene *G3b* (Table 10.3). Here there is a closer correlation in the promoter to the *E. coli* consensus sequence than in that of SPO1, but the relationship between the two phage early regions is confined to the

−35 zone, the AATTGTCGAACA of the present species being largely homologous to the AAGTTTCAAACA of the other. An additional point of resemblance is provided by the ribosomal recognition site. It has been proposed that a possible reason for the inability of *B. subtilis* to express genes from *E. coli,* which processes the other's genes freely, is that the signal mentioned might be too short in the cistrons of the latter for the polymerase to recognize (Murray and Rabinowitz, 1982). Whereas in *E. coli* a four-nucleotide combination such as GGAC suffices, the present form requires sequences like AGAAAGTGGG. Upon examination of the SPO1 cistron and its flanking regions (Greene *et al.,* 1984), that precise combination proves to be absent, but an equivalent, AAAGGGTGG, is found at a corresponding position, suggesting the feasibility of the proposal. To the contrary, in the second set of early genes from ϕ29 (Holder and Whiteley, 1983), AGGAGG is advocated as serving as that signal (Escarmís and Salas, 1982).

The promoters of the four late transcriptional units of phage P2, from *E. coli,* have been the subject of a recent investigation (Christie and Calendar, 1985). In this species expression of the late genes requires the host polymerase and the product of the phage gene *ogr,* which apparently modifies the α subunit of that enzyme. Also requisite are the products of two P2 DNA replication genes, *A* and *B* (Lengyel and Calendar, 1974). Comparisons of the four late sequences demonstrate their variability among themselves (Table 10.3), in which only the occupants of the first and third sites are constant. In the first two cistrons, some resemblances to the *E. coli* consensus structure are shown, but in the remaining pair the sequences are totally different.

The processes of transcription of just one more example of double-stranded DNA phages need attention, that of bacteriophage Mu, which is one of the most active transposable elements known (Krause *et al.,* 1983). At times the viral genome may be transposed to as many as 50 different sites in the host DNA in 1 hr. One of the late genes, *mom,* which modifies certain adenine residues of its genome, is expressed only when the product of a second cistron, *dam,* is present. Also required is the protein encoded by *dad* (Hattman and Ives, 1984). Transcription of *mom* has been demonstrated to occur only after several copies of the tetranucleotide GATC situated upstream of its promoter have been methylated by *dam* (Plasterk *et al.,* 1983). The entire gene and leader have been sequenced by two laboratories independently, and, while the structures of the two are identical, different interpretations have been given as to the start site of transcription and, concomitantly, the identification of the promoters (Hattman and Ives, 1984; Plasterk *et al.,* 1984). Given in Table 10.3 as *mom1* and *mom2,* the two are seen to differ extensively between themselves and also with the corresponding sector of the third late gene, *dam.* The latter correlates most closely with the *E. coli* consensus series of nucleotides, in fact having greater resemblance than the early promoter pC cited there (Krause *et al.,* 1983).

One of the traits peculiar to Mu is a site-specific inversion of a genomic segment, called G, that carries four genes in the sequence 3′ *Sv–U–U′–S′v* 5′, the last two being in opposite orientation from the first pair (Figure 10.6A). All encode tail-fiber components involved in the infection of its host. When segment G becomes inverted, *U′* and *S′v* are transcribed, since they then lie on the same strand as the promoter and adjacent to it, while the other two become unexpressed, since no such promoter then is present in the correct orientation (Figure 10.6B; Craig, 1985). Inversions occur within 34-base-pair inverted repeats, located one at each terminus of G, and are catalyzed by the invertase product of *gin,* which gene lies just downstream of the sector; however, an unidentified host factor is

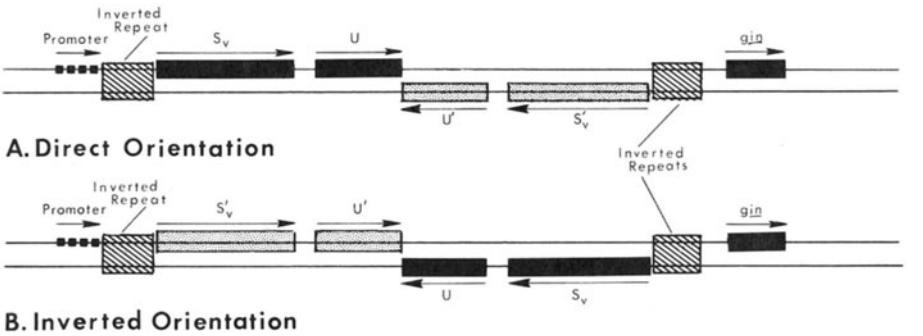


Figure 10.6. A site-specific inversion in the double-stranded DNA bacteriophage Mu. Inversion of a DNA segment changes transcription from that of (A, solid black) the typical genes to (B, stippled) a second set. (After Craig, 1985.)

also requisite (Kahmann *et al.*, 1985). Similar inversions that control gene expression are known in the phages P1 and P7 and in the bacterium *Salmonella typhimurium.*

Promoters of Genes of Single-Stranded DNA Phages. Very little firm information is available regarding promoter structure in single-stranded DNA phages; the main exceptions pertain to two species, ΦX174 and fd (Sanger *et al.*, 1977; Otsuka and Kunisawa, 1982). In Table 10.3 seven from the first virus and two from the second are cited, whose general interrelatedness makes a combined discussion possible. One of the transcription recognition signals, P(X), proves to be virtually identical to the *E. coli* consensus, but the remainder shows at most a 50% correspondence to that sequence, with the majority having only the first nucleotide (T) like that found in the bacterium. Almost all terminate in A rather than G. Like the majority shown in the table, the polymerase can begin transcription with any of the four common nucleotides, as is especially clearly illustrated in the PA′ sequence, where initiation may be at any site in the combination ATCGC.

10.6. PROMOTERS OF EUKARYOTIC DNA VIRUSES

It is not proposed here to examine the transcriptive processes in every type of DNA virus of eukaryotes, nor later to do likewise with the RNA species. Rather, present purposes are met best by providing depth in a few types where the synthesis of mRNAs has been more fully investigated, supplemented by several samples of less well-documented representatives to create the necessary breadth. Because transcription in the relatively rare DNA viruses of seed plants has not been explored sufficiently, the DNA viral group is represented largely by forms from vertebrates. The primary exception among plant DNA viruses is the cauliflower mosaic virus, which will be recalled as a double-stranded type with two breaks in the plus and one (Δ1) in the strand that is transcribed, the minus. Six reading frames have been disclosed, all obviously in the same orientation and together covering 85% of the transcribed molecule (Guilley *et al.*, 1982). While promoters have been detected, they have not been identified, but at least one transcript originates near Δ1.

Early Promoters in Adenoviruses. The several species of adenoviruses, particularly those referred to as 2, 5, and 12, have been relatively abundantly explored as to their processes of transcription and therefore are especially useful among vertebrate viruses in introducing the subject. A degree of confusion results from the employment in the literature of two different forms of designating the genes, the letter E (early) being sometimes followed by Arabic, sometimes by Roman, numerals; the first of these alternatives is followed here. One of the pre-early genes, *E1A*, is of outstanding importance in that its product at times is involved in the transcription of early cistrons. Experiments on transcription of this coding element suggest that a promoter somewhat resembling the typical TA-TA- is present, not at its accustomed −10 position, but farther upstream at −28 (Table 10.4; Osbourne *et al.*, 1982). No region 5′ to this sector influences transcription, nor is the promoter essential, but if it is removed by mutagenesis the level of production of mRNAs from *E1A* is reduced to 10 or 20% of the wild-type level. Other studies have yielded somewhat different results, in that the region above −231 has been reported to enhance transcription from the −28 promoter (Hen *et al.*, 1983; Sassone-Corsi *et al.*, 1983).

Expression of the early gene *E2A* is especially influenced by the 289-amino acid phosphoprotein encoded by the longer (13 S) mRNA produced from the foregoing cistron, a substance requisite for transcription from any early promoter (Gaynor and Beck, 1983). Strikingly, this stimulant of adenovirus transcription represses those processes from SV40 early promoters (Velcich and Ziff, 1985). But there may be some confusion here because of *E1A* having two promoters, for it has been demonstrated that the product of the shorter mRNA (12 S) from *E1A* represses *E2A* activity (Guilfoyle *et al.*, 1985). Thus the existence of two overlapping genes in *E1A* is clearly intimated. The postulated promoter of the early cistron *E2A* may be noted to be partly homologous to that of the pre-early gene (Table 10.4) and lies at a comparable distance from the start site. An ancillary sequence has been detected, located nearly 80 residues upstream of the start (Murthy *et al.*, 1985). It is worthy of note that two other studies of presumably this same gene of adenovirus 2 reported entirely different promoter and surrounding regions, distinguished in Table 10.4 as $E2A_2$ and $E2A_3$ (Langner *et al.*, 1984; Zajchowski *et al.*, 1985). In addition, the latest of these researches reported dual promoters, here distinguished as $\phi 1$ and $\phi 2$. The latter, farthest from the translational start site, bears considerable resemblance to the standard TA-TA- combination, while the first shows none at all. In addition to the promoters, a region lying between the two (italicized in $E2A_3$ $\phi 1$) is essential for transcription from either one of them, and the ancillary site italicized in the $\phi 2$ sequence is required for expression from that promoter. This second one is also necessary for activation by the *E1A* product (Zajchowski *et al.*, 1985). In a promoter of a third gene in this category, *E3*, the sequence (Table 10.4) shows loose kinship to the others from this virus, but the ancillary site displays no similarity in any form (Lee *et al.*, 1982). Despite this variability in the signal, all transcription of the early genes results from activity of the same host-cell RNA polymerase II, possibly with the aid of various ancillary proteins.

Late Promoters in Adenovirus. Some of the problems of initiation of those adenovirus late genes that are transcribed by RNA polymerase III have already received attention in Chapter 3, Section 3.2.4, to which reference should be made for a more complete view of the subject. But those that encode an mRNA are, like the early ones, acted upon by RNA polymerase II. Very few in that category have been examined for transcriptional properties, the major late (*ML*) gene and an associated intermediate one,

Table 10.4
Promoters of DNA-Viral Genes of Mammals[a]

	Ancillary			Promoter			Start Site	
Adenovirus								
E1A[b]	CGTTTTTATT	ATTATAGTCAGC	TGACGTGTAGTG	--TATTTATA-	CTCGGTGAG	(+10)GCC	*A*	CTC
$E2A_1$[c]	AGATGACGT	AGTTT(+23)CG	CGAAACTAGTCC	--TTAAGAGT-	CAGCGCGCA	(+8)TGA	*A*	GAG
$E2A_2$[d]		(Incomplete)	TCAGG	--TACAAATT-	TGCGAAGGT	(+9)TCC	*A*	CAG
$E2A_3\phi1$[e]	AGATGACGT	AGTTT(+21)AG	*GGCGCGAAA*CTA	--GTCCTTAA-	GAGTCAGCG	(+12)TGA	*A*	GAG
$E2A_3\phi2$[e]	CGCC*GGGTG*	*TGGCC*(+6)ACG	TAGTTTTCGCGC	--TTAAATTT-	GAGAAAGGG	(+13)CTT	*A*	AGA
E3[f]	GGCGCAGCT	TGCGG(+23)TC	GCCCGGGCAGGG	--TATAACTC-	ACCTGAAAA	(+9)GAG	*G*	TAT
ML (late)[g]	GTGATTGGT	TTATA(+24)TC	CTGAAGGGGGGC	--TATAAAAG-	GGGGTGGGG	(+11)CTC	*A*	CTC
IVa2 (mid)[g]	CCCCCTAGT	GGACA(+15)CC	CACTTAGCCTCC	--TTCGTGCT-	GGCCTGGAC	(+11)GTC	*T*	CAG
SV40								
φ Early[h]	*CCATTCTCC*	*GCCC*CATGGCTG	ACTAATTTTTTT	--TATTTAT--	GCAGAGGCC	GA(+7)TC*G*	*G*	*CCT*
φ Late[i]	(Incomplete)	CAGCTGGTTCTT	TCCGCCTCAGAA	GGTACCTAACC	AAGTTCCTC	(+8)GTT	*A*	T

[a]Experimentally determined promoters are underscored and initiation sites are italicized. Inc, incomplete.
[b]Osbourne *et al.* (1982).
[c]Murthy *et al.* (1985).
[d]Langner *et al.* (1984).
[e]Zajchowski *et al.* (1985).
[f]Lee *et al.* (1982).
[g]Matsui (1982).
[h]Baty *et al.* (1984).
[i]Brady *et al.* (1982).

IVa2, being among the exceptions (Matsui, 1982; Concino *et al.*, 1984). Although the *ML* leader contains a recognizable TA-TA- sequence, the midphase does not (Table 10.4), yet it is transcribed as faithfully as the other in an *in vitro* system. These two cistrons are in the form of an inverted pair, with divergently arranged promoter regions. The absence of any recognizable promoter has been reported by other laboratories, along with that same condition from *E2A* (Brady *et al.*, 1982; Natarajan and Salzman, 1985). In the later of these investigations, the results suggested the need for a product of the *E1A* along with an enhancer and promoters, but the last two elements were not specifically identified.

The Promoter of SV40 Early Genes. The early genes encoded in the SV40 genome are expressed continuously throughout lytic infection, while the late ones are suppressed until onset of viral DNA replication (Tack and Heard, 1985). After entry into the host's nucleus and after the parasite's DNA has been completely uncoated, cellular RNA polymerase II initiates transcription from the single promoter for the early region, resulting in the eventual production of mRNAs for large and small tumor (T) antigens. When those messengers have been translated, the large T antigen is returned to the nucleus, where it interacts with that same promoter and the region of origin of DNA

replication (Keller and Alwine, 1984). The early promoter sequence shows no unusual traits, being identical to that of adenovirus *E1A* (Table 10.4); however, a distinctive feature is found immediately upstream of that signal in the form of an uninterrupted run of seven Ts (Benoist and Chambon, 1981; Ghosh *et al.*, 1981; Byrne *et al.*, 1983). The location of the promoter, which has been demonstrated to be essential for accurate transcription (Mathis and Chambon, 1981), is about nine base pairs closer to the start site than the adenoviral type, and hence is more similar to the corresponding element of prokaryotes.

However, the processes just described are simplistic to a degree in that they fail to include the elements involved in late-early transcription, applying only to those of the early-early stage. No specific promoter has (or, more likely, promoters have) been identified, but G,C-rich motifs contained within a trio of 21-base-pair repeated elements have proven essential (Baty *et al.*, 1984). The late-early start sites, shown underscored in Table 10.4, are located well before the TA-TA- sequence that serves in the early-early stage. Consequently, the latter is firmly excluded as a factor in this subsequent period.

An Enhancing Sequence. In addition to the promoter, a unique enhancer element has been reported to play an important role in transcription of both early and late genes. The atypical feature of this component is in its structure, which consists of a pair of identical sequences 72 base pairs long, arranged in tandem (Gruss *et al.*, 1981; Byrne *et al.*, 1983). Moreover, the repeated G,C-rich sectors extending upstream from their beginnings in the ancillary site (Table 10.4) also have proven to be of significance in the processes (Everett *et al.*, 1983). Thus, early transcription in SV40 is known to require host RNA polymerase II, the promoter and enhancer, the large T antigen, and then G,C-rich sectors located in 21-base-pair repeats. Late-early stages are known to require the same G,C-rich elements and the 72-base-pair repeated segment, but specific promoters have not been identified. Thus, as seen in earlier chapters, transcription even in these viruses is not a simple, semiautomatic process, but is highly complex, involving many interacting enzymes and structural components.

The Promoter of SV40 Late Genes. Transcription of the late genes of SV40 that encode two mRNAs (one for VP1, the other for VP2 and VP3) involves the same 72-base-pair repeats as the foregoing, but takes place in opposite orientation (Ernoult-Lange and May, 1983; Sassone-Corsi *et al.*, 1984). Also necessary to efficient mRNA synthesis is the same series of G,C-rich repeated elements used in the early phase. Only a single promoter has thus far been proposed, one that is particularly striking in length, containing 11 base pairs (Brady *et al.*, 1982, 1984). Since thus much of the early stage control mechanism also affects late stage equally, it is difficult to account for the temporal effects that have been noted. However, refinement of experimental procedures have permitted the finding that subclasses of the large T antigen exist, at least one particular type of which is active only in regulation of late transcription (Tack and Heard, 1985). Moreover, the DNA structure appears also to be modified in the late period by an unknown substance which acts on the nontranscribed strand.

10.7. TRANSCRIPTION OF SOME RNA-VIRAL GENES

In analyzing transcription initiation of RNA viruses, much the same procedure is followed as in the preceding section, attention being devoted to a few, better documented

examples from different types of the parasites to provide both depth and breadth in an economical format. Here some varieties from seed plants can be included, along with those of vertebrates, thereby filling out the picture more completely.

10.7.1. Transcription of Genes of RNA Viruses

A condition made conspicuous by its absence in the foregoing statement is any mention of RNA bacteriophages. Because all the familiar members of that group have the single-stranded genome of the plus variety, transcription is unnecessary, replication of the entire nucleic acid molecule taking its place. The same statement is equally true for those RNA viruses of eukaryotes that have plus-strand genomes. On the other hand, in those with only minus strands, replication is a separate function and has received attention in the literature, albeit to a limited extent.

Transcription in Vesicular Stomatitus Virus. In all negative-strand RNA viruses. including vesicular stomatitus virus (VSV), that genome is encapsidated with ribonucleoprotein particles to form nucleocapsids, as seen in a preceding section. As there, these structures serve as templates for transcription and also genomic replication (Arnheiter *et al.*, 1985), both processes being carried out by the same RNA polymerase of viral origin, perhaps with the aid of host factors. For early use, some of the enzyme is packaged into the virion. It initiates transcription at the 3′ end of the genome and continues in sequence through the five genes in their structural order 3′ *N–NS–M–G–L* 5′, there being only one functional promoter, located at the 3′ end (Emerson, 1982). Moreover, this activity produces leader RNA from the 3′ terminus, the presence of which appears to shut down host macromolecular syntheses (Wilusz *et al.*, 1983; Grinnell and Wagner, 1984). Hence, the primary transcriptive product is a long, polycistronic mRNA, which is cleaved into the several portions by unidentified processes.

Transcription of Double-Stranded RNA Viral Genes. As described earlier, three groups of eukaryotic RNA viruses are similar in having genomes consisting of multisegmented double-stranded RNA, the cytoplasmic polyhedrosis virus of the silkworm and the reo- and rotaviruses of mammals. In the first two, there are ten segments and in the last 11, but all share such other characteristics as a viral RNA polymerase stored in the capsid and the presence of caps [including $m^7GppN(m)$] at the 5′ ends of their messengers. At least in the insect parasites, the production of the cap is a prerequisite to transcription (Furuichi, 1978). Each genomic segment consists of a monocistronic mRNA (plus strand) united to its complement in an end-to-end base-paired duplex, except the 5′ cap of the plus strand (Imai *et al.*, 1983). Since each segment is transcribed by the viral RNA polymerase into capped mRNAs, which can either be translated or employed as templates for synthesis of minus strands, every one must have promoters and replication signals. Despite the fact that a number of genes from reoviruses and a few from rotaviruses have been fully sequenced (Cashdollar *et al.*, 1982, 1985; Richardson and Furuichi, 1983; Dyall-Smith and Holmes, 1984), numerous cap sites but no promoters have been reported.

Transcription in Influenza Viruses. Discussion of the influenza viruses, which have a minus single-strand RNA genome, might well have preceded that of the double-stranded ones, but has been reserved for this point, since their transcriptive processes may throw light on those others. Like the above, the RNA exists in multiple strands, of eight segments, however, not ten or 11 as there (Huddleston and Brownlee, 1982); an additional resemblance is that a cap of identical structure is present at the 5′ end of each.

Although most segments are monocistronic, at least three, numbers 6–8, encode two proteins. The last of these carries genes for a pair of nonstructural proteins, NS_1 and NS_2 (Lamb and Choppin, 1979; Lamb and Lai, 1980; Lamb *et al.*, 1980; Porter *et al.*, 1980), whereas the first bears those for the neuriminidase (NA) and a glycoprotein (NB) (Shaw *et al.*, 1983). The third exception, segment 7, more recently has been demonstrated to encode two membrane proteins, M_1 and M_2 (Lamb *et al.*, 1985).

Transcription requires the presence of the caps, the nucleocapsid protein (NP), and three P proteins—PB_1 and PB_2 being basic, and PA being acidic (Braam *et al.*, 1983). NP, the predominant protein of the virus, comprising ~90% of the total proteins, is located along each of the eight RNA segments at about 20-nucleotide-residue intervals. Seemingly the three P proteins form a complex, which at the onset of transcription moves as a unit from the 3′ ends of the viral RNA strands down the mRNAs as they are synthesized. PB_2 also interacts with the cap structure, which is then cleaved by a viral endonuclease at a purine located 10–13 residues distant, the resulting capped fragments serving as primers. Transcription proper then is initiated by the addition of a guanosine residue onto the 3′ end of the primers, a reaction guided by the penultimate C residue of the viral RNA strands. Following initiation, the polymerase complex elongates the messengers in typical fashion (Braam *et al.*, 1983). Consequently, specific promoter sequences are not a feature of transcription in these and possibly others that possess capped RNA genomic strands. In one gene whose sequence has been established, that for neuraminidase (Hiti and Nayak, 1985), a TAA translational stop signal immediately precedes the ATG initiation triplet, but whether it plays any significant role is not established.

10.7.2. Transcription of Plant RNA-Viral Genes

Despite the establishment of a number of genomic sequences from RNA viruses of plants, their processes concerned with transcription and with synthesis of the characteristic subgenomic RNAs still remain completely unknown (Watanabe *et al.*, 1984). In the tobacco mosaic virus, the 5′ end of the genome is capped by mGpp (Goelet *et al.*, 1982) and that structure possibly plays a role in its synthesis, no transcription of this plus-strand nucleic acid being necessary. However, the subgenomic strands are transcribed, that for the 30,000-dalton protein being initiated at a guanosine 1550 residues upstream of the terminus (Watanabe *et al.*, 1984), and that for the coat protein at 693 sites from its 3′ end (Guilley *et al.*, 1979). Aside from those few data, the mechanism, enzyme(s), and other factors remain for future investigations to reveal.

10.7.3. An Evolutionary Sequence of Transcriptional Events

Although firm data regarding initiation of transcription in viruses are lacking in the desired abundance, even casual reading of the foregoing descriptions discloses differing levels of complexity in those processes. At present, the events can be arranged at best into a mere skeleton of a phylogenetic succession, but perhaps its presentation will stimulate additional studies to detail its evolutionary progress further.

The starting point in the sequence is obvious, for what can be simpler insofar as transcriptional initiation is concerned than an inheritable genome that consists of messengers ready for translation? Thus, single-stranded RNA viruses whose genome is a plus strand represent the earliest phase in the development of transcription; multiple molecules,

Table 10.5
Terminators of Bacteriophages[a]

Phage T7[b]						
TE	CGTTTATAAGGA-	GACACTTTATGT	TTAA	GAAGGTTGG	TAAATT-C	CTTGCGGCTTTG
Tφ	TGCTGA-AAGGAG	GAACTATATGCG	CTCA	TACGATATG	AACGTTGA	GACTGCCGCTGA
Phage f1						
I[c]	TAAACCGATACA	ATT*AAAGGCTCC*	TTTT	*GGAGCCTTT*	TTTTTTGG	AGATTTTCAAC-
II[d]	CCCTTTGACGTT	*GGAGTCCACGTT*	CTTT	*AATAGTGGA*	*CTCT*TGTT	CCAAACTGGAAC
Phage fd[e]						
I	TAAACCGATACA	ATT*AAAGGCTCC*	TTTT	*GGAGCCTTT*	TTTTTTGG	AGATTTTCAAC-
Phage M13						
I[f]	TAAACCGATACA	ATT*AAAGGCTCC*	TTTT	*GGAGCCTTT*	TTTTT	
II[g]		*AACCTCCCG*	CAAG	*TCGGGAGGT*	*T*CGCT	
Phage IKe[h]						
I	TTTTCAGCGTTA	TTT*AAGGGGCGC*	TATT	*GCGCCCTT*T	TTTTTACT	TTAATTCAGCTA
Phage φX174[i]						
T4	GTATGT	TTTCAT*GCCTCC*	*A*AAT	C*TTGGAGGC*	TTTTTTAT	GGTTCGTTCT--
T2	CAACAATTTTAA	TTGC*AGGGGC*TT	CG*GC*	*CCCT*TACTT	GAGGAT	
T1	ACTATA	GACCAC*CGCCC*	*GAAC*	*GGGACG*AAA	AATGGTTT	TTAGAGAACG
Phage P22						
T *ant*[j]	GATAACCAAC	GCAACG*ACCCAG*	*C*TTC	G*GCTGGGT*T	TTTTTATG	
T *nutR*[k]	CCA	ATCTGA*ACCGC*C	GACA	AC*GCGGT*AA	ACC	
R_1[k]		TCAAA*GCGCA*	*T*-CA	ACGAA*TGCG*	*C*ACAACTA	ACTAT
Phage λ[k]						
T nutR		*GCCCT*G	AAAA	*AGGGC*ATCA	AATTAAAC	CACAC
R_1		CTATGG*TGTATG*	*CA*TT	TATT*TGCAT*	*ACA*TTCAA	TCAATT
Phage Pf3[f]						
CP		TCGTT*ATAAGG*	*GGGC*	TTC*GGCTCC*	*CTTAT*TCG	TTTA

[a]Regions of dyad symmetry are italicized and noncanonical base pairs are underscored.
[b]Dunn and Studier (1983). [c]Hill and Petersen (1982). [d]La Farina and Vitale (1984). [e]Beck *et al.* (1978). [f]Luiten *et al.* (1983). [g]Smits *et al.* (1984). [h]Peeters *et al.* (1985). [i]Otsuka and Kunisawa (1982). [j]Berget *et al.* (1983). [k]Backhaus and Petri (1984).

each encoding a single gene, probably antedated polycistronic types, for the latter require an additional element in the form of a cleaving enzyme. The next logical step is found in those single-stranded RNA species that need to be transcribed into messengers. In the earliest of such minus-stranded forms, the polymerase of transcription is also that for genomic replication, and requires no promoter or other signal, the 5′ cap alone being a prerequisite. Those that require a promoter sequence, and perhaps other factors, are at a still higher level of the processes, followed by DNA-containing species, but the steps by which transcription activities developed their greater complexity beyond that point are too dim to recount now.

10.8. TERMINATION OF TRANSCRIPTION IN VIRUSES

Although data pertaining to termination of transcription in viruses is not yet abundantly available, present knowledge is not so limited as that concerning initiation. At least it is found that many suggestions have been made as to the structure of possible terminators in a diversity of types. As expected, the terminators of bacteriophages provide the richest source of information—but not from the usual reliable forms.

10.8.1. Terminators of Bacteriophages

The T Series of Phages. As just intimated, the standard model of molecular virology, the two T series of bacteriophages, are of little value in supplying data concerning termination of transcription. Studies on the genomic sequence of T7 indicate the location of two terminators, from which the nucleotide structure of those regions given in Table 10.5 have been derived (Dunn and Studier, 1983). Obviously no series of Ts is found here, as is the frequent case in bacteria, nor is any present in the structure downstream of the initial section provided. Since regions of dyad symmetry also are lacking, how transcription is terminated here cannot even be conjectured. Nevertheless, the two examples are not without interest, for the one that lies close to the end of the genome, Tϕ, is seen to be largely homologous to TE, the terminator of early transcription. The similarities of structure are striking through the first 40 residues, but they are especially impressive through the opening 24 sites. The frequent identities of occupants in corresponding positions suggest clearly that that sector may eventually prove to be involved in termination.

Terminators of Filamentous Bacteriophage Genes. The terminator sequences of the first three of the four from filamentous viruses (f1, fd, and M13) add further support to the concept that those organisms have common ancestry (Luiten *et al.*, 1983), for these structures are 100% homologous (Table 10.5). Included in the identities is the region of dyad symmetry (italicized), with a loop composed of four Ts and a run of eight Ts at the 3′ end, thus greatly resembling the corresponding signal of *E. coli*. In the fourth member of that group, IKe, there is general resemblance, but greatly reduced homology, both in the dyad symmetric portion and the upstream sequence. Moreover, one of the four Ts in the hairpin loop of the others is replaced by an A, but the run of eight Ts at its downstream end is conserved. Thus IKe, although related to the other three, is more distantly so.

Apparently more than a single terminator exists per genome, for one of quite distinct structure has been determined at the end of the I region of f1 and at the 3′ end of an RNA

in M13, each referred to as terminator II in Table 10.5 (La Farina and Vitale, 1984; Smits *et al.*, 1984). That of f1, which has been shown to be rho-dependent, diverges from the other in having a noncanonical base pair (T,G, underscored) and an unpaired A in the stem; furthermore, the terminal run of Ts is short and interrupted. Although of the same general format as the rest, that of M13 shows almost no homology with the others and the run of Ts is absent from the 3′ terminus. But what is of greater interest is that the precise sequence given elsewhere as a terminator is reported in the same paper as a promoter at the 5′ end of the same RNA (Smits *et al.*, 1984), a fact that apparently escaped observation.

Still another terminator from a fifth filamentous virus, Pf3, whose host is *Pseudomonas aeruginosa,* not the *E. coli* of all the rest, shows almost nothing in common with the four just discussed insofar as sequence structure is concerned. However, it has a similar long region of dyad symmetry, with ten paired sets of bases instead of their eight, including one noncanonical T,G combination (underscored). Moreover, the downstream run of Ts is partially broken and shortened by an inserted CG combination.

Terminators of Miscellaneous Bacteriophages. In each of three viruses of a miscellany of types, the terminators of two or more genes have been investigated, presenting a further opportunity for comparisons of structures acted upon by the same enzyme. Although it is to be expected that all the terminators of a given bacteriophage should be closely similar, in these there is no meaningful level of constancy. Terminator 4 (T4) of ΦX174, for instance, parallels those of the four filamentous viruses rather closely, except in being shorter, the GCCTCC and GGAGGC of its dyad symmetric sector being recognizable modifications of the GGCTCC and TGGAGC of f1 and kin. Additionally, it is followed by a comparable run of Ts. On the other hand, its T2 sequence shows little in common in the paired sector and downstream structure, and in T1 the base pairing in the stem-and-loop element is interrupted and is followed by a series of five As (Otsuka and Kunisawa, 1982).

Similar variation in structure is seen in the phages P22 and λ, some terminators of each of which have received analysis (Backhaus and Petri, 1984). In that comparative investigation, the terminator R_1 and that of the *nutR* gene were examined from both source organisms. The two from the *nutR* gene had short loops, with only five paired sets of bases, but the two R_1 terminators differed both in length and structure. A third terminator from P22, that associated with the *ant* gene (Berget *et al.*, 1983), showed some sequence homology in the stem structure to that of the *nutR* cistron, but it is followed by a series of seven Ts absent from its mates.

10.8.2. Terminators of Genes in Viruses of Eukaryotes

To date in molecular researches on the vertebrate viruses, termination of transcription seems to have been a minor issue. In part this undoubtedly stems from the circularity of the genomes, in which transcription often is bidirectional, with broad, common initiation and termination zones for early and late gene expression. However, in a few forms, including SV40 and BKV, the typical polyadenylation signal AAUAAA has been detected, which in the first form at least may also serve for termination (Seif *et al.*, 1979; Conway and Wickens, 1985). The later study has demonstrated that the poly(A) itself is not as important in the present function as the 220-residue sequence that surrounds it.

The identical condition prevails in transcriptional termination in plant viruses. In many instances, as in the RNA2 of the tomato strain of TMV and cucumber mosaic virus (Ohno *et al.*, 1984; Rezaian *et al.*, 1984), no evidence of the standard polyadenylation signal can be traced beyond the last coding sequence, whereas in others, including TMV proper (Goelet *et al.*, 1982; Takamatsu *et al.*, 1983), one is present and correctly positioned. Tobacco streak virus RNA3 lacks that nucleotidyl combination, but three sequences in the 5′ leader are repeated in the 3′ region, arranged in similar fashion (Cornelissen *et al.*, 1984). Their particular role has not yet been examined, but they might prove to serve as terminators. It must also be borne in mind that a number of plant viral genomes terminate in tRNA cistrons, whose sequences themselves possibly may be sufficient to end transcription (Kozlov *et al.*, 1984).

10.9. OVERLAPPING (DUAL) GENES

A particularly characteristic feature of viral genomic structures in general is the presence of genes that overlap one another. So common is this arrangement that any attempt at covering examples even from the major types would be futile and largely an exercise in repetition. Instead, the effort here is devoted to the broader picture of this peculiarity, so that the discussion can apply to mitochondrial, chloroplastic, nuclear, and prokaryotic genomes, although the examples cited at this point are confined to the viruses.

10.9.1. Types of Overlapping Genes

Overlapping genes may be considered to fall into seven major types arranged in three great subdivisions. In those larger categories, the polarities and number of interacting sequences provide the criteria for separation. Division A contains pairs of genetic elements of opposite polarity, that is, located on different strands; division B embraces two located on the same strand and therefore having identical orientation; and division C includes sets of three or more jointly overlapping genes.

Division A. In the first of these categories are found two types. In type I only the terminal portions overlap, while in type II a smaller gene is contained within the limits of the coding region of a larger, but on the opposite strand (Figure 10.3C). The first may be considered to be divided into three subtypes, depending on whether the 5′ ends of each overlap (type IA), the 3′ ends do so (type IB), or the 5′ end of one overlaps the 3′ terminus of the other (type IC) (Figure 10.7).

Division B. Whereas the overlapping cistrons in the preceding division are necessarily confined to forms having double-stranded genomes, either RNA or DNA, those of division B can occur in species with any type of genetic apparatus, since the pair of sequences involved, being of like polarity, are situated on the same strand. Four major types appear possible, one of which has three variants. In type III the ends of the two genes overlap, three possible arrangements of which can occur. In subtype IIIA, the 5′ end of one begins just before the 3′ terminus of the other (Figure 10.7); subtype IIIB arrangements have the 3′ end of a shorter genetic element ending somewhat beyond that of a longer component; and in subtype IIIC, the 5′ end of the shorter begins prior to that of a longer one. In type IV, overlaps involve two genes that share a common initiation

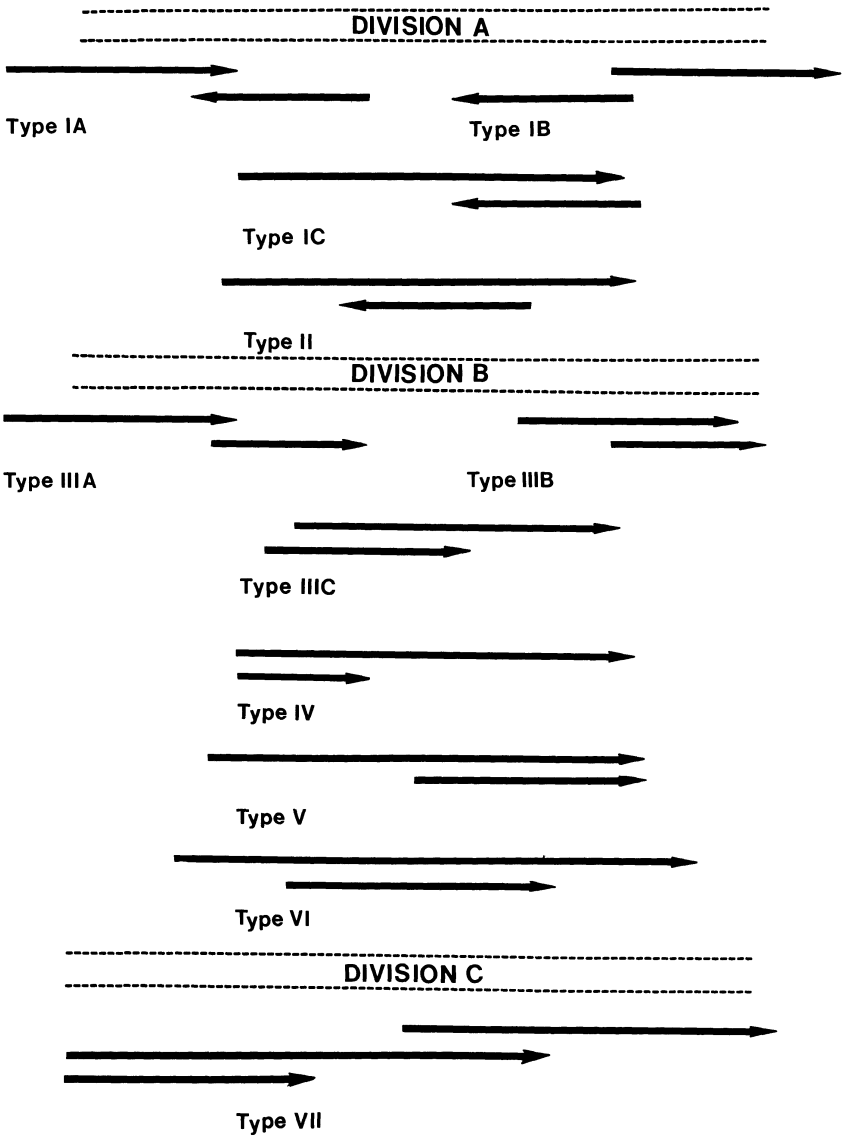


Figure 10.7. The three divisions and seven types of the major varieties of overlapping genes.

codon, but one terminates before the other. Quite in contrast, type V dual genes initiate separately, but terminate jointly, whereas in type VI a smaller genetic sequence lies totally within a larger one (Figure 10.7).

Division C. For the present it is considered expedient to place all sets of multiple overlapping genes into a single unit, type VII, until, with increasing knowledge, a greater abundance of representation than now appears evident may warrant recognition of additional groupings.

Table 10.6
Examples of Overlapping Gene Sequences[a]

T4 *frd*[b]	--- GAA TCA GTA TAT AAA TGA
T4 *ts*[b]	A TGA AAC AAT ACC AAG AT- ---
λ *xis*[c]	--- AAG AGG ATC AGA AAT GGG AAG AAG GCG AAG TCA TGA
λ *int*[c]	AT GGG AAG AAG GCG AAG TCA TGA G--
TMV *30,000*[d]	--- AGT TTG TTT ATA GAT GGC TCT AGT TGT TAA
TMV *28,000*[d]	AT GGC TCT AGT TGT TAA A-- --- ---
M13 *I*[e]	--- AAA GGT AAT TCA AAT GAA ATT GTT AAA TGT AAT TAA
M13 *IV*[e]	AT GAA ATT GTT AAA TGT AAT TAA T--
ΦX174 *C*[f]	--- ATA GGT AAG AAA TCA TGA
ΦX174 *D*[f]	A TGA GTC AAG TTA CTG AAC AAT CC
ΦX174 *A*[f]	--- ACT GCT GGC GGA AAA TGA
ΦX174 *K*[f]	-GG ACT GCT GGC GGA AAA TGA GAA AAT TCG ACC TAT CCT T--
ΦX174 *C*[f]	A TGA GAA AAT TCG ACC TAT CCT TG-

[a]Arcs indicate connections between parts of codons and primes demarcate their termini.
[b]Purohit and Mathews (1984).
[c]R. W. Davies (1980).
[d]Goelet *et al.* (1982).
[e]Van Wezenbeek and Schoenmakers (1979).
[f]Sanger *et al.* (1977).

10.9.2. Examples of Dual Genes

Unfortunately in one of the major sources of information from double-stranded DNA viruses, the genome of bacteriophage T7 (Dunn and Studier, 1983), all the genes are located on the same strand, so that it supplies no representatives of division A dual (overlapping) cistrons. However, it does yield some interesting examples of divisions B and C, as shown shortly, while phage λ supplies a few of the first.

Examples of Type III Dual Genes. By far the commonest kind of dual gene structure in viruses is type III, in which the initiation end of one of a pair of directly oriented genes overlaps the termination region of the other. One example of this condition from phage T4 is of particular interest because the downstream member, which encodes thymidylate synthase, is one of the few viral genes that contains an intron (Chu *et al.*, 1984). In this case overlap is minimal, involving only a single nucleotide, as indicated in Table 10.6, where in the downstream representative the parts of the codons are tied together by arcs, since they are out of frame with the first member (Purohit and Mathews, 1984). While thus extremely simple, this illustration nevertheless makes the problem

clear: Since this is part of a polycistronic transcript, translation of the *frd* gene for dihydrofolate reductase by the usual processes concomitantly destroys the 3′-coding region *ts,* and, conversely, preservation of the latter eliminates the former. How control over expression is exercised has not been established, but it may involve posttranscriptional processing, the subject of the closing chapter. Similar single-residue overlaps are frequent, occurring between genes *IX* and *VIII* of bacteriophage f1 (Hill and Petersen, 1982) and between genes *12* and *ninA* in phage P22 (Backhaus and Petri, 1984), to cite two additional cases.

In one set of dual genes of bacteriophage λ, the overlap involving coding elements for the integrase (*int*) and the excisionase (*xis*) of this temperate form extends through eight codons (Davies, 1980). In this instance a frameshift of two nucleotides is involved, not just one, as may be noted in Table 10.6. A double structure of slightly shorter length exists in the tobacco mosaic viral genome (Goelet *et al.,* 1982), also given in that table, but for clarity the uridyls of its RNA are replaced by the thymidines of DNA in this pair, whose products are known only by the molecular weights. Other examples are given from bacteriophage M13 (van Wezenbeek and Schoenmakers, 1979) an identical set of which occurs also in phage f1 (Hill and Petersen, 1982), but is not shown. Much longer overlaps of type III are known and are not uncommon. In influenza A virus, the NS_1 and NS_2 cistrons overlap by 61 codons (Porter *et al.,* 1980), in the vaccinia genome, *F9* and *F10* share a region of 29 triplets (Plucienniczak *et al.,* 1985), and *p4* and *p3* of phage Φ29 have a 38-codon region in common (Escarmís and Salas, 1982). Here, in each case, as in the first example given, the problem is the same—the expression of one member of the dual genes results in the elimination of the other, unless a special translational device exists. Hence, it is not an efficient arrangement and involves a measure of control that bespeaks the presence of a supramolecular element. In every instance a frameshift exists between the cistrons of each pair, a device that undoubtedly serves in expression control.

Examples of Miscellaneous Types. Type VII, in which three genes form a triple direct structure with two of the components initiating separately but terminating together, has a well-established example in the BK papovavirus of man (Yang and Wu, 1979). Here *VP2* has its ATG translational initiation signal 1350 base pairs prior to that of *VP3,* but the two have a joint termination codon (TAA). Thus, the two coding sectors are in the same frame of reference. However, some 110 base pairs before that point and out of frame with it is the initiation codon for *VP1.* Consequently, three genes are seen to share an extensive coding sector, signifying that only a single member of the trio can be expressed in a given round of translation, resulting in an even lower level of synthesis efficiency than in the dual systems. A duplicate of this that occurs in SV40 is illustrated in Figure 10.3A. In phage T7 early genes (class I) a similar condition exists (Dunn and Studier, 1983), but in this set, one (gene *0.5*) of the three cistrons has a single nucleotide residue at its 3′ end overlapping the initiation codon shared by the other two. The shorter of the latter, *0.6A,* terminates 58 codons before the longer, *0.6B.*

Bacteriophage T7 supplies illustrations of two other types of dual structures, each of which is located in the late (class III) region of its DNA. In the first of these, involving genes *10A* and *10B,* the initiation point is in common, but the second coding region extends 58 nucleotides beyond the first, thus exemplifying type IV duplexes. One example of a type V duality has been well-established in phage f1 (Fulford and Model, 1984). Here gene *II,* which encodes a site-specific endonuclease, has been shown to overlap gene

X, which begins at codon 300 of the former. This cistron, whose encoded product is unidentified, then terminates with gene *II*, the two sequences being completely in frame. Finally, type VI is represented by two sets of adjacent sequences in T7, genes *18.5* and *18.7* on the one hand and *19* and *19.2* on the other (Dunn and Studier, 1983). In both sets the second component listed is much shorter than the first and is enclosed entirely within the larger; each is out of phase with the sequence in which it is embedded by one position downstream. The processes by means of which the particular sequence is selected for translation should provide a fertile and challenging field for investigation by cellular and molecular biologists.

But to make the genomic structure and expression clear, it must be acknowledged that overlapping genes, although frequent, are not a predominant feature. More typically, sequential coding regions are separated by several nucleotide residues. But on abundant occasions spacing involves just one or two sites, and not rarely there may be no spacer at all between the translation stop signal of one and the initiation signal of the second. Moreover, in the numerous plus-stranded RNA species, the polycistronic messengers even lack those translational signals between components. On the other hand, in viruses with larger genomes, some gene sets may be spaced by runs of 20–100 sites, or even several hundreds.