7

Complex Genes

Thus far in this analysis, only those genes have been examined that simply encode a single product, whether an RNA or a protein. Although often the coding sequences have been found to be interrupted by untranslated sectors, the presence of such introns had no effect on the immediate product of translation, because of their removal before that process occurred. But now the point has been reached where the dictum "one gene, one peptide" loses its validity, for in the several major classes of genes that receive attention in this chapter two or more distinct proteins are encoded in every case. Each of these products must undergo processing before the main component (or components) is able to function. The simpler of the diverse complex genes is that class, earlier named diplomorphic (Chapter 1, Section 1.1.3), that codes for a double translational product. As a rule, but not always, the bulk of the transcript becomes translated into an active enzymic or structural protein; in addition, this bears a prefatory peptide that appears to be requisite for the protein to pass through a membrane. The latter may be either the cytoplasmic covering or the sheath that encloses an organelle, such as a mitochondrion, chloroplast, Golgi body, or endoreticulum.

As a result of the analyses of Chapters 2–6, full gene sequences that encode the entire mature product have been surveyed in depth sufficient to provide an adequate understanding of their nature, unique features, and interrelatedness. Consequently, attention here focuses solely on the characteristics of the adjuncts, except as a prominent, unique trait demands otherwise, beginning with those of the simplest nature and confronting the remainder in turn as their complexity increases.

7.1. DIPLOMORPHIC GENES

Diplomorphic genes encode many contrasting families of macromolecules, but in broad terms the class may be considered to embrace two major groups. One includes those whose products are secreted through the cytoplasmic membrane, either into the environment or an enclosing capsule such as exists around many bacteria and algae. The second category viewed here has already received some mention in the preceding chapter, for it embraces those nuclear genes that encode a product that functions in the mitochondrion or chloroplast, like those of cytochromes *f* and c_1. Since more attention has been devoted to

researches on bacterial products than those of other prokaryotes and protistans, these provide a major segment of those examined here, along with hormones, blood proteins, and a diversity of mammalian molecules. To those two large groups are added the few available from insects and fungi, and a larger, but still minor, fraction from seed plants.

7.1.1. Simple Diplomorphic Genes from Mammalian Sources

The primary transcripts and coding regions of many mammalian genes for secretory products have been established, so that they provide a firm foundation for some of the complexities that are to follow. Although the gene structures and the prefatory portions have been determined in great abundance, almost no special attention has been devoted to transcriptional problems, such as initiation and termination. However, near the close of this section is summarized what can be gleaned from the occasional promoter and other signals that have been incidentally garnered along with the main studies.

Although the term "signal peptide" is frequently applied to the temporary portion (presequence) preceding the protein proper, another name in general use, "transit peptide," appears superior in that it describes the main function of the appendage, service in passage through a membrane (Figure 7.1). Throughout much of this chapter the conventions used in the tables are uniform, underscoring indicating codons for polar charged (hydrophilic) amino acids, and italics for the apolar (hydrophobic) ones, groups of the latter being especially characteristic of membrane-enclosed structures. To bring out the critical features in spite of the variations in length that exist, the presequences are aligned both at their 5′ termini and again at the 3′ point where cleavage takes place during processing. The first two codons of the mature coding region are also included with the latter to disclose any clues possibly employed by the cleaving enzymes.

The Glycoprotein Family of Hormones. In vertebrates a unique family of hormones is found, whose members are closely related in quaternary structure. All consist of αβ dimers, the α subunit of which appears to be shared by all types, while the β subunit is distinctive and specifies the biological activity of the particular mature protein. The two types of subunits share one common feature, that of being glycosylated (Ramabhadran *et*

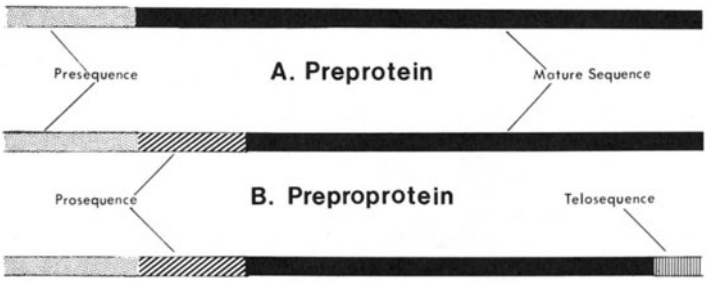


Figure 7.1. (A) Simple and (B,C) complex diplomorphic genes. Complex diplomorphs bear both a pre- and a prosequence, and occasionally (C) a telosequence as well.

Table 7.1
Transit Peptide Genes of the Glycoprotein Hormones[a]

Row 1	
Subunit α	
Mouse TSH[b]	*ATG* GAT TAC TAC AGA AAA TAT *GCA GCT GTC ATT CTG GTC ATG*
Rat GP[c]	*ATG* GAT TGC TAC AGA AGA TAT *GCG GCT GTC ATT CTG GTC ATG*
Human CG[d]	*ATG* GAT TAC TAC AGA AAA TAT *GCA GCT ATC TTT CTG GTC* ACA
Bovine GP[e]	*ATG* GAT TAC TAC AGA AAA TAT *GCA GCT GTC ATT CTG* ACC *ATT*
Subunit β	
Mouse TSH[f]	*ATG* AGT --- --- --- --- *GCT GCC GTC CTC CTC* TCC *GTG CTT*
Bovine TSH[g]	*ATG* ACT --- --- --- --- *GCT* ACC *TTC CTG ATG* TCC *ATG ATT*
Human LH[h]	*ATG* GAG --- --- --- --- *ATG TTC* CAA***GGG CTG CTG CTG TTG*
Rat LH[i]	*ATG* GAG --- --- --- --- AGG *CTC* CAG***GGG CTG CTG CTG* TGG
Row 2	
Subunit α	↓
Mouse TSH	*CTG* TCC *ATG TTC CTG* --- CAT *ATT CTT* CAT TCT *CTT CCT*
Rat GP	*CTG* TCC *ATG GTC CTG* --- CAT *ATT CTT* CAT TCT *CTT CCT*
Human CG	*TTG* TCG *GTG TTT CTG* --- CAT *GTT CTC* CAT TCC *GCT CCT*
Bovine GP	*TTG* TCT *CTG TTT CTG* --- CAA *ATT CTC* CAT TCC *TTT CCT*
Subunit β	
Mouse TSH	*TTT GCT CTT GCT* TGT --- *GGG* CAA *GCA GCA* TCC *TTT* TGT
Bovine TSH	*TTT GGC CTT GCA* TGT --- *GGA* CAA *GCA ATG* TCT *TTT* TGT
Human LH	*CTG CTG CTG* AGC *ATG GGC GGG* ACA TGG *GCA* TCC AAG GAG
Rat LH	*CTG CTG CTG* AGC *CCA* AGT *GTG GTG* TGG *GCC* TCC AGG *GGC*

[a]Codons for apolar (hydrophobic) amino acids are italicized and those for charged ones (hydrophilic) are underscored. Cleavage sites are indicated by the vertical arrow. Asterisk denotes the location of an intron. TSH, thyrotropin; GP, glycoprotein; CG, chorionic gonadotropin; LH, luteinizing hormone.
[b]Chin *et al.* (1981).
[c]Godine *et al.* (1982).
[d]Fiddes and Goodman (1979).
[e]Erwin *et al.* (1983).
[f]Gurr *et al.* (1983).
[g]Maurer *et al.* (1984).
[h]Boorstein *et al.* (1982); Policastro *et al.* (1984).
[i]Chin *et al.* (1983); Jameson *et al.* (1984).

al., 1984). Three species of the mature proteins are produced in the pituitary, including luteinizing (LH or lutropin), follicle-stimulating (FSH or follitropin), and thyroid-stimulating hormones (TSH or thyrotropin) (Van Heuverswyn *et al.*, 1984), while the fourth, chorionic gonadotropin (CG), is secreted by the placenta of mammals.

Table 7.1 includes four representative presequences of the genes encoding each subunit, although examples only from two hormone species can be listed for the β peptide. When the structures of the α presequences are examined, their virtual identity is at once apparent, despite the fact that their sources are from four diverse species of mammals and from different hormones. Consequently, there can be no doubt that the same α subunit serves in each dimeric protein, ragardless of the latter's function. Very little variation from one sequence to another can be noted; the only difference of any consequence is in the bovine glycoprotein (GP) presequence just before the end of row 1, where a neutral amino acid is encoded instead of the apolar one of the others. A similar distinction occurs in the human CG peptide at the termination of that same row. In contrast, the first codon for the mature coding sequence is seen to vary widely, whereas the second is constant.

There is some confusion in the literature regarding the correct identity of the two β subunits given here as luteinizing hormones, as well as the precise location of the cleavage site. Two articles pertaining to the human gene originally described it as encoding chorionic gonadotropin (Boorstein *et al.*, 1982; Policastro *et al.*, 1983), but in a footnote at the end of the first of these references, the encoded product was reidentified as luteinizing hormone. Since the rat LH cistron given is perceived to correspond closely to that of the human (Chin *et al.*, 1983; Jameson *et al.*, 1984), it appears that the structure of chorionic gonadotropin remains unestablished. The location of the cleavage site was arbitrarily decided as being at the position suggested in the table, but perhaps it will prove to be before the TTC and TCC, as in the reference on the human gene (Policastro *et al.*, 1983).

The presequences for the four β subunits are also highly conserved evolutionarily, but not to the extent found in the preceding ones. The two representative genes of TSH transit peptides are obvious homologs, as are those encoding that peptide for LH, but identities between the pairs are rare. The most universal feature is the length, which is almost identical in all four sequences. Near the end of row 1 are two columns of similar codons, one involving CTC, ATG, and CTG, the other GTG, ATG, and CTG, and in row 2 there is one more, containing CTT and CTG. The only other resemblance in codon makeup is at the very 3′ end, where TCT, TTC, and TCC are found. Because the TSH structures represent cDNA corresponding to the mRNA, not the gene proper, any consistency in the placement of the intron with those of the two LH genes that may exist is not disclosed at present.

Although the mature sequences of these genes are of no concern here, it is pertinent to note that that of human chorionic gonadotropin β subunit stands out from the remainder of the family in having an extension of 87 nucleotides at the 3′ end (Lentz *et al.*, 1984). This addition encodes 29 amino acids, including four serines to which oligosaccharide moieties become attached after translation has been completed. The function of this peculiarity has not been determined.

The Prolactin Family of Hormones. The prolactin family of hormones includes a small number of polypeptides that have rather extensive sequence homology as well as a

degree of overlapping in biological activity (Martial *et al.*, 1979). Embraced in this category in addition to the type form are growth hormone (GH), chorionic somatomammotropin (CS; placental lactogen), and proliferin (PLF), all except the placental being secreted by the pituitary gland (Table 7.2). The importance of this group of hormones is reflected in the fact that in the bovine anterior pituitary the messenger for prolactin constitutes 60% of the total mRNA on polysomes (Sasavage *et al.*, 1981). At least in cattle, there appear to be multiple loci for these genes, but their location and arrangement in the genome have yet to be determined.

The degree of homology between genes depends on both the species being compared and the source organisms (Linzer and Talamantes, 1985). Among the prolactin presequences, the level of agreement between corresponding sites is particularly outstanding near the end of row 1 and again just after the onset of row 2. Comparable consistency can be observed in the transit peptide cistrons of the growth hormones at similar points, both among themselves and with those of prolactins. Here, as in the preceding family, the greatest points of fluctuation are those ending the presequence and the pair beginning the mature coding region, sites where least variation might be expected because of their involvement in cleavage during processing. As a whole in the entire presequence the proportion of codons for apolar amino acids is surprisingly low for a peptide of this function, running close to 55%, except in the bovine growth hormone sequence, in which a level of 73% is found.

The unexpected high degree of identity between the human growth hormone and chorionic somatomammotropin presequences led to a comparison of the two mature coding regions (Table 7.3). After viewing numbers of highly diverse macromolecules under identical names distinguished only by subscripts, as in the cytochromes P-450, finding two very similar ones bearing such distinctive designations comes as a jolt. As may be noted in Tables 7.2 and 7.3, not only do the presequences and mature genes display close resemblances, but so do the 3′ trains, including the rare, true palindromic sequences of some length shown boxed. Moreover, the level of kinship extends even into the 5′ leader (Table 7.10), typically a region of low evolutionary conservation. Hence, it would appear to be far more realistic to refer to these closely related hormones as growth hormones (or somatotropins) A and B, or perhaps pituitary and chorionic growth hormones. With the suggested change in names the similarities between the pair certainly could be better appreciated than under their current designations. Previously the somatomammotropin gene had been compared only with a prolactin sequence, from which it is obviously distinct (Cooke *et al.*, 1981).

Other Simple Diplomorphic Genes from Vertebrates. Quite a few genes of vertebrates encode products that bear transit peptides, and many more are becoming apparent as additional sequences are established. Among those also known to fall here is α_1-antitrypsin (Long *et al.*, 1984a), thymosin β_4 (Goodall *et al.*, 1985), several varieties of lipoproteins, such as E, A-IV, C-I, and C-II (Boguski *et al.*, 1984; Fojo *et al.*, 1984; Knott *et al.*, 1984a; Zannis *et al.*, 1984), the β subunit of muscle acetylcholine receptor (Tanabe *et al.*, 1984), and many other receptors. It should be noted, however, that not all lipoprotein genes are to be classified here, because some, including A-I and A-II, are more complexly structured and are accordingly examined in a later section. Still further vertebrate genes that current knowledge suggests belong here are those of the various caseins (L. Hall *et al.*, 1984b; Stewart *et al.*, 1984), but in reality they may not. Because

Table 7.2
Presequence Genes of the Prolactin Family of Hormones[a]

Row 1	
Prolactins	
Rat[b]	*ATG* AAC AGC --- CAG *GTG* TCA *GCC* CGG AAA *G GG* --- ACA --- --- *CTC CTC CTG CTG ATG*
Human pituitary[c]	*ATG* AAC *ATC* AAA AAA *GGA* TCG *CCA* TGG AAA *G***GG* --- TCC --- --- *CTC CTG CTG CTG CTG*
Human decidua[d]	(Incomplete) TCC --- --- *CTC CTG CTG CTG CTG*
Bovine[e]	*ATG* GAC AGC AAA --- *GGT* TCG TCG CAG AAA *G GG* --- TCC CGC *CTG CTC CTG CTG CTG GTG*
Growth hormones	
Bovine[f]	*ATG ATG GCT GCA* --- *GGC* --- *CCC CCG* --- ACC --- TCC --- --- *CTG CTC CTG GCT TTC*
Human[g]	*ATG* --- *GCT* ACA --- *GGC* TCC --- CGG --- ACG --- TCC --- --- *CTG CTC CTG GCT TTT*
Proliferins	
Murine[h]	*ATG CTC CCT* --- --- --- TCT *TTG ATT* CAA *CCA* TGC TCC --- --- TGG *ATA CTG CTC CTA*
Chorionic somatomammotropin	
Human[i]	*ATG* --- *GCT CCA* --- *G***GC* TCC --- CGG --- ACG --- TCC --- --- *CTG CTC CTG GCT TTT*

Row 2

Prolactins	
Rat	*ATG* TCA AAC *CTT CTG TTC* TGC --- --- --- --- CAA AAT *GTG* CAG ACC ↓ *CTG CCA*
Human pituitary	*GTG* TCA AAC *CTG CTG CTG* TGC --- --- --- --- CAG AGC *GTG GCC CCC* *TTG CCC*
Human decidua	*GTG* TCA AAC *CTG CTC CTG* TGC --- --- --- --- CAG AGC *GTG GCC CCC* *TTG CCC*
Bovine	*GTG* TCA AAT *CTA CTC TTG* TGC --- --- --- --- CAG *GGT GTG GTC* TCC ACC *CCC*
Growth hormones	
Bovine	*GCC* --- --- --- *CTG CTC* TGC *CTG CCC* TGG --- ACT CAG *GTG GTG GGC* *GCC TTC*
Human	*GGC* --- --- --- *CTG CTG* TGC *CTG CCC* TGG *CTT* CAA GAG *GGC* AGT *GCC* *TTC CCA*
Proliferin	
Mouse	*CTA CTG* --- *GTG* AAC AGC TCG *TTA TTG* TGG --- AAG AAT *GTT GCC* TCA *TTT CCC*
Chorionic somatomammotropin	
Human	*GCC* --- --- --- *CTG CTC* TGC *CTG CCC* TGG *CTT* CAA GAG *GCT GGT GCC* *GTC* CAA

[a]Codons for hydrophobic amino acids are italicized and those for hydrophilic ones underscored. Vertical arrow points to cleavage sites. Asterisks indicate location of introns.
[b]Gubbins *et al.* (1979); Cooke *et al.* (1980).
[c]Takahasi *et al.* (1984).
[d]Cooke *et al.* (1981); Truong *et al.* (1984).
[e]Sasavage *et al.* (1981).
[f]Miller *et al.* (1980).
[g]Martial *et al.* (1979); Roskam and Rougeon (1979).
[h]Linzer and Nathans (1984).
[i]Selby *et al.* (1984).

Table 7.3
Relationships between Growth Hormone (GH) and Somatomammotropin (SMT) Genes[a]

Row 1	
Human GH[b]	TTC CCA ACC ATT CCC TTA TCC AGG CCT TTT GAC AAC GCT ATG CTC CGC GCC CAT CGT CTG CAC CAG CTG
Human SMT[c]	GTC CAA ACC GTT CCC TTA TCC AGG CTT TTT GAC CAC GCT ATG CTC CAA GCC CAT CGC GCG CAC CAG CTG
Row 2	
Human GH	GCC TTT GAC ACC TAC CAG GAG TTT GAA GAA GCC TAT ATC CCA AAG GAA CAG AAG TAT TCA TTC CTG CAG
Human SMT	GCC ATT GAC ACC TAC CAG GAG TTT*GAA GAA ACC TAT ATC CCA AAG GAC CAG AAG TAT TCA TTC CTG CAT
Row 3	
Human GH	AAC CCC CAG ACC TCC CTC TGT TTC TCA GAG TCT ATT CCG ACA CCC TCC AAC AGG GAG GAA ACA CAA CAG
Human SMT	GAC TCC CAG ACC TCC TTC TGC TTC TCA GAC TCT ATT CCG ACA CCC TCC AAC ATG GAG GAA ACG CAA CAG
Row 4	
Human GH	AAA TCC AAC CTA GAG CTG CTC CGC ATC TCC CTG CTG CTC ATC CAG TCG TGG CTG GAG CCC GTG CAG TTC
Human SMT	AAA TCC*AAT CTA GAG CTG CTC CGC ATC TCC CTG CTG CTC ATC GAG TCG TGG CTG GAG CCC GTG CGG TTC
Row 5	
Human GH	CTC AGG AGT GTC TTC GCC AAC AGC CTG GTG TAC GGC GCC TCT GAC AGC AAC GTC TAT GAC CTC CTA AAG
Human SMT	CTC AGG AGT ATG TTC GCC AAC AAC CTG GTG TAT GAC ACC TCG GAC AGC GAT GAC TAT CAC CTC CTA AAG

Row 6	
Human GH	GAC CTA GAG GAA GGC ATC CAA ACG CTG ATG GGG AGG CTG GAA GAT GGC AGC CCC CGG ACT GGG CAG ATC
Human SMT	GAC CTA GAG GAA GGC ATC CAA ACG CTG ATG GGG*AGG CTG GAA GAC GGC AGC CGC CGG ACT GGG CAG ATC
Row 7	
Human GH	TTC AAG CAG ACC TAC AGC AAG TTC GAC ACA AAC TCA CAC AAC GAT GAC GCA CTA CTC AAG AAC TAC GGG
Human SMT	CTC AAG CAG ACC TAC AGC AAG TTT GAC ACA AAC TCG CAC AAC CAT GAC GCA CTG CTC AAG AAC TAC GGG
Row 8	
Human GH	CTG CTC TAC TGC TTC AGG AAG GAC ATG GAC AAG GTC GAG ACA TTC CTG CGC ATC GTG CAG TGC CGC TCT
Human SMT	CTG CTC TAC TGC TTC AGG AAG GAC ATG GAC AAG GTC GAG ACA TTC CTG CGC ATG GTG CAG TGC CGC TCT
Row 9	
Human GH	GTG GAG GGC AGC TGT GGC TTC TAG CTG CCC GGG TGG CAT CCC TGT GAC CCC TCC CCA GTG CCT CTC
Human SMT	GTG GAG GGC AGC TGT GGC TTC TAG GTG CCC GAG TAG CAT CCT -GT GAC CCC TCC CCA GTG CCT CTC
Row 10	
Human GH	CTG GCC CTG GAA GTT GCC ACT CCA GTG CCC ACC AGC CTT GTC CTA ATA AAA TTA AGT TGC ATC AAA AAA
Human SMT	TTG GCC CTG -AA GGT GCC ACT CCA GTG CCC ACC AGC CTT GTC CTA ATA AAA TTA AGT TGT ATC ATT TCA

[a]Asterisk indicates the location of an intron. Vertical arrow indicates cleavage site.
[b]Roskam and Rougeon (1979).
[c]Selby *et al.* (1984).

of the complex cleavages that can be produced by plasmin, at least β-casein may eventually prove to be encoded by a member of the cryptomorphic genes (Section 7.6.1). A similar condition is prevalent in the haptoglobin genes of man (Bensi *et al.*, 1985), and possibly also in that for human pancreatic polypeptide (Boel *et al.*, 1984).

7.1.2. Simple Diplomorphic Genes from Seed Plants

In the seed plants, too, there are a number of newly translated proteins that bear transit sequences. These include not only those encoded by nuclear genes for use in the mitochondrion or chloroplast, as mentioned in the preceding chapter, but numerous types transcribed or processed within the endoreticulum or dictyosomes also have been shown to possess that feature. In addition, it is not unlikely that some types that are excreted through the cell membranes also have appended presequences, but no gene structures for any in that category have been determined, except perhaps the lectin cistron discussed below. Undoubtedly the most thoroughly documented are the seed-storage proteins, which receive attention first, while some of the energy-related types then supplement earlier discussions.

Seed-Storage Proteins. The seed-storage proteins comprise several large, complex families, since each major group of plants appears to have one or more distinctive types. Some of these are of a more complex organization and are analyzed in Section 7.6.2, but many appear to belong in the present category. In maize, zein, accounting for 50% of the total endosperm protein, is the primary representative, being synthesized in the endosperm from 14 to 55 days after fertilization (Pedersen *et al.*, 1982; Spena *et al.*, 1983). Within the haploid genome of this monocot are found about 120 copies of the gene, located on at least three of the ten chromosomes, with much deviation from one copy to another. The two presequences given in Table 7.4 are from a pair of adjacent cistrons. It is improbable that the translation termination signal TAA actually exists at the 3′ end of this sequence in the E19 species as given in the original reference (Spena *et al.*, 1983), for that would end the translational processes. In all likelihood this represents either a typographical error or a misreading of the chromatographic blocks and probably should be TAC as in the E25 structure.

The next most abundant protein in the endosperm of maize, glutelin-2, accounts for 15% of the total. Its presequence is quite unlike that of the zeins in being much shorter, in this respect more closely resembling the corresponding parts of dicotyledonous seed-storage proteins (Prat *et al.*, 1985), with which it is placed in the table. Homology with those sequences is perceived to be only occasional, so that no kinship is in evidence. A second monocot, barley, has the seed endosperm enriched in an alcohol-soluble class of proteins known as hordeins; these comprise a complex mixture of polypeptides that fall into four principal subdivisions, the B and C types providing 95% of the total present (Kreis *et al.*, 1983; Forde *et al.*, 1985). However, the gene structure has yet to be fully determined. In a third grain, hexaploid wheat, gliadins translated on the endoreticulum of endosperm cells provide the major storage component. These, too, are a complex group, being separable into 35–50 components that can be arranged into three subfamilies (Kasarda *et al.*, 1984; Okita *et al.*, 1985). Only two of the trio have had a gene of a representative sequenced, one of the α subfamily and another of the α/β type (Rafalski *et al.*, 1984; Sumner-Smith *et al.*, 1985), both of which are included in Table 7.4. The

Table 7.4
Transit-Peptide Genes of Seed Storage Proteins[a]

Row 1

Zein E19[b]	*ATG GCA GCC* AAA *ATA TTT* TGC *CTC CTT ATC CTC CTT GGT CTT* TCT *GCA* AGT *GCT GCT* ACG *GCG*
Zein E25[b]	*CTG GCA GTC* AAA *ATA TTT* TGC *CTC CTT ATG CTC CTT GCT CTT* TCT *GCA* AGT *GCT GCT* AAC *GCG*
Gliadin α/β[c]	*ATG* --- --- AAG ACC *TTT CTC ATC CTT GTC CTC CTT GCT ATT GTG GCG* --- --- --- --- ---
Gliadin α[d]	*ATG* --- --- AAG ACC *TTT CTC ATC CTT GCC CTC CTT GCT ATC GTG GCA* --- --- --- --- ---
Phaseolin[e]	*ATG ATG* AGA *GCA* AGG *GTT CCA CTC CTG TTG CTG GGA* --- *ATT CTT* --- --- --- --- --- ---
Lectin[f]	*ATG* CAT GAT CAT *GGC TTC CTC* CAA *GTT* ACT *CTC CCT* AGC *CCT CTT CCT* --- --- --- --- ---
Glutelin-2[g]	*ATG* --- AGG --- *GTG TTG CTC* --- *GTT GCC CTC* --- *GCT CTC TTG GCT* --- --- --- --- ---

Row 2

Zein E19	ACC *ATT TTC CCG* CAA TGC TCA CAA *GCT CCT ATA GCT* TCC *CTT CTT CCC CCG* TAA ↓ *CTC* TCA
Zein E25	ACC AAT *TTT CTG* CAA TGC TCA CAA GAT *CCA ATT GCT* TCC *CTT CTT CCC* TCA TAC *CTC* TCA
Gliadin α/β	ACC ACC --- --- --- --- --- --- *GCC* ACA ACT *GCA* --- --- --- --- --- --- *GTT* AGA
Gliadin α	ACC ACC --- --- --- --- --- --- *GCC* ACA ACT *GCA* --- --- --- --- --- --- *GTA* AGA
Phaseolin	--- --- *TTC CTG GCA* --- TCA --- *CTT* TCT *GCC* --- TCA *TTT GCC* --- ACT TCA *CTC* CGG
Lectin	--- --- --- --- --- TGC *GCT* TCT CAG *CCA* --- CGC --- AAA *CTC* --- --- --- AGC CAC
Glutelin-2	--- --- --- *CTC* --- --- *GCT* --- *GCG* --- AGC *GCC* --- --- --- --- --- TCC ACG CAT

[a]Codons for hydrophobic amino acids are italicized and those for hydrophilic ones are underscored. Arrow indicates the cleavage site.
[b]Spena *et al.* (1983).
[c]Rafalski *et al.* (1984).
[d]Kasarda *et al.* (1984).
[e]H. Paaren, personal communication (1985).
[f]Hoffman *et al.* (1982).
[g]Prat *et al.* (1985).

genomic numbers often vary with the cultivar and have not been firmly documented; the location of one gene, that of α-1Y, has been demonstrated to be on chromosome 6A of wheat (Anderson *et al.*, 1984).

The dicotyledonous seed-storage proteins have not been so thoroughly explored at the level of the gene as have their counterparts; however, in addition several gene structures that have been established are of the more complex type mentioned earlier. Hence, only a single representative of the simple group from this source is currently available. That one, from the French bean (*Phaseolus vulgaris*), is one of a group of polypeptides termed phaseolin that comprises about half of the storage proteins of the seed (Talbot *et al.*, 1984). These are deposited rapidly, beginning when the cotyledons are about 7 mm in length and continuing until they have attained 17–19 mm (Slightom *et al.*, 1983). Now that the presequence as well as the full coding region of the gene has been determined (H. Paaren, personal communication), the former can be noted to be quite distinct from the four from monocotyledonous sources. Among the especially outstanding distinctions is the presence of two methionine codons at the 5′ end, followed by two for arginine separated by one triplet, but none for lysine. Despite the sequence being aligned as fully as possible with the others, very few corresponding sites can be noted. Still one other dicotyledonous member is known in part, conglycinin of the soybean, a trimeric protein of αα′β constitution; while much of the gene structure has been determined, the presequence is missing (Schuler *et al.*, 1982). Since this protein is trimeric rather than monomeric like the rest, it obviously represents a family separate from those given here. Finally, among the albumins that are also abundant in beans is a substance called lectin, a protein widely spread throughout the living world. This really serves as an immune-related substance rather than as a source of food for the growing embryo (Hoffman *et al.*, 1982), but nevertheless, shares many features with the others of Table 7.4.

Considering the diversity of the sources, an unexpectedly high level of homology among the transit-peptide genes of the three species of seed-storage proteins is revealed by examination of the tabulation, but it is particularly those from grains that display close kinship (Reeck and Hedgcoth, 1985). The initiating codon, which is CTG for leucine in zein E25, is usually followed shortly by a codon for a charged amino acid, even in the lectin. Thereafter in row 1 a number of further correspondences can be noted among the triplets, nearly all of which specify hydrophobic monomers. In the second row the level of evolutionary conservation is greatly reduced, as is the proportion of codons for apolar amino acids. One feature that particularly merits notice is the relative frequency of the CT- family of codons for leucine, while the others for that amino acid, TTA and TTG, are not employed at all.

The Ribulose-Bisphosphate Carboxylase Family. In the preceding chapter one of the important energy-related families of proteins was seen to be that of ribulose-1,5-bisphosphate carboxylase, one of the principal genes involved in carbon dioxide fixation. This enzyme will be recalled as being comprised of eight copies each of large and small subunits, the former being encoded in the chloroplast genome, the latter in the nucleus. Hence, it is the minor component whose gene is provided with a presequence, and thus requires attention here. Only four full gene structures for the transit peptide have been determined, plus an additional partial one. Since three of the total are from wheat, only three different source organisms are thus represented in Table 7.5.

Here, as in the seed proteins of monocotyledons, sequence homology is at a high

Table 7.5
Gene Sequences for the Transit Peptides of Ribulose-1,5-Bisphosphate Carboxylase[a]

Row 1	
Pea *rbcS*[b]	*ATG GCT* TCT *ATG ATA* TCC TCT TCC *GCT GTG* ACA ACA *GTC* AGC CGT *GCC* TCT AGG *GGC* CAA TCC
Lemna G-3[c]	*ATG GTT* TCC ACC *GCC GCC GTG GCC* CGC *GTC* CGC CCT *GCC* CAG ACC AAC *ATG GTG GGC* --- ---
Wheat *W9*[d]	*ATG GCC* --- --- --- --- --- --- --- --- --- --- --- *CCA GCC GTG ATG GCT* TCT TCC
Wheat *WS4.3*[d]	*ATG GCC* --- --- --- --- --- --- --- --- --- --- --- *CCC GCC GTG ATG GCT* TCG TCG
Row 2	
Pea *rbcS*	*GCC* --- *GCA CTG GGC CCA TTC GGC GGC CTC* AAA TCC *ATG* ACT *GGA TTC CCA GTG* --- AAG AAG
Lemna G-3	*GCC* --- --- --- --- --- *TTC* AAC *GGG* TGC CGC TCC TCC *GTC GCC TTC CCC GCC* ACC CGC AAG
Wheat *234*[e]	(Incomplete)*GTC GCT CCT TTC* CAG *GGG CTC* AAG TCC ACC *GCC GGC CTC CCC GTC* AGC CGC CGC
Wheat *W9*	*GCC* ACC ACC *GTC GCG CCC TTC* CAG *GGG CTC* AAG TCG ACC *GCC GGC CTC CCC ATC* AGC TGC CGC
Wheat *WS4.3*	*GCT* ACC ACC *GTC GCA CCC TTC* CAG *GGT CTC* AAG TCC ACA *GCC GGC CTG CCC GTC* AGC CGC CGC
Row 3	
Pea *rbcS*	*GTC* AAC ACT GAC --- *ATT* ACT TCC *ATT* ACA AGC AAT *GGT GGA* AGA *GTA* AAG TGC *ATG* CAG*
Lemna G-3	*GCC* AAC AAC GAT *TTG* TCG ACT --- *CTC CCC* AGC TCC *GGC GGC* AGG *GTT* AGC TGC *ATG* CAG
Wheat *234*	--- TCC AAC *GGC GCT* AGC *CTC GGC* AGC *GTC* AGC AAC *GGT GGA* AGG *ATC* AGG TGC *ATG* CAG
Wheat *W9*	--- TCC *GGC* AGC ACC *GGC CTC* AGC AGC *GTC* AGC AAT *GGC GGA* AGG *ATC* AGA TGC *ATG* CAG
Wheat *WS4.3*	--- TCC --- AGG GGC AGC *CTC GGC* AGC *GTC* AGC AAC *GGC GGA* AGG *ATC* AGG TGC *ATG* CAG

[a]Codons for hydrophilic amino acids are underscored and those for hydrophobic ones are italicized. Asterisk indicates the location of an intron. The arrow indicates cleavage site.
[b]Coruzzi *et al.* (1983, 1984). [c]Stiekema *et al.* (1983). [d]Broglie *et al.* (1983). [e]Smith *et al.* (1983).

Table 7.6
Presequences of Cytochromes and Related Proteins[a]

Row 1	
Yeast$_N$ cyt c_1[b]	*ATG TTT* TCA AAT *CTA* TCT AAA CGT TGG *GCT* CAA AGG ACC *CTC* TCG AAA AGT *TTC* TAC TCT ACC *GCA* ACA
Pea$_c$ cyt f[c,e]	*ATG* GAT AGG GAA *CTG* --- --- --- AGT AAC *CTA CCT* AAT *CTT ATT GTA* GAA *ATT TTC* AGG *ATC* AAG GAT
Spinach$_c$ cyt f[d]	*GTG* GAT AGG GAA *CTT* --- TAC --- TAG CAA *CCT* ACC CAA *TTT ATT GTA* TAA *ATT TTC GGA ATC* AAT *GGT*
Wheat$_c$ cyt f[e]	*GTG* TAT AGG GAA *CTA* GAT TAC *GTT* AGC TAC *CTA* TCT AAT *XTT ATT GTA* GAA AXX *TTC* TGG *ATC* TGC GAT
Yeast cyt c peroxidase[f]	*ATG* ACT ACT *GCT GTT* AGG *CTT TTA CCT* TCA *CTG GGC* AGA ACC *GCC* CAT AAG AGG TCT *CTC* TAC *CTG TTC*
Bovine cyt P-450$_{scc}$[g]	*ATG CTA GCA* AGG *GGG CTT CCC CTC* CGC TCA *GCC CTG GTC* AAA *GCC* TGC *CCA CCC ATC CTG* AGC ACA *GTG*
Yeast cyt c oxidase IV[h]	*ATG CTT* TCA *CTA* CGT CAA TCT *ATA* --- --- --- --- AGA *TTT TTC* --- AAG *CCA GCC* --- ACA --- AGA
Bovine cyt c oxidase IV[i]	*ATG TTG GCA* ACC AGA *GTA TTT* AGC --- --- --- --- *CTG ATT GGT* --- AGG CGT *GCA* --- *ATC* --- ---
Yeast cyt c oxidase VI[j]	*ATG TTA* TCA --- AGG *GCC* --- *ATA* --- --- --- *TTC* AGA AAT --- --- --- *CCA GTT* --- *ATA* AAT AGA
Row 2	
Yeast$_N$ cyt c_1	*GGT GCT GCT* AGT AAA TCT *GGC* AAG *CTT* ACT CAA AAG *CTA GTT* ACA *GCG GGT GTT GCT GCC GCC GGT ATC*
Pea cyt f	TGT ACC *ATG* CAA ACT AGA AAT *GCT TTT* TCT TGG *ATA* AAG AAA GAG *ATT* ACT CGA TCT *ATT* TCC *GTA TTG*
Spinach cyt f	TGG ACT *ATG* CAA ACT *ATA* AAT ACC *TTT* TCT TGG *ATA* AAA GAA CAG *ATT* ACT CGA TCC *ATT* TCC *ATA* TCA
Wheat cyt f	TGG ACT *ATG* GAA AAT AGA AAT ACT *TTT* TCT TGG *GTA* AAG GAA CAG *ATA* ACT CGA TCG *ATT* TCT *GTA* TCG
Yeast cyt c peroxidase	TCC *GCT GCT GCT GCT GCT GCT GCT GCT GCT GCA* ACT *TTT GCT* TAC TCG CAA TCC CAC AAG AGA TCA TCG
Bovine cyt P-450$_{scc}$	*GGG* GAG *GGC* TGG *GGC* CAC CAC AGG *GTG GGC* --- --- --- --- --- --- --- --- --- --- --- --- ---
Yeast cyt c oxidase IV	--- ACT *TTG* TGT --- AGC --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
Bovine cyt c oxidase IV	--- --- --- --- --- --- --- --- --- TCC --- --- --- --- --- --- ACC --- --- --- TCG *GTG* ---
Yeast cyt c oxidase	--- ACT *TTA TTG* --- AGA *GCC* AGA *CCT GGT GCT* TAT CAT *GCA* ACT AGA *TTG* ACT --- AAA --- --- ---

Row 3

Yeast cyt c_1	ACC *GCA* TCG ACT *TTA* *CTC* TAT *GCA* GAC TCA *TTA* ACT --- --- --- --- --- --- *GCC* --- GAA *GCT*	*ATG* ACC
Pea cyt *f*	*CTC* *ATG* *ATC* TAT *ATA* *ATA* ACT CGA *GCA* CCC *ATT* TCA --- --- --- --- --- --- --- --- AAT *GCA*	TAT *CCC*
Spinach cyt *f*	*CTT* *ATA* TTA TAT *ATA* *ATA* ACT CGG TCA TCC *ATT* *GCG* --- --- --- --- --- --- --- --- AAT *GCC*	TAT *CCC*
Wheat cyt *f*	*ATC* *ATG* ATA TAC *GTA* *ATA* ACT CGG ACA TCT *ATT* TCA --- --- --- --- --- --- --- --- AAT *GCA*	TAT *CCC*
Yeast cyt *c* peroxidase	TCT TCT *CCT* *GGG* *GGT* *GGT* AGT AAC CAC *GGA* TGG AAC AAC TGG *GGG* AAG *CCA* *GCT* *GCT* *TTG* *GCT* TCC	ACT ACA
Bovine cyt P-450$_{scc}$	--- --- --- --- --- --- ACT *GGA* GAG *GGA* --- --- --- --- --- --- --- --- --- --- *GCT* *GGC*	*ATC* TCC
Yeast cyt *c* oxidase IV	--- --- --- --- --- --- TCT AGA TAT --- --- --- --- --- --- --- --- --- --- --- *CTG* *CTT*	CAG CAA
Bovine cyt *c* oxidase IV	--- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- TGT *GTT* CGG	*GCC* CAT
Yeast cyt *c* oxidase VI	--- --- --- --- --- AAT ACG *TTT* *ATT* CAA AGT --- --- --- --- --- --- --- --- AGG AAG TAT	TGT GAC

[a]Codons for apolar (hydrophobic) amino acids are italicized and those for hydrophilic ones are underscored. N, nuclear; C, chloroplast; cyt, cytochrome. Vertical arrow shows the cleavage site.
[b]Sadler *et al.* (1984).
[c]Willey *et al.* (1984a).
[d]Alt and Herrmann (1984).
[e]Willey *et al.* (1984b).
[f]Kaput *et al.* (1982).
[g]Morohashi *et al.* (1984).
[h]Maarse *et al.* (1984).
[i]Lomax *et al.* (1984).
[j]Wright *et al.* (1984).

level, especially at the 3′ end, where a number of sites are invariant or virtually so. In addition, several unusual features for a peptide gene of this sort are to be observed, chief among them being the consistent placement of codons for arginine, two double columns of which occur, one set at the end of row 2, and another, which is interrupted, similarly placed in row 3. The greatest distinctions between sequences from the dicotyledonous and monocotyledonous sources are in the 5′ end, where the pea and duckweed representatives are greatly elongated relative to the others. All these presequences are obviously longer than any that have been viewed earlier. As a whole, codons for hydrophobic amino acids are more prolific in the third row, but the CT- family for leucine is not especially abundant proportionately with the rest.

7.1.3. Presequences of Cytochrome-Related Genes

As pointed out in the preceding chapter, many cytochromes and related proteins that function in the mitochondrion or chloroplast are encoded in the nuclear genome, such as the ribulose-1,5-bisphosphate carboxylase small subunit in the foregoing section, and similarly are translated in the cytoplasm of the cell. Hence, many of the genes whose structures were examined in Chapter 6 receive attention here in connection with their presequences, whose products function in the penetration through the organellar membrane. As may be seen in Table 7.6, quite a diversity of types bear this appendage, the nine structures shown representing six different species.

General Characteristics. The coding regions for the transit peptides vary extensively in length, the yeast cytochrome *c* peroxidase, with 67 codons, being the longest (Kaput *et al.*, 1982), and the bovine cytochrome *c* oxidase subunit IV, with 22, the shortest (Lomax *et al.*, 1984). Even the three representatives of cytochrome *f* from chloroplasts display unexpected slight differences in length, having a range from 57 in the pea to 60 in the wheat (Willey *et al.*, 1984a,b). The spinach example as given in the table differs from that indicated in the paper describing the gene sequence (Alt and Herrmann, 1984), since the investigation apparently believed the first ATG in row 2 to be the initiating codon, whereas in all likelihood it is the GTG of row 1, as in the wheat example (Willey *et al.*, 1984b).

Homologies of Structure. Homologies among the entire assemblage of nine are obviously lacking. Despite the differences in length, which permit much freedom in aligning the shorter components with the larger, it is not possible to bring sites together to produce a single full column of either identical or related codons. Not even the underscored triplets for lysine and arginine form vertical series in excess of four representatives. Furthermore, as a whole the cytochrome and related gene presequences are not outstandingly rich in codons of apolar amino acids. Indeed, one of those cited, the cistron for yeast cytochrome *c* oxidase subunit VI, consists of only 40% such triplets, and almost all have 50% or less codons for those hydrophobic amino acids. The sole exception is the gene for bovine P-450_{scc}, which encodes at a 70% level of apolarity (Morohashi *et al.*, 1984). The highest concentration of these units is medially, where the yeast cytochrome *c* peroxidase, for an extreme example, has an unbroken series of ten. As a general rule, the codon in the ultimate position at the cleavage site specifies a small apolar amino acid, but that of bovine cytochrome *c* oxidase subunit IV is for arginine and that of yeast subunit VI signals tyrosine. Similarly, the character of those in the penultimate site is variable, four being for hydrophobic, four for polar uncharged, and one for polar charged amino acids.

The three representatives for cytochrome *f* do not display a high level of evolutionary conservation, but vary rather freely. Perhaps the strangest condition is the rather frequent homologies between the pea sequence and the wheat, the spinach example often differing from the two others. While variation exists, the spinach sequence resembles that of wheat at eight sites, and that of the pea the same number of times; in contrast, the pea agrees with the wheat at 14 sites. As on other occasions, single characteristics here are disclosed as being totally unreliable as indicators of phylogenetic kinships.

Between the two representatives of cytochrome *c* oxidase subunit IV the level of homology is astonishingly low, complete agreement between corresponding sites occurring only at two or three triplets. Perhaps the two are actually two different subunits, which happen to show a like chromatographic pattern in the complete enzyme and consequently bear the same numerical designation.

7.1.4. Simple Diplomorphic Genes from Miscellaneous Sources

Comparatively few gene sequences have been established among eukaryotes other than yeasts, vertebrates, and seed plants, and of those that have been established, most either cannot be fitted into the "simple" category or lack transit-peptide portions, or the latter are too incomplete to be meaningful. Thus the four of Table 7.7 containing a pair each from fungal and insect sources must be considered representative on a preliminary basis. Only the two from the insect (*Bombyx mori*) are related, but even there the level of homology is not high. Nevertheless, those available contribute to the fabric of the total picture of presequence structure.

Two Unrelated Fungal Genes. Cutinase, one of the two simple diplomorphic genes whose sequences have been established from fungi, is an enzyme secreted by cells of a pathogenic species involved in the penetration of the host seed plant. The sequence of the transit peptide of this glycoprotein, which hydrolyzes the cutin coat of leaf cuticle and is essential for successful invasion, is from *Fusarium solani,* a pest of potatoes (Soliday *et al.*, 1984). The second is from a basidiomycete, *Schizophyllum commune,* and encodes an unnamed glycoprotein involved in the development of fruiting bodies (Dons *et al.*, 1984).

One unusual feature of the *Fusarium* sequence is the location of an intron between the transit peptide and mature coding portions. Of course, this would have no effect upon the proteolytic removal of the presequence, because the intron would have been removed prior to translation. Both genes are remarkable for the high concentration of codons for hydrophobic amino acids in their 5′ portions (row 1), where only one triplet for a polar charged type is present and two or four for uncharged polar amino acids. To the contrary, on the 3′ portion half or less of the codons encode apolar species. Here, too, only four triplets encoding a charged monomer are to be noted in the *Fusarium* gene and three in that of *Schizophyllum.*

Two Related Insect Genes. As stated before, the pair of genes from insects encode chorion proteins of the silkmoth, but the two presequences do not display the expected high level of homology (Iatrou *et al.*, 1984). Because of the extreme richness in cysteine (30%), the substances encoded by the mature genes are believed to contribute to the formation of the outer shell of the egg. Although both share this compositional feature, the mature coding sequences differ extensively, A being rich also in codons for glycine, while B is not so heavily marked with triplets for that amino acid. Thus they really form a small family of genes, rather than multiple copies of a single unit. As a result, it is less

Table 7.7
Presequence Genes from Miscellaneous Eukaryotes[a]

Row 1

Fungal genes	
Schizophyllum lG2[b]	*ATG CCC TTC* TCG *CTC GCC ATC CTT GCT CTC CCC GTC CTC GCG GCT GCG* ACT *GCG GTT CCC* CGC *GGC*
Fusarium cutinase[c]	*ATG* AAA *TTC TTC* --- *GCT* --- --- --- --- --- --- *CTC* ACC ACA *CTT CTC GCC GCC* ACG *GCT* TCG
Insect chorion genes	
Bombyx A[d]	*ATG TTT* ACC *TTC* --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- *GCT CTT CTC*
Bombyx B[d]	*ATG* --- --- --- --- *GCC* --- --- --- --- --- --- --- --- --- --- --- --- --- *GCT* --- ---

Row 2

Fungal genes	
Schizophyllum lG2	*GGC GCT TCC* AAG TGC AAC AGC *GGT CCC GTC* CAG TGC TGC AAC ACC *CTG GTC* GAC* ACT AAG
Fusarium cutinase	*GCT CTG CCT* ACT TCT AAC --- --- *CCT GCC* CAG GAG *CTT* GAG *GCG* CGC CAG *CTT* GGT AGA
Insect chorion genes	
Bombyx A	*CTT CTC* TGC *GTT* CAG *GGC* TGC *CTG ATC* CAA*AAT *GTG* TAC --- --- --- --- *GGT* CAG TGC
Bombyx B	AAA *CTC ATT GTC TTC GTC* TGC *GCC ATC GCC CTC GTG GCT* CAG*TCC *GTT TTG GGC* ACT *GGT*

[a]Codons for hydrophobic amino acids are italicized and those for hydrophilic ones are underscored. Asterisk indicates the location of an intron. Arrow indicates cleavage site.
[b]Dons *et al.* (1984).
[c]Soliday *et al.* (1984).
[d]Iatrou *et al.* (1984).

difficult to understand the low grade of homology that exists between the sequences for the transit peptides shown here. However, the correspondences are sufficient, especially in the similar—but not identical—location of an intron, to suggest that the two have had common ancestry. Codons for hydrophobic amino acids are numerous and distributed fairly uniformly through both presequences, but only one triplet for a charged monomer can be observed in the B gene and none in the other.

7.2. SIMPLE DIPLOMORPHIC GENES FROM PROTISTS AND PROKARYOTES

Although many interesting variations in presequence structure were found in the foregoing pages, no common theme was in evidence among the higher forms described there. However, it is possible that at the lower levels of phylogeny, the simpler organization that might exist in transit-peptide structures may better reveal the fundamentals. Accordingly, this section centers on presequences from the true yeasts and bacteria as the basis for comparison, using genes that encode products secreted from the cells into the environment. Since not all such secretions are provided with the simpler type of gene being considered at this time, only those cistrons whose products do not undergo further processing after removal of the transit protein receive attention now. Eventually many lower eukaryotic genes of this category should be available, for most, like the green alga *Chlamydomonas,* release an abundance of substances into the milieu (Voigt, 1985). Moreover, many antigens, either secreted or membrane-embedded, doubtlessly will prove to bear prepeptides, as one from *Plasmodium falciparum* has already proven to do (Hope *et al.*, 1985).

7.2.1. Diplomorphic Genes of Yeast and Bacteria

Yeast Secreted-Protein Genes. The first yeast genes for proteins secreted through the cell membrane consist of two cistrons for acid phosphatase, *PHO3* and *PHO5,* a tandemly repeated duplication (Arima *et al.*, 1983; Bajwa *et al.*, 1984). In addition, a third copy may be present, but its existence has not been firmly determined. No organization into operons exists, the *PHO* genes being located on linkage group II, while those for their controlling elements are dispersed throughout the genome. After translation (on endoreticulum?) the acid phosphatase is glycosylated and then transported through the cell membrane into the periplasmic space that lies beneath the cell wall. As may be observed in Table 7.8, the sequences of the transit peptides display approximately the same level of homology (82%) that has been reported for the mature coding portion (Bajwa *et al.*, 1984). Despite the similarity of structure, which extends deeply into the flanking regions, the controls differ widely, for *PHO3* is independent of exogenous influences, whereas *PHO5* is governed by a series of positively and negatively acting products of regulatory genes. The noteworthy features of the presequences include a codon for a basic amino acid near the 5′ end and a level of 75% or less presence of triplets for hydrophobic monomeric units. Of the latter class, those for valine (GTN) and alanine (GCN) are outstandingly abundant.

The presequence of the third gene (*SUC2*) in Table 7.8 has much the same structural

Table 7.8
Transit Peptide Sequences of Yeast Secreted-Protein Genes[a]

Row 1	
Yeast PHO3[b]	*ATG TTT* --- AAG --- TCT *GTT GTT* TAT TCG *GTT*
Yeast PHO5[b,c]	*ATG TTT* --- AAA --- TCT *GTT GTT* TAT TCA *ATT*
Yeast SUC2[d]	*ATG CTT TTG* CAA *GCT TTC CTT TTC CTT TTG GCT*
Row 2	↓
Yeast PHO3	*CTA GCC GCT GCT TTA GTT* AAT *GCA* *GGT* ACA
Yeast PHO5	*TTA GCC GCT* TCT *TTG GCC* AAT *GCA* *GGT* ACC
Yeast SUC2	*GGT TTT GCA GCC* AAA *ATA* TCT *GCA* TCA *ATG*

[a]Codons for hydrophilic amino acids are underscored and those for hydrophobic ones are italicized.
[b]Bajwa *et al.* (1984).
[c]Arima *et al.* (1983).
[d]Taussig and Carlson (1983).

properties as those just described, including the relatively short total length and single codon for a basic amino acid (Taussig and Carlson, 1983). The latter, however, is located close to the 3′ end, rather than near the 5′ end. Additionally, the level of the hydrophobicity is higher in the present case, all but three of the 19 codons being for apolar monomers. Triplets for leucine (CTN, TTA, and TTG) constitute more than 25% of the entire structure and those for alanine (GCN) about 20%, but none for valine (GTN) is to be seen. SUC2 encodes invertase, which is glycosylated when secreted, but not when retained intracellularly.

Bacterial Secreted-Protein Genes. Any illusions about primitive presequences being consistently short as implied by the yeast examples are quickly dispelled when the gene structures for bacterial secreted proteins are viewed. While the first three listed in Table 7.9 are of comparable length (18–20 codons), that of the *E. coli* gene for F pilin (*traA;* Frost *et al.*, 1984) has 50 and that of the *Bacteroides nodosus* pilin cistron only seven (Elleman and Hoyne, 1984). Nevertheless, some similarity in structure can be noted, particularly in the presence of at least one codon for a basic amino acid in proximity to the 5′ end. But again there are exceptions, for three of those given (*Pseudomonas aeruginosa* exotoxin A and *E. coli ompT* and *traA*) lack that feature (Gordon *et al.*, 1984; Gray *et al.*, 1984). The first of these lacks such triplets completely, but *traA* has a pair medially and *ompT* has two spaced ones before and at the 3′ terminus. Thus the presequences of bacteria are not consistently divided into charged and hydrophobic sectors as sometimes described (Michaelis and Beckwith, 1982; Emr and Silhavy, 1983).

Among those included in the table are several enterotoxins secreted through the cell walls into the environment, such as that of *Vibrio cholerae* (*V.c.*) and two from *E. coli* (*E.c.*). The cholera enterotoxin is comprised of a single subunit A and five of the much smaller B, the transit peptide portion of each of which has been established (Gennaro and

Table 7.9
Presequences of Bacterial Genes for Secreted Products[a]

Row 1	
Toxins	
V.c. enterotoxin A1[b]	*ATG --- --- --- GTA AAG --- --- --- --- --- --- --- --- --- ATA ATA --- TTT*
V.c. enterotoxin B[e]	*ATG --- --- --- ATT AAA TTA AAA --- --- --- --- --- --- --- --- --- --- --- TTT*
E.c. enterotoxin *LTA*[d]	*ATG --- --- --- --- AAA --- AAT --- --- --- --- --- --- --- --- ATA* ACT *--- TTC*
E.c. enterotoxin *LTB*[d]	*ATG --- --- --- AAT AAA GTA AAA --- --- --- --- --- --- --- --- TGT TAT --- GTT*
P.a. exotoxin A[e]	*ATG ---* CAC CTG *--- --- ATA CCC* CAT *---* TGG *--- ATC CCC CTG GTC GCC* AGC *CTC*
C.d. toxin 228[f]	*GTG ---* AGC *---* AGA AAA *--- --- --- CTG TTT --- GCG* TCA *ATC TTA ATA GGG GCG*
Enzymes	
B.p. xylanase A[g]	*ATG* AAT *TTG ---* AGA AAA *TTA* AGA *CTG TTG TTT GTG ATG ---* TGT *ATT GGA CTG ---*
E.c. lamB[h]	*ATG ATG ATT* ACT CTG CGC *---* AAA *--- CTT CCT CTG GCG GTT GCC GTC GCA --- ---*
B.s. α-amylase[i]	*ATG --- TTT GCA* AAA CGA *TTC* AAA ACC TCT *TTA CTG CCG TTA TTC GCT GGA TTT TTA*
B.e. α-amylase[j]	*ATG --- --- ---* AAA CAA CAC AAA CGC *CTT* TAT *GCC* CGA *TTG CTG CCG CTG TTA TTT*
Outer membrane proteins	
E.a. ompA[k]	*ATG --- --- ---* AAA AAG ACA *--- --- --- GCT --- ATC --- GCG ATT GCA --- GTG*
S.t. ompA[l]	*ATG --- --- ---* AAA AAG ACA *--- --- --- GCT --- ATC --- GCG ATT GCA --- GTG*
E.c. ompF[m]	*ATG ATG --- ---* AAG CGC *---* AAT *--- --- ATT CTG GCA --- GTG ATC GTC --- CCT*
E.c. ompC[n]	*ATG --- --- ---* AAA *--- GTT* AAA *GTA CTG* TCC *--- --- CTC CTG --- GTC --- CCA*
E.c. ompT[o]	*ATG* TGT *CTC* AGT *--- --- TTT --- GTC CCT CTT --- --- --- --- --- --- --- TTT*

(*continued*)

Table 7.9 (Continued)

E.c. traA[p]	*ATG* AAT *GCT GTT TTA* AGT *GTT* CAG *GGT GCT* TCT *GCG CCC GTC* AAA AAG AAG TCG *TTT*[u]
B.n. pilin[q]	*ATG* --- --- --- AAA AGT --- --- --- --- --- --- --- --- --- --- --- --- ---
E.c. fimA[r]	*ATG* --- --- --- AAA --- *ATT* AAA ACT *CTG GCA* --- *ATC GTT GTT* --- --- *CTG* TCG
E.c. fimpA[s]	*ATG* --- --- --- AAA AAA *GCA* --- --- *TTC TTA* --- --- *TTA GCA* *GTT TTT TTT*
S.a. protein A[t]	*TTG* --- --- --- AAA AAG --- AAA AAC *ATT* TAT TCA *ATT* CGT AAA *CTA GGT GTA GGT*[v]

Row 2

Toxins

V.c. enterotoxin A1	*GTG* --- *TTT TTT ATT* --- *TTC TTA* TCA TCA *TTT* --- TCA TAT *GCG* AAT GAT
V.c. enterotoxin B	*GGT GTT TTT TTT* ACA *GTT TTA CTA* TCT TCA *GCA* TAT *GCA* CAT *GGA* ACA *CCT*
E.c. enterotoxin *LTA*	*ATT* --- *TTT TTT ATT* --- *TTA TTA GCA* TCG *CCA* --- *TTA* TAT *GCA* AAT *GGC*
E.c. enterotoxin *LTB*	*TTA* --- *TTT ACG GCG* --- *TTA CTA* TCC TCT *CTA* TGT *GCA* TAC *GGA* *GCT CCC*
P.a. exotoxin A	*GGC* --- *CTG CTC GCC* --- *GGC GGC* TCG TCC --- --- *GCG* TCC *GCC* *GCC* GAG
C.d. toxin 228	*CTA CTG GGG ATA GGG* --- *GCC CCA CCT* TCA --- --- *GCC* CAT *GCA* *GGC GCT*

Enzymes

B.p. xylanase A	ACG *CTT* --- *ATA CTG* --- --- ACG *GCT GTA CCA* --- *GCC* CAT *GCG* AGA ACC
E.c. lamB	--- --- --- *GCG GGC* --- *GTA ATG* TCT *GCT* CAG --- *GCA ATG GCT* *GTT* GAT
B.s. α-amylase	*TTG CTG TTT* TAT *TTG GTT* --- *CTG GCA GGA CCG* --- *GCG GCT GCG* AGT *GCT*

B.l. α-amylase	*GCG CTC* --- *ATC TTC TTG CTG CCT* CAC TCT *GCA* --- *GCT GCG GCG*	*GCA* AAT
Outer membrane proteins		
E.a. ompA	*GCA CTG GCT GGC TTC* --- *GCT* --- ACC *GTA* --- --- *GCG* CAG *GCC*	*GCT CCG*
S.t. ompA	*GCA CTG GCT GGT TTC* --- *GCT* --- ACC *GTA* --- --- *GCG* CAG *GCC*	*GCT CCG*
E.c. ompF	*GCT CTG TTA GTA GCA* --- *GGT* --- ACT --- --- --- *GCA* AAC *GCT*	*GCA* GAA
E.c. ompC	*GCT CTG CTG GTA GCA* --- *GGC* --- --- *GCA* --- --- *GCA* AAC *GCT*	*GCT* GAA
E.c. ompT	--- --- TGT ACT --- --- --- AAA AAC *ATA* --- --- *GTA TTG* AGG	*ATA* ACC
E.c. traA	*CCG GCT GCT GTT CTG ATG ATG TTC TTC CCG* CAG *CTG GCG ATG GCC*	*GCC GGC*
B.n. pilin	--- --- --- --- --- --- --- --- *TTA* --- CAA --- --- AAA *GGT*	*TTC* ACC
E.c. fimA	*GCT CTG* TCC --- CTC --- --- AGT TCT --- ACA *GCG GCT CTG GCC*	*GCT GCC*
E.c. fimpA	*CTC* --- --- ACT *GGG* --- *GGC GGG* --- *GTT* --- --- TCT CAC *GCT*	*GCG GTT*
S.a. protein A	*GGT* ACA *TTA* CTT *ATA* TCT *GGT GGC GTA* ACA *CCT GCT GCA* AAT *GCT*	*GCG* CAA

[a]Codons for hydrophilic amino acids are underscored and those for hydrophobic ones are italicized. Abbreviations: *V.c., Vibrio cholerae; E.c., Escherichia coli; P.a., Pseudomonas aeruginosa; C.d., Corynebacterium diphtheriae; B.p., Bacillus pumilus; B.s., Bacillus subtilis; B.l., Bacillus licheniformis; E.a., Enterobacter aerogenes; S.t., Salmonella typhimurium; B.n. Bacteroides nodosus; S.a., Staphylococcus aureus.*

[b]Lackman *et al.* (1984). [c]Gennaro and Greenaway (1983). [d]Yamamoto *et al.* (1984a,b). [e]Gray *et al.* (1984). [f]Kaczorek *et al.* (1983). [g]Fukusaki *et al.* (1984). [h]Emr and Silhavy (1983). [i]Ohmura *et al.* (1984). [j]Sibakov and Palva (1984). [k]Braun and Cole (1983). [l]Freudl and Cole (1983). [m]Inokuchi *et al.* (1982). [n]Mizuno *et al.* (1983) [o]Gordon *et al.* (1984). [p]Finlay *et al.* (1984); Frost *et al.* (1984). [q]Elleman and Hoyne (1984). [r]Klemm (1984). [s]Mooi *et al.* (1984). [t]Uhlén *et al.* (1984). [u]Insert *TTT* TCC AAA TTC ACT CGT CTG AAT *ATG CTT* CGC *CTG GCT* CGC *GCA GTG ATC*. [v]Insert *ATT GCA* TCT *GTA* ACT *TTA*.

Greenaway, 1983; Lockman *et al.*, 1984). Actually the first of these gene products undergoes further processing into smaller components and thus is not truly a simple gene, but since this action is conducted only after the enterotoxin has entered a host cell, it appears valid to consider it simple as here, at least on a tentative basis. The two genes from *E. coli, LTA* and *LTB,* encode subunits corresponding to those from the cholera organism, which unite in the identical AB_5 ratio to produce the holoenzyme (Yamamoto *et al.*, 1984a,b). Since the end product has a comparable effect on the host, the two proteins are obviously members of the same family, a condition reflected in the presequences (Table 7.9). Although all four are distinctly related, the structures within each corresponding pair display a higher level of kinship to one another than between contrasting types. One of the outstanding chemical properties of this quartet is the inclusion of numerous codons for phenylalanine (TTT, TTC), especially in the A subunit of *V. cholerae.*

Two additional toxins are represented in the table by sequences of their transit peptide genes, one from *Pseudomonas aeruginosa* (*P.a.;* Gray *et al.*, 1984), the other from *Corynebacterium diphtheriae* (*C.d.;* Kaczorek *et al.*, 1983). Despite the similarities in subunit structure and activities that appear between the components of this pair and the preceding, very few homologies in nucleotide sequences exist. The diphtherial gene is somewhat similar to the others in having codons for a basic amino acid near the 5′ terminus, but here two are found side by side. Among the numerous additional divergences evidenced is the nearly complete lack of codons for phenylalanine, which are displaced by those for serine (TCN, AGC, AGT), alanine (GCN), and leucine (CTN). Glycine triplets are also moderately abundant in the *Pseudomonas* cistron.

Bacterial Simple Enzyme Genes. Among the enzymes secreted into the environment or deposited in the outer membrane are several from diverse bacteria that are encoded by simple genes including presequences (Table 7.9). The first of these, from *Bacillus pumilus* (*B.p.*), encodes a xylanase that breaks down xylans extracellularly (Fukusaki *et al.*, 1984); two for α-amylase from *Bacillus subtilis* and *B. lichenformis* that break down starches are similarly deposited in the medium (Ohmura *et al.*, 1984; Sibakov and Palva, 1984). On the other hand, a fourth representative, *lamB* from *E. coli,* is a component of the outer membrane that facilitates the passage of maltose from the surroundings into the cell (Emr and Silhavy, 1983). All are seen to encode largely hydrophobic amino acids, and the trio secreted into the milieu are alike in having three codons for basic components near the 5′ end. That which remains in the cell wall is distinct in having only two, but these are correspondingly located. It is of great importance to note that, when cloned in *E. coli,* the α-amylase gene of *B. subtilis* was produced and transported to the exterior in normal amounts, but the xylanase gene of *B. pumilus,* although transcribed at usual levels, accumulated in the cytoplasm instead of being secreted (Fukusaki *et al.*, 1984; Ohmura *et al.*, 1984).

Outer Membrane Proteins. Another large group of proteins that bear presequences are those of the cell coat, which must pass through the cytoplasmic membrane to attain their ultimate location. Out of the fairly large number whose gene structures have been determined, ten have been selected for Table 7.9 to show both the similarities and differences that characterize them. The two extremes of length of the presequences that have already been mentioned are found here and afford a first view of the contrasts that exist. Seven of the ten are of typical structure in having two or three codons for basic

amino acids close to the 5′ end; that of *B. nodosus* for pilin has only one, but an additional one occupies the penultimate site. And, as already noted, the *E. coli* cistrons for outer membrane protein T (*ompT*) and for pilin F (*traA*) are likewise irregular in the placement of such triplets.

The two sequences for outer membrane protein A (*ompA*) from *Enterobacter aerogenes* and *Salmonella typhimurium* that head the list exemplify the high degree of evolutionary conservation that usually exists between corresponding transit peptides, for they differ at only two sites (Braun and Cole, 1983; Freudl and Cole, 1983). In addition, they illustrate the conceptual model of a presequence, for they have, in order, the initiation codon ATG, a pair of triplets for lysine, followed by a single one for a neutral amino acid, an uninterrupted series of 12 or 13 for hydrophobic amino acids, and an absence of any for additional charged monomeric units. Similar continuous runs of 12 or 13 codons for hydrophobic amino acids occur in the *E. coli ompT* and *ompC* sequences (Inokuchi *et al.*, 1982; Mizuno *et al.*, 1983) and *traA* for pilin (Finlay *et al.*, 1984). The remaining four presequences do not fit this supposed standard pattern so well. Although largely endowed with codons for hydrophobic units, there are no such long, uninterrupted series, the longest stretch being six in *E. coli's fimA*, five in its *fimpA*, and six also in the *S. aureus* cistron for protein A (Klemm, 1984; Mooi *et al.*, 1984; Uhlén *et al.*, 1984). In contrast, the *B. nodosus* pilin gene has only a single such triplet in its interior—three of its total of seven codons are for hydrophobic, two for basic, and two for neutral amino acids. Consequently, this presequence is less than 45% hydrophobic.

Although not included in the table, presequences of related bacterial genes are not without interest. That of a strain of *E. coli* pathogenic in the urinary tract of man, encoding the $F7_2$ fimbrial subunit, consists of 21 codons, all but four of which are for hydrophobic amino acids (van Die and Bergmans, 1984). In contrast, that for a vitamin B_{12} receptor from the same organism has only 11 of its 20 codons for that type of monomer and two for basic amino acids near its 5′ end instead of only one (Heller and Kadner, 1985).

7.2.2. Transcriptional Control Signals of Diplomorphs

While there is no firm reason to suspect that initiation and termination of transcription in diplomorphic genes would differ from those of the simple type, the possibility exists nevertheless, and hence requires investigation. Regrettably, however, the same situation prevails here as in too many preceding aspects, in that too little experimental work has been conducted to present a clear picture, even preliminarily. Since the reports of the vast majority of eukaryotic diplomorphic genes provide neither promoter nor terminator sequences, even of a presumptive nature, attention here is necessarily confined to those from bacterial sources. Almost all of those provided in Table 7.10 are proposed signals and are given in italics, while the very few that have been experimentally determined are underscored. To the contrary, in the case of the trains, underscoring indicates nucleotide residues that are unpaired, while the arrows display regions of dyad symmetry.

Possible Promoters of Diplomorphic Genes. In Table 7.10 the promoters in the first 11 leader sections from simple diplomorphic genes of bacteria are presumptive only, being based on a resemblance to the TA-TA- box mentioned in many preceding discussions. Similarly, the ancillary signals are based on these suggested promoters and may or

Table 7.10
Flanking Sequences of Bacterial Simple Diplomorphic Genes[a]

	Ancillary site		Promoter	Start site
Leaders				
E.a. ompA P1[b]	GAGTTCACA CTTGTA--	----------	AGTTTCTAAC TAAGTT	GTAGAC TTT ACATCG
E.a. ompA P2[b]	CACTTGTAA GTTTCT--	----------	AACTAAGTTG TAGACT	TTACAT CGC CAGGGG
S.t. ompA P1[c]	GAGTTCACA CTTGTA--	----------	AGTTTCCAAC TACGTT	GTAGAC TTT ACATCG
S.t. ompA P2[c]	CACTTGTAA GTTTCC--	----------	AACTACGTTG TAGACT	TTACAT CGC CAGGGG
E.c. ompT[d]	GACTTAGAA GTTCCTAG	----------	AACGACATT- TTAAGT	CAACAA CTT ACCGCG
E.c. fimA[e]	TGTTTGATA TGTAAATT	----------	ATTTCTATTG TAAATT	AATTTC AC- ATCACC
E.c. fimpA[f]	GAGTTGTGT ATTC----	----------	GCTGGCACCT TATTAT	GCAGAT CCG GGCAAG
B.l. α-amylase[g]	TATTTGTT- AA------	----------	AAATTCAAAA TATTTA	TACAAT AGC ATGTGT
S.a. protein A P1[h]	ATTTTAGTA TTGCAATA	----------	CATAATTCGT TATATT	ATGATG ACT TTACAA
S.a. protein A P2[h]	GTATTGCAA TACATAA-	----------	TTCGTTATAT TATGAT	GACTTT ACA AATACA
P.a. exotoxin A[i]	ACATTCACC ACTCTG--	----------	CAATCCAGTT CATAAA	TCCCAT AAA GCCCTC
E.c. malK[j]	GGATTTAAG CCATCT--	----------	CCTGATGACG CATAGT	CAGCCC A(+48)
E.c. malE[j]	AGGAGGATG GAAAGAGG	TT--------	GCCGTATAAA GAAACT	AGAGTC CGT (+45)
Trains				
B.n. pilin[k]	TAGCTAGCTCT TAAAATGC	GAAAGCCTCTC	TCT TGAGAGGCTTTT	TTATGGTTTATTGTT
E.a. ompA[b]	GTTTCCTACGA TAA-----	-AAAACCCGCT	CGA TGCGGGTTTTTT	TTGGCCTGATTCTTG
S.t. ompA[c]	GTTTCCGTCTG ATA-----	AAAAACCCCGC	GTC -GCGGGTTTTTT	GCTCTGGTCTGGATG
E.c. fimA[e]	CCTACCCAGGT TCAGGA[l]+8	CGGGCAGGG	ATG CCCACCCTTGTG	CGATAAAAATAACGA
S.a. protein A[h]	CCTTAGGTGCA CGCT[m]+128	CTAAATGCACG	AGC AACATCTTTTGT	TGCTCAGTGCATTTT
P.a. exotoxin A[i]	CTGCCGCGACC	GGCCGGCT	CCCTTCGCAGG	AGC CGGCCTTCTCGG GGCCTGGCCATACAT

[a]Abbreviations: *E.a., Enterobacter aerogenes; S.t., Salmmella typhimurium; E.c., E. coli; B.l., Bacillus licheniformis; S.a., Staphylococcus aureus; P.a., Pseudomonas aeruginosa; B.n., Bacteroides iodosus.* P1, promoter 1; P2, promoter 2. Italics indicate presumptive, and underscores established, signals.
[b]Braun and Cole (1983).
[c]Freudl and Cole (1983).
[d]Gordon *et al.* (1984).
[e]Klemm (1984).
[f]Mooi *et al.* (1984).
[g]Sibakov and Palva (1984).
[h]Uhlen *et al.* (1984).
[i]Gray *et al.* (1984).
[j]Bedouelle *et al.* (1982).
[k]Elleman and Hoyne (1984).
[l]Insert seven residues.
[m]Insert 128 residues.

may not have any relation to the actual functional sequence. Hence, they should be viewed only as interesting possibilities and, as such, do not merit detailed attention at this time. Examination of the two gene leaders of the *E. coli* maltose system, *malE* and *malK*, is more meaningful, since the promoters and ancillary signals are deduced from actual experimentally determined start sites of transcription (Bedouelle *et al.*, 1982). Consequently, although they are not indisputably established, at least they have some basis in fact.

The location of the two start sites ~50 positions upstream of the initiator codon of translation is one reason for not taking the 11 presumptive ones too seriously, for most of those lie closer to that point. In neither of the two *mal* promoters is much resemblance displayed to the canonical TA-TA-, nor can standard features be found in the abutting nucleotides. The upstream ancillary signal shows still less constancy between the two underscored, and no similarity of any sort to the CAAT- box that is supposed to mark this site. As the result of all the uncertainties that exist, it is not possible to deduce whether these simple diplomorphs are transcribed by special mechanisms or whether the usual processes are active.

Possible Terminators of Transcription. The terminators of transcription are similarly hypothetical, since no experimental studies of any sort have been conducted on termination in this type of gene. Nevertheless, the ones that have been proposed have structural features identifiable with those that have been determined *in vitro* or *in vivo*. In all six examples of trains from diplomorphic genes given in Table 7.10, a stem-and-loop region is present, which almost always ends downstream in a series of Ts. Additionally, the first three representatives show a surprisingly high level of homology within this structure, but not elsewhere in the train. The implication of this degree of localized conservation is that these regions of dyad symmetry have a particularly important function, probably in termination as proposed. Only in three of them are unpaired bases (underscored) present, the *fimA* train of *E. coli* being unusual in having such unmated sites on each side of the stem.

7.3. ANALYSIS OF PRESEQUENCES OF SIMPLE GENES

In order to bring out any general trends that may exist in the transit-peptide portion of simple diplomorphic genes, the chief characteristics of those that have been reviewed are analyzed in Table 7.11. For the sake of clarity, apolar (hydrophobic) amino acids (Rose *et al.*, 1985) are indicated by the abbreviation A, charged polar (hydrophilic) ones by the letter C, and polar uncharged (neutral) ones by N. At the cleavage point, the nature is indicated of the monomers encoded by two codons on each side, that is, the last two of the presequence and the first couple of the mature coding sector of the gene. In the table, the presequences are divided into approximate halves to bring out differences in distribution that may exist, with the percent codons for apolar amino acids calculated for each "half" separately and again for the total structure.

Mean Properties of Presequences. When the properties of the presequences are viewed as described, it is noted that, in the vertebrate hormones of Table 7.11, the 3′ half encodes a higher proportion of hydrophobic amino acids than does the 5′ half, except for the β subunits of the two LH genes. This condition, however, is exceptional, for the mean

Table 7.11
Summary of Presequence Structure of Simple Diplomorphic Genes[a]

Table	Gene[b]	Number of codons	5'-Half				3'-Half				Total % Apolar	Cleavage site			
												Transit peptide		Mature sequence	
			A	N	C	%A	A	N	C	%A		Pen	Ult	Site 1	Site 2
7.1	Murine TSH α	24	7	3	3	54	7	4	0	64	58	N	N	A	N
	Rat GP α	24	7	3	3	54	7	4	0	64	58	N	N	A	N
	Human CG α	24	7	3	3	54	7	4	0	64	58	N	N	A	N
	Bovine GP α	24	7	3	3	54	7	4	0	64	58	N	N	A	N
	Murine TSH β	20	6	3	0	67	8	3	0	73	70	A	N	A	A
	Bovine TSH β	20	5	4	0	56	8	3	0	73	65	A	N	A	A
	Human LH β	20	7	1	1	77	8	3	0	73	75	N	A	A	C
	Rat LH β	20	7	1	1	77	7	4	0	64	70	N	A	N	C
7.2	Rat prolactin	28	9	5	2	56	5	7	0	44	50	N	N	A	A
	Human pituitary prolactin	29	9	5	3	53	7	5	0	56	55	A	A	A	A
	Bovine prolactin	30	9	5	4	50	7	5	0	56	53	A	N	N	A
	Bovine GH	26	11	2	1	79	8	4	0	63	73	A	A	A	A
	Human GH	26	8	4	1	62	7	5	1	52	58	N	A	A	A
	Murine proliferin	29	9	6	0	60	7	6	1	50	55	A	N	A	A
	Human somatomammotropin	26	9	3	1	70	9	3	1	68	69	A	A	A	N

7.4	Zein E19	39	15	5	1	70	11	7	0	61	67	A	N	A	N
	Zein E25	39	15	5	1	70	8	10	0	44	58	N	N	A	N
	Gliadin α/β	20	12	1	1	86	2	4	0	33	70	N	A	A	C
	Gliadin α	20	12	1	1	86	2	4	0	33	70	N	A	A	C
	Phaseolin	26	12	0	2	86	7	5	0	56	73	N	N	A	N
	Lectin	24	10	4	2	63	3	3	2	33	54	C	A	A	C
7.5	Pea *rbc*	57	17	10	2	59	12	10	6	44	50	C	N	A	N
	Lemna	52	15	6	2	70	13	11	5	45	50	A	N	A	N
	Wheat *W9*	47	13	5	0	72	12	13	4	43	53	C	N	A	N
	Wheat *WS4.3*	46	13	5	0	72	13	9	6	48	56	C	N	A	N
7.6	Yeast$_N$ cytochrome c_1	61	13	13	6	41	17	10	2	60	49	N	A	A	N
	Pea$_C$ cytochrome *f*	57	13	8	8	45	14	9	5	50	47	N	A	N	A
	Spinach$_C$ cytochrome *f*	59	14	12	5	45	12	12	4	43	44	N	A	N	A
	Wheat$_C$ cytochrome *f*	60	12	13	7	38	11	13	4	40	39	N	A	N	A
	Yeast *c* peroxidase	67	20	7	4	69	15	18	3	42	55	A	N	N	N
	Bovine cytochrome P-450$_{scc}$	39	16	4	3	69	9	4	3	56	63	A	A	A	N
	Yeast cytochrome *c* oxidase IV	25	9	3	4	56	3	5	1	33	48	A	A	N	N
	Bovine cytochrome *c* oxidase IV	22	9	3	3	60	2	4	1	29	49	A	C	A	N
	Yeast cytochrome *c* oxidase VI	41	11	6	4	52	4	11	5	20	36	C	N	N	C

(*continued*)

Table 7.11 (Continued)

Table	Gene	Number of codons	5'-Half				3'-Half				Total % Apolar	Cleavage site			
												Transit peptide		Mature sequence	
			A	N	C	%A	A	N	C	%A		Pen	Ult	Site 1	Site 2
7.7	*Schizophyllum commune* 1G2	40	19	2	1	86	7	10	1	38	60	N	N	N	C
	Fusarium solani C3	31	10	4	1	67	8	7	1	50	58	N	A	A	C
	Bombyx mori A	21	8	2	0	80	6	5	0	55	66	A	A	N	A
	Bombyx mori B	21	8	1	1	80	9	3	0	75	80	N	A	N	N
7.8	Yeast PHO3	17	5	3	1	56	7	1	0	88	70	N	A	A	N
	Yeast PHO5	17	5	3	1	56	6	2	0	75	65	N	A	A	N
	Yeast SUC2	19	10	1	0	90	6	1	1	75	84	N	A	N	A
7.9	*V.c.* toxin A1	18	7	0	1	87	6	4	0	60	72	N	A	A	C
	V.c. toxin B	21	7	2	2	64	6	4	0	60	62	N	A	A	A
	E.c. toxin LTA	18	6	2	1	67	7	2	0	78	72	N	A	N	A
	E.c. toxin LTB	21	5	4	2	47	6	4	0	60	53	N	A	N	A
	P.a. toxin A	25	9	4	0	86	9	3	0	75	72	N	A	N	C
	C.d. toxin 228	25	8	2	2	67	11	2	0	84	76	N	A	A	A
	B.p. xylanase	27	9	2	3	60	10	3	0	46	70	N	A	C	N
	E.c. lamB	25	12	1	2	80	8	2	0	80	80	A	A	A	C

B.s. **amylase**	29	10	2	3	67	15	1	0	94	86	A	A	N	A
B.l. **amylase**	29	9	3	4	56	11	2	0	84	68	A	A	A	N
E.a. ompA	21	7	1	2	70	9	2	0	82	76	N	A	A	A
S.t. ompA	21	7	1	2	70	9	2	0	82	76	N	A	A	A
E.c. ompF	22	9	1	2	75	8	2	0	80	78	N	A	A	C
E.c. ompC	21	8	1	2	73	9	1	0	90	80	N	A	A	C
E.c. ompT	17	8	1	0	89	4	2	2	50	70	A	C	A	N
E.c. traA	50	16	7	4	58	20	1	2	87	72	A	A	A	A
B.n. **pilin**	7	1	1	1	33	2	1	1	50	43	C	A	A	N
E.c. fimA	23	8	2	2	67	7	4	0	64	65	A	A	A	A
E.c. fimpA	21	9	1	2	75	6	3	0	67	72	N	A	A	A
S.a. **protein A**	36	12	4	5	58	11	4	0	74	64	N	A	A	N
Total 61		541	220	129		502	303	62			20A, 6C, 35N	31A, 2C, 28N	44A, 1C, 16N	23A, 13C, 25N
Mean:	30%				66%				59%	63%				

[a]Abbreviations: A, apolar (hydrophobic); C, polar, charged (hydrophilic); N, neutral (polar, uncharged); Pen, penultimate; Ult, ultimate.
[b]For additional abbreviations see the respective tables.

hydrophobicity of the 5′ portion of the total of 61 sequences is 66%, whereas that of the 3′ part is 59%. As a whole, codons for charged amino acids are sparse and usually characterize the 5′ half, where two-thirds of their total occur; there are some cases, nevertheless, where they abound in the downstream portion, the ribulose-1,5-bisphosphate carboxylase genes of plants (Table 7.5) being particularly outstanding examples. Triplets encoding uncharged polar amino acids are more frequent in the 3′ sector, being about 50% more abundant there than in the upstream part. In total hydrophobicity, the range is from the 50% level of the ribulose-1,5-bisphosphate carboxylases to 80 and 86% in two bacterial enzymes, for an overall mean rate of 63%.

In connection with the degree of hydrophobicity, one factor that has previously passed unnoticed in the literature needs to be brought out, an observation based on the number of codons for the several categories of amino acids. The apolar group of amino acids is encoded by a total of 28 codons, or 46% of the 61 that signal monomers, that is, exclusive of the three for stop combinations. Half as many, 14 (23%), code for charged polar types, and only slightly more, 19 or 31%, indicate uncharged polar varieties. Hence, a structure of 50% or less codons for apolar units is at the random-chance level and cannot validly be considered especially hydrophobic. Thus the 59% mean of 3′ halves of Table 7.11 represents only a mild level of hydrophobicity; only the 65% average of the 5′ halves appears to offer much conviction.

At the cleavage site, codons for hydrophobic protein monomers also are the most abundant, followed closely by those of uncharged ones, but the distribution is decidedly uneven. In the penultimate position of the presequence, those for neutral amino acids are prevalent, with a total of 35, against 20 for apolar and six for charged. But at the closing site the situation is reversed, with those encoding hydrophobic amino acids leading by a slim majority. The near absence here of triplets for charged amino acids is particularly surprising, as is their virtual nonexistence at the first location of the mature coding sequence. At the latter site codons for hydrophobic monomers are again the rule, in the ratio of 44 to 16 for neutral ones. Only at the second site of the mature sequence are codons for charged amino acids frequent, but not nearly as abundantly as those of the neutral and apolar types, which are subequal in numbers.

Importance of the Transit Peptide. Obviously the transit peptides encoded by presequences of numerous simple diplomorphic genes vary widely, as is made clear by Table 7.11. Aside from the very evident hydrophobicity that usually prevails, few common traits are in view—even the extent of hydrophobic genes is under 50% in some cases. Nevertheless, the transit peptides have been demonstrated to be requisite as a whole for the transport of proteins through various membranes (Austen *et al.*, 1984; Hannink and Donoghue, 1984; Hurt *et al.*, 1984; Horwich *et al.*, 1985a; Takahara *et al.*, 1985). But what specifically qualifies a peptide sequence to serve in this capacity remains confused at the moment. As was seen, much diversity exists among presequences, as a consequence of which experimental results are also highly varied. One series of analyses on mutations within the transit-peptide region indicated that the length of its hydrophobic sector was the major determinant of its functionality (Bankaitis *et al.*, 1984). In some cases a specific essential sequence has been identified. For instance, the β-galactosidase of *E. coli* could be targeted to the nucleus in yeast cells, provided a segment as small as 13 amino acids of the presequence was present (M. N. Hall *et al.*, 1984). An important part of this appeared to be the series lysine, isoleucine, proline, isoleucine, lysine, that is, three apolar types

between two charged units. But as the tables show, few presequences possess such a series. Moreover, arginine has been stated to be essential (Horwich *et al.*, 1985b). Still another proposed requisite feature was the secondary structure, as in the appendage of the E. *coli lamB* gene (Table 7.9; Emr and Silhavy, 1983). In this investigation the results indicated that genetic shortening of the helical region between remote codons for proline and glycine induced a coiled configuration in the encoded product that prevented transit through the cell membrane. But again this arrangement of codons for amino acids that interfere with α-helix construction is far from being a universal feature.

Absence of a Presequence in Exported Gene Products. Although a transit peptide characterizes the majority of products that must pass through a membrane before being fully processed, that condition, too, lacks catholicity. Indeed, a recently identified peptide involved in the processing of the transit portion of lipoproteins in *E. coli* has proven to be devoid of this trait, in spite of its being exported (Innis *et al.*, 1984). Such *E. coli* outer membrane proteins as the products of the *ompR* and *envZ* genes and outer membrane phospholipase A, which obviously must pass through that of the cell, lack identifiable presequences (Mizuno *et al.*, 1982; Wurtzel *et al.*, 1982; de Geus *et al.*, 1984). Among the most evident members of this category are the cytochrome *c* genes, which, as seen earlier, are located in the nucleus and translated in the cytoplasm, but their products function only in the mitochondrion. Yet they universally have no presequence (Limbach and Wu, 1985a,b; Scarpulla, 1985). This absence, however, is readily understood in the present example, because cytochrome *c* does not actually penetrate the mitochondrion, remaining outside in close association with the outer membrane.

Additionally, a number of proteins that remain embedded in the cytoplasmic membrane do not possess a presequence. One such has recently come to light in rat liver, a peptide that serves as the receptor for asialoglycoprotein (Holland *et al.*, 1984). But a number of similar structures are well known, including genes for *E. coli* lactose permease and members of the *Salmonella* histidine- and *E. coli* maltose-transport systems (Ehring *et al.*, 1980; Higgins *et al.*, 1982; Froshauer and Beckwith, 1984). Several genes for viral products that become located in prokaryotic or eukaryotic membranes but lack a presequence also have been sequenced, among which are the gene III of bacteriophage f1 that infests *E. coli* and the E1 glycoprotein of a coronavirus that reproduces within the endoreticular membranes of the laboratory mouse (Boeke and Model, 1982; Armstrong *et al.*, 1984). To the contrary, the fusion protein gene of the human respiratory syncytial virus has a presequence of 26 codons, despite the fact that mature product is embedded in the cell membrane (Elango *et al.*, 1985). Consequently, it may be deduced that proteins penetrating into and becoming fixed within a membrane require factors different from those that pass through a membrane.

Multiplicity of Factors. Since the proteins that travel through membranes are so diverse in structure of the transit presequence, or, in the latter's absence, in the structure of the mature protein, it appears unrealistic to suppose that only one or a few membrane components carry out the function of recognizing and transporting proteins of all descriptions. Rather, it seems far more logical to propose that a large number of such transit-peptide-recognizing substances are present that react with one or two families of products and no other (Rapoport, 1985). Indeed, one such protein has recently been purified from endoreticulum of canine pancreas and visualized by electron microscopy. This component, which is comprised of six polypeptide subunits plus one molecule of 7 S RNA,

proved to be a narrow cylinder 24 nm long and 6 nm wide (Andrews *et al.*, 1985). However, this component recognizes only the presequences of messengers that are to be processed within the endoreticulum and functions in establishing the ribosomal connection with that organelle's membrane to initiate translation. Hence, presequences and transit peptides destined for processing or functioning in other organelles would be recognized by other proteins. As a consequence of this diversity in the transit agents, no great uniformity in structure, either of the transit presequence or transported mature product, would be expected; rather, a diversity, such as that which exists, would be predicted.

The comparable multiple-factor condition that is coming to light for the peptidases that remove the transit peptide correlates well with the foregoing proposals. In *E. coli* one such peptidase that acts upon the cleavage site of precoat proteins has long been recognized (Zwizinski and Wickner, 1980; Wolfe *et al.*, 1983). As pointed out before, a second one, active only on lipoproteins, has now been characterized (Innes *et al.*, 1984). Just as the visualized multiplicity of presequence and transit-peptide-recognizing proteins may explain the structural diversity of the latter, this possible many-factored condition now being revealed affords an explanation of the inconstancy that prevails at the cleavage site.

Evolutionary Implications. If the hypothesized abundance of protein types becomes more firmly established, as seems to be the strong likelihood, then current explanations of distribution of mitochondrial and chloroplastic genes between the nuclear and organellar genomes need to be rethought. Obviously, removal of a gene from the organellar DNA for insertion into the former requires far more than is evident from the superficial data on which that type of proposal is based. The need for a preexisting presequence has already been pointed out (Chapter 6, Section 6.1.3), but now it appears that each such transplanting requires the presence in the proper location of two additional substances, a particular transit-peptide-recognition protein and a specific transit peptidase. Both of the latter as well as the presequence would have to be present immediately following the translocation of the given gene, otherwise the product would be unable to reach its destination and become functional. Translocating a gene from an organelle to the nucleus may be readily performed, but making it able to carry out its usual activity in the needed site requires the simultaneous acquisition of three additional genes, if the predicted multiple-factor theory proves to be correct.

7.4. PRE- AND PROSEQUENCES OF MORE COMPLEX DIPLOMORPHS

A large number of genes, especially those for secreted products, not only encode a transit peptide in a presequence, but also have another expendable sector known as a prosequence, therefore being rather more complex than the preceding types (Figure 7.1). Like the former, this latter section is removed proteolytically, its removal bringing about the activation of the mature product. Thus, so long as the two portions are attached, it suppresses the activity of the principal protein and accordingly is here named the inhibitor peptide. In examining the structure of these two classes of appended gene sequences, the procedures parallel those of the foregoing sections, with vertebrate genes being viewed first.

7.4.1. Transit and Inhibitor Sequences of Vertebrate Genes

An unexpectedly large proportion of genes for proteins in vertebrates bear pre- and prosequences at the 5′ end. Since the presence of an inhibitor peptide provides the organism with a means of controlling the activity of the substances, its occurrence on genes for such digestive enzymes as pepsinogen and chymotrypsinogen comes as no surprise. What is unexpected is that inhibitors also are found on such chemically mild substances as albumins, as is seen in the first set of examples that follows.

The Albumin Family of Genes. Included within a common protein family are four diverse major categories, albumins, α-fetoproteins (really embryonic albumins), parathyroid hormone (PTH), and lysozymes. Although the gene sequences of representatives from each of these groups have been established, the precise limits of the prosequence are not always provided. Consequently, the genes for such proteins as rat α-lactalbumin and chicken lysozyme (Jung *et al.*, 1980; Qasba and Safaya, 1984) and the α-fetoproteins of mouse and human (Gorin *et al.*, 1981; Law and Dugaiczyk, 1982; Morinaga *et al.*, 1983; Sakai *et al.*, 1985) could not be included in Table 7.12. However, the five contained therein serve well in introducing this topic, for they fall into two contrasting sets, which nevertheless share some general trends.

The structures of the three PTH genes of mammals are especially helpful because they are largely homologous. As a whole, their presequences display the same major features as did those of simpler genes, namely a high content of codons for hydrophobic amino acids and the presence of one or two triplets encoding basic monomers near the 5′ end. The two codons for methionine at the extreme 5′ terminus in the bovine and rat genes (Heinrich *et al.*, 1984a; Weaver *et al.*, 1984), but not in the human (Vasicek *et al.*, 1983), while not of rare occurrence, is certainly not commonplace. This same codon is highly conserved at three other places in the presequence, where it may hold functional importance. In each member of the trio, the 3′ half is distinctive in having codons for acidic amino acids in the penultimate site, often preceded by one for a basic monomer two sites upstream. Their prosequences are uniformly short, consisting of only 18 nucleotide residues, and are remarkable for their basicity, each containing four triplets encoding lysine or arginine, all of which are located at the terminal portions.

The other two sequences are from human sources, the first encoding blood serum albumin (Mita *et al.*, 1984) and the second encoding that protein from liver (Morinaga *et al.*, 1983). In the 5′ halves, several relationships to the PTH presequences can be perceived, although homologous sites are far from abundant. Only single codons for basic amino acids lie near the upstream termini, differently situated in each case. Distinctions between the pair of presequences exist also in the 3′ halves, for whereas the liver albumin gene resembles the termini of the parathyroid hormones, the serum transit sequence does not. However, in the prosequence, that situation is reversed, with the serum protein gene more closely approaching the hormonal. In neither case is the extreme basicity of the latter approached, the sector from the serum cistron having a 50% basic ratio and that of the liver only 12½%. Thus, here it would seem, as in those of the first sections, that variability around the cleavage sites is their only constant feature.

The Trypsinogen Family of Vertebrate Genes. The present family of proteases whose active sites contain a serine residue includes trypsinogen, chymotrypsinogen,

Table 7.12
Transit and Inhibitor Peptide Genes of the Albumin Family[a]

Row 1

Parathyroid hormone	Presequence
Bovine PTH[b]	*ATG ATG* TCT *GCA* AAA GAC *ATG GTT* AAG *GTA ATG ATT GTC ATG CTT GCC ATC* TGT *TTT*
Rat PTH[c]	*ATG ATG* TCT *GCA* AGC ACC *ATG GCT* AAG *GTG ATG ATC CTC ATG CTG GCA GTT* TGT *CTC*
Human PTH[d]	*ATG ATA CCT GCA* AAA GAC *ATG GCT* AAA *GTT ATG ATT GTC ATG TTG GCA ATT* TGT *TTT*
Albumins	
Human serum A[e]	*ATG* AAG TGG *GTA* --- ACC --- --- --- *TTT* --- *ATT* --- TCC *CTT* --- --- *CTT TTT*
Human liver A[f]	*ATG* GCT TCT CAT CGT *CTG* --- --- --- *CTC* --- *CTC CTC* TGC *CTT* --- --- *GCT GGA*

Row 2

Parathyroid hormone	↓	Prosequence	↓	Mature gene
Bovine PTH	*CTT GCA* AGA TCA --- GAT *GGG*	AAG TCT --- *GTT* --- AA*G AAG AGA		*GCT GTG*
Rat PTH	*CTT* ACC CAG *GCA* --- GAT *GGG*	AAA *CCC* --- *GTT* --- AA G AAG AGA		*GCT GTC*
Human PTH	*CTT* ACA AAA TCG --- GAT *GGG*	AAA TCT --- *GTT* --- AA*G AAG AGA		TCT *GTG*
Albumins				
Human serum A	CTC *TTT* AGC TCG *GCT* TAT TCC	AGG *GGT* --- *GTG TTT* -- - CGT CGA		GAT *GCA*
Human liver A	*CTG GTA TTT GTG* TCT GAG *GCT*	*GGC CCT* ACG *GGC* ACC *GG T* GAA TCC		AAG TGT

[a]Codons for hydrophobic amino acids are italicized and those for hydrophilic ones are underscored. PTH, parathyroid hormone. Asterisks indicate location of an intron; arrows mark cleavage sites.

[b]Weaver *et al.* (1984). [c]Heinrich *et al.* (1984a). [d]Vasicek *et al.* (1983). [e]Mita *et al.* (1984). [f]Morinaga *et al.* (1983).

elastases, and kallikreins. Typically each is represented in the genome by multiple isozymic species, with kallikreins being especially highly diversified. The members of that category form a distinct subfamily that processes the precursors of polypeptide hormones, each having a limited substrate specificity. One representative is a subunit of the enzyme that processes nerve growth factor, another activates epidermal growth factor, and a third removes the repressor peptide from angiotensinogen to produce angiotensin II. In the mouse the 25–30 kallikrein genes have been located on chromosome 7 (Mason *et al.*, 1983). At this early stage in the sequencing of genes, only a few members of the family have had their structures determined; however, some of these, like the trypsinogen gene and two elastase cistrons, have been sequenced, all from the rat (Craik *et al.*, 1984; Swift *et al.*, 1984), but the precise limits of the prosequences have not been established. Consequently, just three representatives of the group are included in Table 7.13, cistrons for canine chymotrypsinogen (Pinsky *et al.*, 1983) and rat and mouse kallikreins (Swift *et al.*, 1982; Mason *et al.*, 1983).

The three examples are remarkably similar in structure, including chymotrypsinogen, despite the latter's strongly contrasting activity. Both the pre- and prosequences are rather constant in length, the most deviant of the former being that of the second species of kallikrein. Probably the greatest distinctive feature is the presence of an intron near the 5′ end of row 2 in the mouse representative, which is not indicated in the others. Undoubtedly this is the result of faulty knowledge, rather than a basic difference, since the other two sequences have been derived from mRNAs, the third alone being from the genomic DNA. Two cleavage sites for the presequences are given in the table, the more upstream one being derived from the canine and rat structures, the second from that of the mouse. Since the chemical nature of the encoded amino acids abutting the sites is similar in each case, it is not possible to determine the one more likely to prove valid. The 3′ end of the prosequence, however, is firmly established, a codon for an arginine base being uniformly at the cleavage point, with a triplet for serine preceding it in all cases. At the 5′ end of the mature coding sequence, two codons for apolar amino acids border the site.

Vertebrate Lipoproteins. The plasmolipoproteins of vertebrates are grouped into four main classes, chylomicrons, very low density, low density, and high density. At least nine species have been identified, A-I, A-II, A-IV, B, C-I to C-III, D, and E, of which the first two make up most of the high-density fraction (Shoulders *et al.*, 1983). This pair also represents the only two that are known to bear both pre- and prosequences, the remainder having only the former. Consequently, these are probably the sole representatives of the family that are potent enzymes. The A-I species is involved as a cofactor of the lecithin-cholesterol acyltransferase activity (Karathanasis *et al.*, 1983), a reaction in which A-II may also participate (Moore *et al.*, 1984).

The representatives of those two species included in Table 7.13 are from human sources, only the first of which reflects the structure of the DNA. Hence, the seeming absence of an intron in A-II results from all three of the established sequences of this species having been derived from mRNA (Knott *et al.*, 1984b; Lackner *et al.*, 1984; Moore *et al.*, 1984). Although no relationship is implied between the present proteins and the proteases also shown in the table, an unexpectedly high number of sites do display homologies in the presequence portion—even the cleavage site of that appendage shows some similarities. However, there is a codon for a basic amino acid at the immediate 5′ end that has no counterpart in the others. The prosequences are completely different from

Table 7.13
Pre- and Prosequences of the Vertebrate Trypsinogen and Liproprotein Families[a]

Row 1

Proteases	Prosequence	↓
Canine chymotrypsinogen[b]	*ATG GCT TTC* --- *CTC* TGG *CTC CTC* TCC TGC *TTC GCC CTC CTG GGC* ACA *GCC* --- *TTC GGC*	TGC *GGG*
Rat kallikrein[c]	*ATG CCT GTT* ACC *ATG* TGG *TTC CTG ATC CTG TTC CTC GCC CTG* --- --- TCC --- *CTG GGA*	CGG *AAT*
Mouse kallikrein[d]	*ATG* --- --- --- --- TGG *TTC CTG ATC CTG TTC CTA GCC CTG* --- --- TCC --- *CTA GGA*	*GGG ATT*
Lipoproteins		
Human A-I[e]	*ATG* AAA *GCT* GCG --- --- *GTG CTG* ACC *TTG GCC GTG CTC TTC CTG* ACG*GGG* AGC CAG *GCT*	CGG CAT
Human A-II[f]	*ATG* AAG *CTG* --- *CTC* --- *GCA GCA* ACT *GTG CTA CTC CTC* ACC *ATC* TGC AGC *CTT* GAA *GGA*	*GCT TTG*

Row 2

Proteases	↓	Presequence	↓	
Canine chymotrypsinogen	*G TC CCT GCC ATC*	CAG *CCG GTG TTA* AGT *GGC CTG* TCC AGG		*ATC GTC*
Rat kallikrein	G AT --- *GCT GCA*	*CCT CCC GTC* --- --- --- CAG TCT CGG		*GTT GTT*
Mouse kallikrein	G*AT --- *GCT GCA*	*CCT CCT GTC* --- --- --- CAG TCT CGA		*ATA GTT*
Lipoproteins				
Human A-I	--- --- --- *TTC*	TGG --- --- --- --- --- --- CAG CAA		GAT GAA
Human A-II	--- --- --- *GTT*	--- --- --- --- --- --- --- CGG AGA		CAG *GCA*

[a]Codons for hydrophobic amino acids are italicized and those for hydrophilic ones are underscored. Asterisks indicate the location of an intron. Arrows mark cleavage sites; alt, alternative cleavage site.
[b]Pinsky *et al.* (1983). [c]Swift *et al.* (1982). [d]Mason *et al.* (1983).
[e]Shoulders *et al.* (1983); Law and Brewer (1984).
[f]Knott *et al.* (1984b); Lackner *et al.* (1984); Moore *et al.* (1984).

those of the trypsinogen family. In the first place, here they are extremely short, at most consisting of only five or six codons, whereas in the others they range from 11 to 15 triplets. Second, the chemical nature of the encoded amino acids at the downstream cleavage site differs strongly, involving the presence of at least one pair of charged monomers.

The Renin Family of Vertebrate Genes. In order to provide a more accurate picture of the vertebrate pre- and prosequences, one additional important family of their genes needs to be analyzed, that including the cistrons for renin and pepsinogen. It is at once apparent that the presequences of the pair given in Table 7.14 have no outstanding features. As in the preceding group, the second site is occupied by a codon for a charged amino acid, in one case for an acid species, in the other for a basic species (Sogawa *et al.*, 1983; Hobart *et al.*, 1984; Miyazaki *et al.*, 1984). At least five codons for leucine (CTN) are present, often in tandem fashion, and the final site at the cleavage point is occupied by a triplet for cysteine. The prosequences are strikingly longer than any that have been viewed to this point, consisting of 46–50 codons. Moreover, an unusually high percentage of these encode alkaline amino acids, ranging from 20 to 25%. These two sectors are definitely not hydrophobic, since less than half of the constituents encode amino acids of that character.

7.4.2. More Complex Diplomorphic Genes of Bacteria

Although an occasional gene for a product referred to as a preproprotein has been sequenced from seed-plant sources (e.g., Lycett *et al.*, 1983), none apparently has actually been established, nor do any appear to be known from yeast. Even the bacteria do not provide a plethora of the type sought for present purposes, although often in bacteriology the presequence is incorrectly referred to as a prosequence; all those available are from bacilli, not the usual *E. coli*. These three encode endoproteases, two of which are for the acid type known as subtilisin, the other for neutral protease. Both enzymes are largely secreted into the environment, only 5% being retained within the organism. The gene for the first of these is activated, along with those of two minor proteases, only when the bacteria are about to sporulate, but the role of the enzyme in spore formation remains unclear (Wells *et al.*, 1983; Wong *et al.*, 1984).

The Presequences. Although it is to be expected that the presequences for the same gene from two species of a given genus, such as those for subtilisin from *Bacillus subtilis* and *B. amyloliquefaciens*, should be largely homologous, what is surprising is that the structure from a different cistron, that for neutral protease, should also show a number of similarities to that pair (Table 7.15). To begin with, all have GTG as the initiation codon, translated in each case as methionine, not valine as elsewhere. Just downstream from this, all three sequences have two contiguous triplets for lysine. Then for a space no kinship between the two types is found, until beyond the middle of the first row, where a broken series of five correspondences exists, beginning with TT- triplets. The extreme 3′ terminus also is constant in each case, the penultimate codon being CAG followed by GC- for alanine. The presequences are moderately short, consisting of 27–29 triplets, about two-thirds of which encode hydrophobic species of amino acids.

The Prosequences. The prosequences encoding the inhibitor peptide greatly exceed in length any that have been reported in preceding pages (Table 7.15). Of the three,

Table 7.14
Pre- and Prosequences of the Renin Family of Genes[a]

Row 1	Presequence	↓ Prosequence
Human renin[b]	*ATG* GAT *GGA* TGG AGA AGG *ATG CCT* CGC TGG *GGA CTG CTG CTG CTG CTC* TGG *GGC* TGG TGT	ACC *TTT* ---
Human pepsinogen[c]	*ATG* AAG --- TGG --- --- *CTG CTG CTG CTG GGT CTG GTG GCG CTC* TCT --- --- GAG TGC	*ATC ATG* TAC
Row 2		
Human renin	*GG T* --- *CTC CCG* ACA GAC ACC ACC ACC *TTT* AAA CG*G *ATC TTC CTC* AAG AGA *ATG CCC* TCA *ATC* CGA GAA AGC	
Human pepsinogin	AA*G *GTC CCC CTC ATC* AGA AAG AAG TCC *TTG* AGG CG C ACC *CTG* TCC GAG CGT *GGC CTG CTG* --- AAG GAC *TTC*	
Row 3		↓ Mature gene
Human renin	*CTG* AAG GAA CGA *GGT GTG* GAC *ATG GCC* --- AGG *CTT GGT CCC* GAG TGG AGC CAA *CCC ATG* AAG AGG	*CTG* ACA
Human pepsinogin	*CTG* AAG AAG CAC AAC *CTC* AAC *CCA GCC* AGA AAG TAC *TTC CCC* CAG TGG GAG *GCT CCC* --- ACC *CTG*	*GTA* GAT

[a]Codons for hydrophobic amino acids are italicized and those for hydrophilic ones are underscored. Asterisks indicate the location of an intron. Arrows mark the cleavage sites.
[b]Hobart *et al.* (1984); Miyazaki *et al.* (1984).
[c]Sogawa *et al.* (1983).

Table 7.15
Pre- and Prosequences from Bacterial Genes for Secreted Proteins[a]

Row 1	Presequence
B.s. subtilisin[b]	GTG AGA --- AGC AAA AAA TTG TGG ATC AGC TTG TTG TTT GCG TTA ACG TTA ATC TTT ACG ATG GCA TTC AGC
B.a. subtilisin[c]	GTG AGA --- GGC AAA AAA GTA TGG ATC AGT TTG CTG TTT GCT TTA GCG TTA ATC TTT ACG ATG GCG TTC GGC
B.a. neutral protease[c]	GTG GGT TTA GGT AAG AAA TTG TCT GTT GCT GTC GCC GCT TCC TTT ATG --- AGT TTA ACC ATC --- --- AGT
Row 2	Prosequence ↓
B.s. subtilisin	AAC ATG TCT GCG --- CAG GCT GCC GGA AAA --- --- AGC AGT ACA --- GAA AAG AAA TAC --- ATT GTC GGA
B.a. subtilisin	AGC ACA TCC TCT GCC CAG GCG GCA GGG AAA --- --- TCA AAC GGG --- GAA AAG AAA TAT --- ATT GTC GGG
B.a. neutral protease	CTG CCG GGT GTT --- CAG GCC GCT GAG AAT +351 GCG CTG GAT CAT GCT TAT AAA GCG ATC GGC AAA TCA CCT
Row 3	
B.s. subtilisin	TTT --- AAA CAG ACA --- ATG AGT GCC ATG AGT TCC GCC AAG AAA AAG GAT GTT ATT TCT GAA AAA GGC GGA
B.a. subtilisin	TTT --- AAA CAG ACA --- ATG AGC ACG ATG AGC GCC GCT AAG AAG AAA GAT GTC ATT TCT GAA AAA GGC GGG
B.a. neutral protease	GAA GCC GTT TCT AAC GGA ACC GTT GCA AAC AAA AAC AAA GCC GAG CTG AAA GCA GCA GCC ACA AAA GAC GGC
Row 4	
B.s. subtilisin	AAG GTT CAA AAG CAA TTT AAG TAT GTT AAC GCG GCC GCA GCA ACA TTG GAT GAA AAA GCT GTA AAA GAA TTG
B.a. subtilisin	AAA GTG CAA AAG CAA TTC AAA TAT GTA GAC GCA GCT TCA GCT ACA TTA AAC GAA AAA GCT GTA AAA GAA TTG
B.a. neutral protease	AAA --- --- TAC CGC CTC GCC TAT GAT GTA ACC ATC CGC TAC ATC GAA CCG GAA CCT GCA AAC TGG GAA GTA
Row 5	↓
B.s. subtilisin	AAA AAA GAT CCG AGC GTT GCA TAT GTG GAA GAA GAT CAT ATT GCA CAT GAA TAT GCG CAA
B.a. subtilisin	AAA AAA GAC CCG AGC GTC GCT TAC GTT GAA GAA GAT CAC GTA GCA CAT GCG TAC GCG CAG
B.a. neutral protease	ACC GTT GAT GCG GAA ACA GGA AAA ATC CTG AAA AAG CAA AAC AAA GTG GAG CAT GCC GCC

[a]Codons for hydrophilic amino acids are underscored and those for hydrophobic ones are italicized. Arrows mark the cleavage sites.
[b]*Bacillus subtilis;* Wong *et al.* (1984).
[c]*B. amyloliquefaciens;* Vasantha *et al.* (1984).

that encoding subtilisin in *B. subtilis,* with 73 condons, is the shortest, its parallel from *B. amyloliquefaciens* being just one codon longer. In the cistron for neutral protease, the prosequence is 194 condons in length (Vasantha *et al.,* 1984), but 117 (351 nucleotides) of these are excluded from the table as meaningless in the absence of comparative material. Insofar as their general chemical natures are concerned, all are similar in having 31 codons for apolar amino acids in the regions given. The two subtilisin genes encode 15 lysines, but no arginines and either nine or ten acidic protein monomers. The neutral protease has triplets for 11 lysines and two arginines and ten for acidic components. At the 5′ terminus the latter has a triplet for an acid amino acid, whereas the other two have codons for lysine. What is particularly striking, however, is that these codons very frequently lie at coinciding points in the sequence, especially beginning with the end of the third row, where three AAAs for lysine comprise a column. In row 4 this is followed almost immediately by another column of lysine codons, then by one for tyrosin (TAT), and toward the end by two sets of GAA for glutamic acid. All in all, nevertheless, this prosequence of the gene for neutral protease cannot be homologized with those of the two subtilisin cistrons.

7.5. GENES FOR CRYPTOMORPHIC PROTEINS

The several preceding sections encompassed genes for many contrasting types of proteins, which shared common distinctive structural features. All had a presequence whose product after translation was removed to release either a mature protein in simpler cases, or, in more complex instances, those that also bore an inhibitor peptide, whose removal then gave rise to the functional product. Although each pre- and prosequence was seen to be unique, they shared a number of common properties, at least functionally. In the present section, however, much more complex gene structures are considered, whose organizational traits are highly divergent from one representative subclass to another. Presequences and inhibitor sections are often present, but one or the other, or even both, may occasionally be missing. Those features thus are of minor importance in the present class of genes. What is of significance here is the posttranslational fragmentation of the product encoded by the mature coding region into one or more definitive peptides; those actual functional parts encoded by the smaller fractions of the mature region are referred to here as the "ultimate" genes. As a rule the sequences between their encoded peptides are called connectors. The posttranslational aspect of cleavage of the major sectors is an important consideration, for the members of one subclass grade into multigenic precursorial transcripts. In the latter, however, the several parts are cleaved while in the form of RNA and by an RNase, not after translation and by a protease. Because the sequence encoding the actual active substances is concealed within a far longer mature region, the genes of this type are said to be "cryptomorphic," as pointed out in Chapter 1, Section 1.1.3. Insofar as is known, the cryptomorphic class is confined to eukaryotes, reflecting the complicated requirements of more complexly organized cells.

7.5.1. Cryptomorphic Genes Encoding Multiple Identical Proteins

Probably the simplest group of cryptomorphic genes (designated here as subclass I) is that which embraces those encoding multiple copies of a single peptide; consequently,

Table 7.16
Presequences of Simple Cryptomorphic Genes (Subclass I)[a]

Row 1	
Yeast *MFα1*[b]	*ATG* AGA *TTT CCT* TCA *ATT TTT* ACT *GCA GTT TTA TTC GCA GCA*
Yeast *MFα2*[b]	*ATG* AAA *TTC ATT* TCT ACC *TTT CTC* ACT *TTT ATT TTA GCG GCC*
Rat enkephalin[c]	*ATG GCG* CAG *TTC CTG* AGA *CTT* TGC *ATC* TGG *CTG CTA GCG CTT*
Human enkephalin[d]	*ATG GCG* CGG *TTC CTG* ACA *CTT* TGC ACT TGG *CTG CTG TTG CTC*
Bovine enkephalin[e]	*ATG GCG* CGG *TTC CTG GGA CTC* TGC ACT TGG *CTG CTG GCG CTC*
Row 2	↓
Yeast *MFα1*	TCC TCC *GCA TTA GCT GCT* --- --- *CCA GTC* AAC ACT
Yeast *MFα2*	*GTT* TCT *GTC* ACT *GCT* --- --- --- *AGT* TCC GAT GAA
Rat enkephalin	*GGG TCC* TGC *CTC CTG GCT* ACA *GTG* CAG *GCA* GAC TGC
Human enkephalin	*GGC CCC GGG CTC CTG GCG* ACC *GTG* CGG *GCC* GAA TGC
Bovine enkephalin	*GGC CCC GGG CTC CTG GCG* ACC *GTC* AGG *GCA* GAA TGC

[a]Codons for hydrophilic amino acids are underscored and those for hydrophobic ones are italicized. Arrow indicates the cleavage site.
[b]Singh *et al.* (1983).
[c]Rosen *et al.* (1984); Yoshikawa *et al.* (1984).
[d]Legon *et al.* (1982).
[e]Gubler *et al.* (1982); Noda *et al.* (1982).

complete processing of the mature coding sector releases several active molecules of identical (or near-identical) structures. But this simplicity is only relative, even in the most primitive eukaryotes, as seen in the discussion that follows immediately.

The α-Pheromones of Yeast. Mating in yeast is coupled to the production of specific pheromones, or mating factors, there being two contrasting types, *a* and α. One population of cells produces factor *a*, and the other α, each of which is capable of arresting cells of opposite type in G_1 of the cell cycle, thereby preventing cell division and asexual multiplication (Siliciano and Tatchell, 1984). Since the gene structure of the α pheromone is by far the better documented, attention is confined entirely to that product and its gene. Recently it has been established that yeast possesses two cistrons for this substance, *MFα1* encoding four copies and *MFα2* only two (Singh *et al.*, 1983). Both are simple cryptomorphs, bearing a presequence encoding a transit peptide that enables the product to be conducted through the cell wall into the medium, where it can act upon cells of opposite type. For ease of comparison, the transit sectors are tabulated separately from the body of the cistrons, which are shown in Table 7.17.

The sequences for the two transit peptides (Table 7.16) are of nearly equal length (21 and 22 codons) and show extensive homology. Each has a codon for lysine (α1) or arginine (α2) directly after the initiation triplet. Although codons for hydrophobic amino acids predominate throughout, there is a continuous series of six near the middle of each presequence and also a run of similar length in α1 at the 3′ terminus.

Table 7.17
Mature Coding Sectors of Mfα Genes of Yeast[a]

Row 1		
MFα1	AAC ACT ACA ACA GAA GAT GAA ACG *GCA* CAA *ATT CCG GCT* GAA *GCT GTC* ATC *GGT* TAC *TTA*	
MFα2	--- --- --- --- GAT GAA GAT *ATC GCT* CAG *GTG CCA GCC* GAG *GCC ATT ATT GGA* TAC *TTG*	
Row 2		
MFα1	GAT *TTA* GAA *GGG* GAT *TTC* GAT *GTT GCT GTT TTG CCA TTT* TCC AAC AGC ACA AAT AAC *GGG*	
MFα2	GAT *TTC GGA GGT* GAT CAT GAC *ATA GCT TTT TTA CCA TTC* AGT AAC *GCT* ACC *GCC* AGT *GGG*	
Row 3		
MFα1	*TTA TTG TTT ATA* AAT ACT ACT *ATT GCC* AGC *ATT GCT GCT* AAA GAA GAA *GGG GTA* TCT *TTG* GAT	
MFα2	*CTA TTG TTT ATC* AAC ACC ACT *ATT GCT* GAG *GCG GCT* GAA AAA GAG CAA AAC ACC ACT *TTG GCG*	
Row 4	Connecting Sequence	
MFα1	AAA AGA GAG --- --- *GCT* GAA *GCT* TGG CAT TGG *TTG* CAA *CTA* AAA *CCT GGC* CAA *CCA ATG* TAC	α1A
MFα2	AAA AGA GAG *GCT GTT GCC* GAC *GCT* TGG CAC TGG *TTA* AAT *TTG* AGA *CCA GGC* CAA *CCA ATG* TAC	α2A
Row 5	Connecting Sequence	
MFα1	AAG AGA GAA *GCC* GAA *GCT* GAA *GCT* TGG CAT TGG *CTG* CAA *CTA* AAG *CCT GGC* CAA *CCA ATG* TAC	α1B
MFα2	AAG AGA GAG *GCC* AAC *GCT* GAT *GCT* TGG CAC TGG *TTG* CAA *CTC* AAG *CCA GGC* CAA *CCA ATG* TAC	α2B

Row 6	Connecting sequence	
MFα1	AAA AGA GAA *GCC* GAC *GCT* GAA *GCT* TGG CAT TGG *CTG* CAA *CTA* AAG *CCT GCC* CAA *CCA ATG* TAC	α1C
MFα2	TGA	
Row 7	Connecting sequence	
MFα1	AAA AGA GAA *GCC* GAC *GCT* GAA *GCT* TGG CAT TGG *TTG* CAG *TTA* AAA *CCC GCC* CAA *CCA ATG* TAC TAA	α1D

[a]Codons for hydrophilic amino acids are underscored and those for hydrophobic ones are italicized. Arrows indicate cleavage sites.
[b]Singh *et al.* (1983).

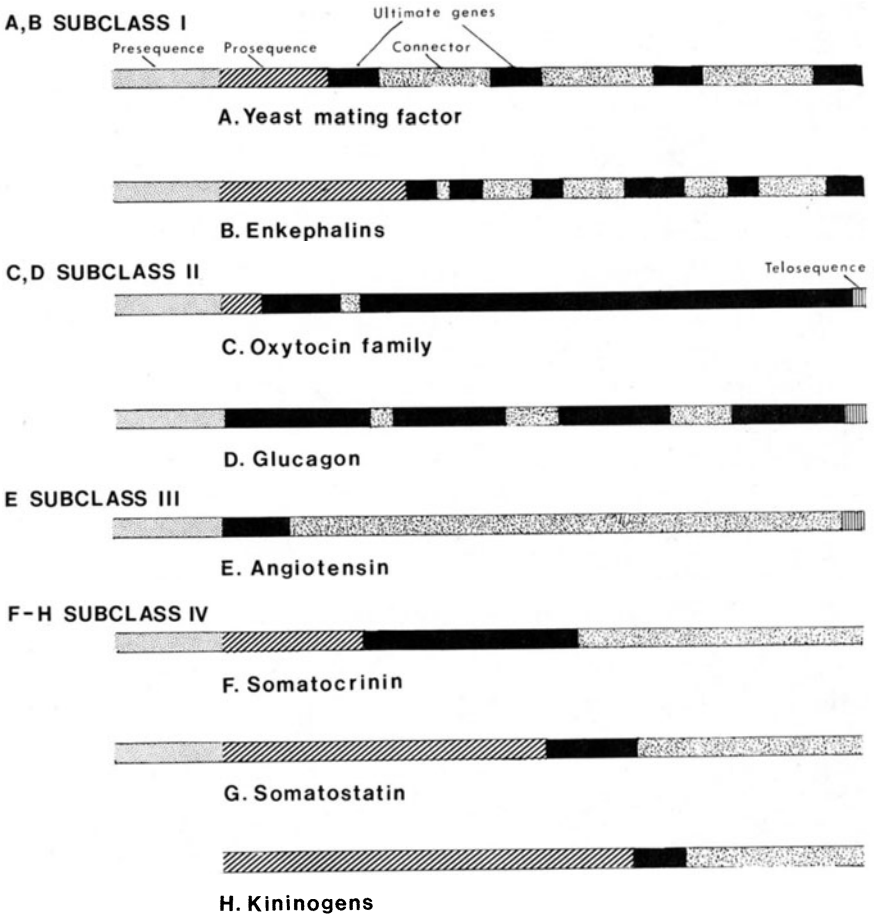


Figure 7.2. Typical structures of four subclasses of cryptomorphic genes; the fifth is shown in Figures 7.3 and 7.4.

Mature Coding Sectors. Aside from the distinctions in gene content just mentioned, the mature coding sectors of the two cistrons differ in several outstanding ways, although numerous homologies exist on a site-to-site basis (Table 7.17). The first of these is the absence of a four-codon sector at the 5′ end of α2, and later, a shorter region of two triplets that is lacking in α1 (Kurjan and Herskowitz, 1982; Singh *et al.*, 1983). But the greatest contrast between the two is in total length, α1 containing 144 codons to the 97 of α2, a condition contributing to the loss of two ultimate coding sequences from the latter. The ultimate genes that encode the actual α factors form a relatively small fraction of the whole, because each consists of only 13 codons, which is a total of just over a third of the mature coding region in α1 and about 25% in α2 (Figure 7.2A). When the mature product is activated, the processing seemingly takes place in two steps, the first proteolysis

occurring 5′ to the series of three charged amino acids at the left of each of rows 4–7. At present no further function is known for the lengthy peptide occupying rows 1–3 in the table that is thus removed. During the second step of processing, the six- to eight-codon-long connecting peptides are removed from the ultimate products, the genes for which are shown stippled in the table.

Distinctions between the ultimate genes of the two cistrons are few, but those that do exist present a challenging basis for speculation as to the origin of each series. The second codons, CAC in α2 and CAT in α1, provide one constant difference, but the fourth triplets, TTA and TTG in α2, TTG and CTG in α1, are problematic, a difficulty accentuated by the triplet that follows. Here one finds CAA in α1A to 1C, with CAG in D, while in α2 the signals are AAT and CAA. The next one, too, is enigmatic, but the genes thereafter are largely identical.

7.5.2. Subclass I Cryptomorphs of Vertebrates

Although only a single group of members of subclass I cryptomorphic genes has yet been sequenced from vertebrates, there can be little doubt that additions will be made to the list as knowledge of the synthesis of secreted control elements advances. Even in these representatives, the small peptides known as enkephalins, understanding of their production is limited, for their presence in various tissues has only been detected within the last decade.

The gene sequences of three enkephalins have been established, one each from the rat, cattle, and human; typically they are derived from the adrenal medullary cells, but the rat sequence is from brain neurons. Two main varieties are found in those tissues, named after the fifth (terminal) amino acid in the pentapeptides, methionine-enkephalin and leucine-enkephalin; other variants result from processing variations, one possibility of which is suggested in a later discussion. Though the actual physiological role of the group remains unsolved, they have been demonstrated to product effects similar to opiate drugs, with which they compete *in vivo* (Hughes *et al.*, 1979).

Although its final assignment to class and subclass depends upon the establishment of the complete structure, the gene encoding vitellogenin in the chicken and most likely also in *Xenopus* probably is a member of the present subclass. This is an immense cistron, embracing over 20,000 base pairs, the whole being divided into 34 exons. Recently the determination of a sector containing exons 23 and 24, which together encode phosvitin, revealed characteristics that can only be interpreted in light of the entire gene being cryptomorphic (Byrne *et al.*, 1984). The short translated sector at the 5′ end of phosvitin shows no indication of representing a transit peptide, since only eight of its 20 codons are for apolar amino acids; rather, it appears more like a connector, as does also the ten-codon sector that follows the final CAG triplet (for glutamine) that marks the 3′ terminus of the mature coding region. Nor are any of the usual transcriptional and translational signals present in either flanking segment. Indeed, termination of transcription of the ovalbumin gene has now been demonstrated to occur ~900 base pairs downstream from the last exon (LeMeur *et al.*, 1984).

Presequences of Mammalian Enkephalins. The three presequences of enkephalin genes (Table 7.16) are homologous to a degree expected only of those encoding

regions of critical importance; consequently, despite their ephemeral existence, their peptide products must be viewed in that light. If only a predominantly hydrophobic constitution is the requirement for passage through a membrane, the question arises as to why the constancy in codon structure exhibited by this trio. The only reasonable conclusion that comes to mind is that passage either to or through the membrane requires a specific factor that reacts to and transports the particular transit peptide and the protein it bears, and no other. Thus the implication is that each cryptomorph and diplomorph is coupled to a specific factor of either the cytoplasm or membrane, as pointed out in another context earlier. In other words, for each transit-peptide-bearing species, or at least species group, of proteins, there is a membrane-transiting protein that enables it to reach its ultimate destination.

The Ultimate Enkephalin Functional Genes. Unlike the α-pheromones of yeasts, all ultimate functional enkephalin sectors of a given gene are not identical, two diverging distinctly from the rest, another having the potential for special processing, and a fourth ending in a codon for a different amino acid (Table 7.18), as already pointed out. Ultimate genes 1–3 and 5 encode the typical met-enkephalin and 6 the somewhat rarer leu-enkephalin, while the product of gene 4 is known as met-enkephalin-arg^6-gly^7-leu^8 and that of gene 7 is met-enkephalin-arg^6-phe^7. In addition, under certain, unknown conditions, ultimate gene 5 may undergo alternative processing as indicated in the table to produce met-enkephalin-arg^6-arg^7-val^8-gly^9, but the possibility remains that this route is followed only under pathological conditions.

Also, in contrast to the earlier example, the connector sequences vary in length, so that ultimate gene 2 follows almost directly after 1 (Figure 7.2B). The third one is preceded by 20 codons in addition to the cleavage sites, the fourth by a variable number exceeding 40 triplets, and the fifth to seventh by 12, 11, and 22, respectively. It is interesting that the longest connector, that between genes 3 and 4, shows a slight decrease in length with increase in phylogenetic position in the Mammalia. Perhaps this sector will prove to be successively longer in bats, marsupials, and monotremes when the gene has been sequenced from those source organisms. One feature of these genes that agrees with those of the yeast is the great length of the prosequence, running to 73 codons before the first cleavage site, although that of cattle is reduced by the loss of three codons in row 3.

Cleavage sites, each consisting of two codons for basic amino acids, are located both before and after all the ultimate genes. It should be especially noted that the paired basic amino acids following the terminus of gene 1 do dual service in also providing for cleavage at the beginning of gene 2. Consequently, at least in this instance, proteolytic cleavage must be able to occur at each end of the double basic amino acids. The pair directly preceding the several ultimate coding regions constantly consists of one triplet for lysine and a second for aginine, but those that begin the connector strands may encode any combination of two basic protein monomers.

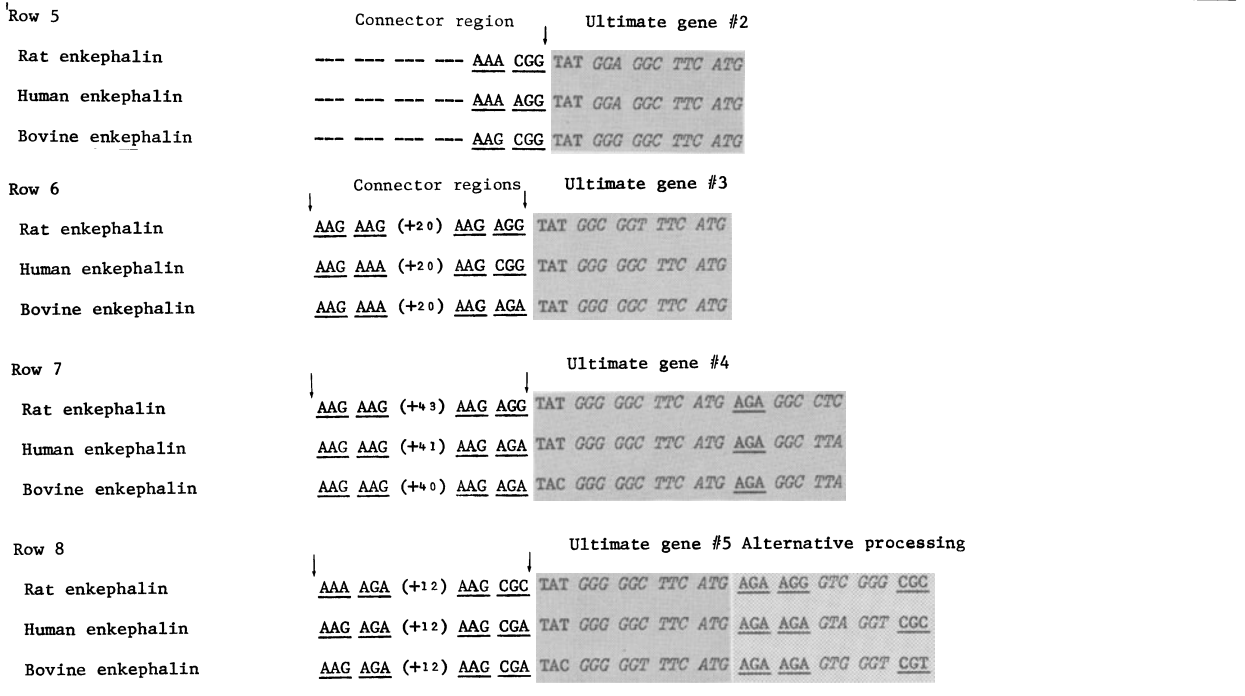
To judge from what can be gleaned from the current literature, quite a few, mostly small hormone cistrons will be found to be representatives of this subclass. For example, the gene for thyrotropin-releasing hormone, a substance abundant in the skin of *Xenopus*, contains four or more spaced repeats of codons for the tripeptide glutamine–histidine–proline that constitutes the active hormone (Richter *et al.*, 1984). Each set is preceded and followed by triplets encoding the lysine–arginine signaling combination.

Table 7.18
Mature Coding Regions of Enkephalin Genes[a]

Row 1	
Rat enkephalin[b]	GAC TGC AGC CAG GAC TGC *GCT* AAA TGC AGC TAC CGC *CTG GTA* CGT *CCC GGC* GAC *ATC* AAC *TTC CTG** *GCA*
Human enkephalin[c]	GAA TGC AGC CAG GAT TGC *GCG* ACG TGC AGC TAC CGC *CTA GTG* CGC *CCG GCC* GAC *ATC* AAC *TTC CTG GCT*
Bovine enkephalin[d]	GAA TGC AGC CAG GAC TGC *GCC* ACG TGC AGC TAC CGC *CTG GCG* CGC *CCG* ACT GAC *CTC* AAC *CCG CTG GCT*
Row 2	
Rat enkephalin	TGC ACA *CTC* GAA TGT GAA *GGG* CAG *CTG CCT* TCT *TTC* AAA *ATC* TGG GAG ACC TGC AAG GAT *CTC CTG* CAG
Human enkephalin	TGC *GTA ATG* GAA TGT GAA *GGT* AAA *CTG CCT* TCT *CTG* AAA *ATT* TGG GAA ACC TGC AAG GAG *CTC CTG* CAG
Bovine enkephalin	TGC ACT *CTG* GAA TGT GAG *GGG* AAA *CTA CCT* TCT *CTC* AAG ACC TGG GAA ACC TGC AAG GAG *CTT CTG* CAG
Row 3	
Rat enkephalin	*GTG* TCC AAG *CCG* GAG *TTC CCT* TGG GAT AAC *ATC* GAC *ATG* TAC AAA GAC AGC AGC AAA CAG GAT GAG AGC
Human enkephalin	*CTG* TCC AAA *CCA* GAG *CTT CCT CAA* GAT *GGC* ACC AGC ACC *CTC* AGA GAA AAT AGC AAA *CCG* GAA GAA AGC
Bovine enkephalin	*CTG* ACC AAA *CTA* GAA *CTT CCT CCA* GAT *GCC* ACC AGT *GCC CTC* --- --- --- AGC AAA CAG GAG GAA AGC
Row 4	Ultimate gene #1 ↓
Rat enkephalin	CAC *TTG CTA GCC* AAG AAG TAT *GGA GGG TTC ATG*
Human enkephalin	CAT *TTG CTA GCC* AAA AGG TAT *GGG GGC TTC ATG*
Bovine enkephalin	CAC *CTG CTT GCT* AAG AAG TAC *GGG GGC TTC ATG*

(*continued*)

Table 7.18 (Continued)



Row 9		Ultimate gene #6
Rat enkephalin	AGA AGG (+11) AAG AGA	TAC *GGA GGC TTC CTG*
Human enkephalin	AGA AGA (+11) AAA CGG	TAT *GGA GGT TTC CTG*
Bovine enkephalin	AGA AGA (+11) AAA AGG	TAC *GGT GGC TTC CTC*
Row 10		Ultimate gene #7
Rat enkephalin	AAG CGC (+22) AAA AGA	TAC *GGA GGC TTT ATG CGG TTT* TGA
Human enkephalin	AAG CGC (+22) AAA AGA	TAC *GGA GGA TTT ATG AGA TTT* TAA
Bovine enkephalin	AAG CGC (+22) AAA AGA	TAT *GGA GGA TTT ATG AGA TTT* TAA

[a]Codons for hydrophilic amino acids are underscored and those for hydrophobic ones are italicized. Asterisk indicates the location of an intron. Arrows indicate cleavage sites. The usual ultimate genes are shaded but a section that can be added to ultimate gene 5 is indicated by lighter stippling.
[b]Rosen *et al.* (1984); Yoshikawa *et al.* (1984).
[c]Legon *et al.* (1982).
[d]Gubler *et al.* (1982); Noda *et al.* (1982).

7.5.3. Subclass II Cryptomorphic Genes

In a sense, the enkephalin family of cryptomorphs just described introduces those of subclass II, for the major distinguishing character here is the combination of two or more ultimate genes encoding different products concealed within a large primary translational product. There can be little doubt that this group is artificially composite, for, as is shown shortly, those classed here fall into two clusters, each containing closely related vertebrate genes, which lack any indication of kinship to those of the other group, although both in part encode important hormones. Here, too, belong the vasoactive intestinal polypeptide genes and their associated proteins (Hefford *et al.*, 1985; Nishizawa *et al.*, 1985).

The Oxytocin Family of Genes. Oxytocin and vasopressin, two structurally closely related nonapeptides, are synthesized by the magnocellular neurons of the supraoptic and suprachiasmatic nuclei of the hypothalamus, whence they are conducted to the neurohypophysis for release into the bloodstream. Both also are synthesized in the corpora lutea of the ovary (Ivell and Richter, 1984b). Chiefly this pair influences water balance in the body, along with cardiovascular functions and smooth muscle contractions (Majzoub *et al.*, 1984). After each presequence, described later, is a short prosequence of four codons, the first and last of which encode hydrophobic amino acids (Table 7.19). No triplet for a charged amino acid is contained here, or in the calf sequence, which is not given (Land *et al.*, 1982). Immediately following this is an ultimate gene encoding either of the hormones oxytocin or vasopressin (Figure 7.2C).

The coding properties of the four shown in the table differ at only two points, the rat vasopressin gene (Ivell and Richter, 1984a; Majzoub *et al.*, 1984) having triplets designating phenylalanine (TTC) and arginine (AGR) in place of those encoding isoleucine (ATY) and leucine (CTG), respectively, in the two oxytocin cistrons (Ivell and Richter, 1984a,b). In the short connecting section that ensues are three codons, the first specifying glycine, the second lysine, and the third arginine; the two examples from rat are identical throughout, while that of the calf oxytocin cistron has CGC in place of the AGA of the others. Beyond this sector is a long region of ~92 triplets that codes for a polypeptide called neurophycin. This protein has been stated to be involved in the transport of oxytocin or vasopressin to the posterior pituitary (Ivell and Richter, 1984a), but it is not clear whether this activity is carried out before or after translation and processing. If afterward, then that is a valid activity; if before processing, then the neurophycin is merely part of the precursor that is conducted, and its function, if any, remains unknown.

In the vasopressin gene a coding region for a glycoprotein of undetermined activity abuts against the 3′ end of the neurophycin sequence, but in those for oxytocin only a single codon intervenes between the latter and the translational termination signal. This can be considered the shortest possible telosequence, a feature absent from the genes for vasopressin. In summary, it may be perceived that both the oxytocin and vasopressin genes consist of six parts, a presequence, a prosequence, an ultimate section, a connector, the neurophycin region, and either a telosequence or a glycoprotein area, in addition to the universal termination signal of all protein genes.

The connector sector, where the two peptides are cleaved proteolytically, is of particular interest in that its structure differs from those of subclass I. Here the neurophycin region is preceded by a pair of codons for lysine and arginine, as is typical; however, the hormone coding sector has only a single codon following it, one that

Table 7.19
Mature Coding Sections of Subclass II Cryptomorphic Genes[a]

Row 1

Prosequence ↓ Hormone ↓ ↓ Neurophycin ⟶

Rat oxytocin[b] CTG ACC TCC GCC TGC TAC ATC CAG AAC TGC CC C CTG GGC GGC AAG AGG GCT GCG CTA GAC CTG GAT ATC

Bovine oxytocin[c] TTG ACC TCC GCC TGC TAC ATT CAG AAC TGC CC C CTG GGC GGC AAA CGC GCG GTG CTG GAC CTC GAC GTG

Rat vasopressin[b] CTC ACC TCT GCC TGC TAC TTC CAG AAC TGC CC A AGA GGA GGC AAG AGG GCC ACA TCC GAC ATG GAG CTG

Bovine vasopressin[d] TTC ACC TCT GCT TGC TAC TTC CAG AAC TGC CC A AGG GGC GGC AAG AGG GCC ATG TCC GAC CTG GAG CTG

Anglerfish glucagon[e] CGG GTT CTT ATG [illegible] GAG GCA GAC AGC ACA CTG AAG GAC GAG CCG

Rat glucagon[f] CAT GCC CCT CAA GAC ACG GAG GAG AAC GCC AG*A TCA TTC CCA GCT TCC CAG ACA GAA CCA CTT GAA GAC

Hamster glucagon[g] CAT TCC CTT CAG GAC ACG GAG GAG AAA TCC AG*A TCA TTC CCA GCT TCC CAG ACA GAC CCA CTC GAG GAC

Human glucagon[h] CGT TCC CTT CAA GAC ACA GAG GAG AAA TCC AG*A TCA TTC TCA GCT TCC CAG GCA GAC CCA CTC AGT GAT

Bovine glucagon[i] CGT TCC CTT CAG AAC ACA GAG GAG AAA TCC AG T TCA TTC CCA GCT CCG CAG ACC GAC CCG CTC GGC GAT

Glicentin ⟶

Row 2

Rat oxytocin CGC AAG*TGT CTT CCC TGC GGA CCC GGC GGC AAA GGG CGC TGC TTC GGG CCG AGC ATC TGC TGC GCG GAC

Bovine oxytocin CGC ACG TGT CTC CCC TGC GGC CCC GGG GGC AAA GGC CGC TGC TTC GGG CCC AGC ATC TGC TGC GGG GAC

Rat vasopressin AGA CAG*TGT CTC CCC TGC GGC CCT GGC GGC AAA GGG CGC TGC TTC GGG CCG AGC ATC TGC TGC GCG GAC

Bovine vasopressin AGA CAG TGT CTC CCC TGC GGC CCC GGG GGC AAA GGC CGC TGC TTC GGG CCC AGC ATC TGC TGC GGG GAC

Anglerfish glucagon AGA GAG CTT TCA AAC ATG --- AAG AGA CAC TCG GAG GGA ACT TTC TCC AAC GAC TAC AGC AAA TAC CTG

Rat glucagon CCT GAT CAG ATA AAC GAA GAC AAA CGC CAT TCA CAG GGC ACA TTC ACC AGT GAC TAC AGC AAA TAC CTA

Hamster glucagon CCT GAT CAA ATA AAT GAA GAC AAG CGC CAT TCA CAG GGA ACA TTC ACC AGT GAC TAC AGC AAA TAC CTG

Human glucagon CCT GAT CAG ATG AAC GAG GAC AAG CGC CAT TCA CAG GGC ACA TTC ACC AGT GAC TAC AGC AAG TAT CTG

Bovine glucagon CCA GAT CAG ATC AAT GAA GAT AAG CGC CAC TCG CAG GGC ACA TTC ACC AGT GAC TAC AGC AAG TAC CTG

⟵ Glicentin ↑ ↑ Glucagon ⟶

(continued)

Table 7.19 (Continued)

Row 3

Rat oxytocin	GAG CTG GGC TGC TTC GTG GGC ACC GCC GAG GCG CTG CGC TGC CAG GAG GAG AAC TA C CTG CCC TCG CCC
Bovine oxytocin	GAG CTG GGC TGC TTC GTG GGC ACG GCC GAG GCG CTG CGC TGC CAA GAG GAG AAC TA C CTG CCG TCG CCC
Rat vasopressin	GAG CTG GGC TGC TTC CTG GGC ACC GCC GAG GCG CTG CGC TGC CAG GAG GAG AAC TA C CTG CCC TCG CCC
Bovine vasopressin	GAG CTG GGC TGC TTC GTG GGC ACG GCC GAG GCG CTG CGC TGC CAA GAG GAG AAC TA C CTG CCG TCG CCC
Anglerfish glucagon	GAG GAC AGG AAG GCA CAG GAG TTT GTT CGG TGG CTG ATG AAC AAC AAG AGG AGC GG T GTG GCA GAA ---
Rat glucagon	GAC TCC CGC CGT GCT CAA GAT TTT GTG CAG TGG TTG ATG AAC ACC AAG AGG AAC CG*G AAC AAC ATT GCC
Hamster glucagon	GAC TCC CGC CGA GCC CAA GAT TTT GTG CAG TGG CTG ATG AAC ACC AAG AGG AAC AG G AAC AAC ATT GCC
Human glucagon	GAC TCC AGG CGT GCC CAA GAT TTT GTG CAG TGG TTG ATG AAT ACC AAG AGG AAC AG*G AAT AAC ATT GCC
Bovine glucagon	GAC TCC AGG CGT GCC CAG GAC TTC GTG CAG TGG TTG ATG AAT ACC AAG AGA AAC AA G AAT AAC ATT GCC
	←Glucagon

Row 4

Rat oxytocin	TGC CAG TCT GGC CAG AAG CCT TGC GGA AGC GGA GGC CGC TGC GCC ACC GCG GGC ATC TGC TGT AGC CCG
Bovine oxytocin	TGC CAG TCC GGC CAG AAG CCC TGC GGG AGC GGG GGC CGC TGC GCC GCC GCC GGC ATC TGC TGC AGC CCG
Rat vasopressin	TGC CAG TCT GGC CAG AAG CCT TGC GGA AGC GGA GGC CGC TGC GCC ACC GCG GGC ATC TGC TGT AGC GAT
Bovine vasopressin	TGC CAG TCC GGC CAG AAG CCC TGC GGG AGC GGG GGC CGC TGC GCC GCC GCC GGC ATC TGC TGC AAC GAT
Anglerfish glucagon	AAG CGT CAC GCT GAT GGG ACC TTC ACC AGC --- --- --- GAT GTC AGC TCC TAC CTC AAA GAC CAG GCA
Rat glucagon	AAA CGT CAT GAT GAA TTT GAG AGG CAT GCT GAA GGG ACC TTT ACC AGT GAT GTG AGT TCT TAC TTG GAG
Hamster glucagon	AAA CGC CAC GAT GAG TTT GAG AGG CAC GCT GAA GGG ACC TTT ACC AGC GAT GTG AGC TCT TAC TTG GAG
Human glucagon	AAA CGT CAC GAT GAA TTT GAG AGA CAT GCT GAA GGG ACC TTT ACC AGT GAT GTA AGT TCT TAT TTG GAA
Bovine glucagon	AAA CGT CAT GAT GAA TTT GAG AGA CAT GCT GAA GGG ACC TTT ACC AGT GAT GTA AGT TCT TAT TTG GAA
	↑Glucagonlike protein I ——→

Row 5

	← Neurophysin
Rat oxytocin	G*AT GGC TGC CGC ACC GAC CCC GCC TGC GAC CCT GAG TCT GCC TTC TCC GAG CGC TG A
Bovine oxytocin	G AC GGC TGC CAC GAG GAC CCC GCC TGC GAC CCT GAG GCC GCC TTC TCC CAG CAC TG A
Rat vasopressin	G*AG AGC TGC GTG GCC GAG CCC GAG TGT CGA GAG GGT TTT TTC CGC CTC ACC CGC GC T CGG --- --- GAG
Bovine vasopressin	G AG AGC TGC GTG ACC GAG CCC GAG TGC CGG GAA GGT GTC GGC TTC CCC CGC CGC GT T CGC GCC AAC GAC
Anglerfish glucagon	A TC AAA GAC TTT GTG GAC AGG CTC AAG GCT GGA CAA GTC AGA AGA GAG TAG Glycoprotein ⟶
Rat glucagon	G GC CAG GCA GCA AAG GAA TTC ATT GCT TGG CTG GTG AAA GGC CGA GGA AGG CGA GA*C TTC CCG GAA GAA
Hamster glucagon	G GC CAG GCT GCA AAG GAA TTC ATT GCT TGG CTG GTG AAA GGC AGA GGA AGG CGG GA C TTC CCA GAA GAA
Human glucagon	G GC CAA GCT GCC AAG GAA TTC ATT GCT TGG CTG GTG AAA GGC CGA GGA AGG CGA GA*T TTC CCA GAA GAG
Bovine glucagon	G GC CAA GCT GCC AAG GAA TTC ATT GCT TGG CTG GTG AAA GGC CGA GCA AGG CGA GA T TTC CCA GAA GAA
	← Glucagonlike peptide I

Row 6

Rat vasopressin	CAG AGC AAC GCC ACG CAG CTG GAC GGG CCA GCC CGG GAG CTG CTG CTT AGG CTG GTA CAG CTG GCT GGG
Bovine vasopressin	CGG AGC AAC CGC ACC CTG CTG GAC GGG CCG AGC GGG GCC TTG TTG CTG CGG CTG GTG CAG CTG GCG GGG
Rat glucagon	GTC GCC ATA GCT GAG GAA CTT GGG CGC AGA CAT GCT GAT GGA TCC TTC TCT GAT GAG ATG AAC ACG ATT
Hamster glucagon	GTC ACC ATT GTT GAA GAA CTC GGC CGC AGA CAT GCG GAC GGC TCC TTC TCC GAT GAG ATG AAC ACG ATT
Human glucagon	GTC GCC ATT GTT GAA GAA CTT GGC CGC AGA CAT GCT GAT GGT TCT TTC TCT GAT GAG ATG AAC ACC ATT
Bovine glucagon	GTC AAC ATC GTT GAA GAA CTC CGC CGC AGA CAC GCC GAT GGC TCT TTC TCT GAT GAG ATG AAC ACT GTT
	↑ Glucagonlike peptide II ⟶

(*continued*)

Table 7.19 (Continued)

Row 7	Glycoprotein
Rat vasopressin	ACA CAA GAG TCC GTG GAT TCT GCC AAG CCC CGG GTC TAC TGA
Bovine vasopressin	GCG CCG GAG CCC GCG GAG CCC GCC CAG CCC GGC GTC TAC TGA
Rat glucagon	CTC GAT AAC CTT GCC ACC AGA GAC TTC ATC AAC TGG CTG ATT CAA ACC AAG ATC ACT GAC AA*G AAA TAG
Hamster glucagon	CTC GAT AGT CTT GCC ACC AGG GAC TTC ATC AAC TGG CTG ATT CAA ACC AAA ATC ACT GAC AA G AAA TAA
Human glucagon	CTT GAT AAT CTT GCC GCC AGG GAC TTT ATA AAC TGG TTG ATT CAG ACC AAA ATC ACT GAC AG G --- TGA
Bovine glucagon	CTC GAT ACT CTT GCC ACC CGA GAC TTT ATA AAC TGG TTG CTT CAG ACG AAA ATT ACT GAC AG G AAG TAA
	← Glucagonlike peptide II — Telopeptide

[a]Asterisks indicate the location of an intron. Vertical arrows indicate cleavage sites. Codons for the basic amino acids (lysine and arginine) are underscored. Mature genes are enclosed in open boxes.
[b]Ivell and Richter (1984a).
[c]Ivell and Richter (1984b).
[d]Land *et al.* (1982).
[e]Lund *et al.* (1983).
[f]Heinrich *et al.* (1984b); Patzelt and Schiltz (1984).
[g]Bell *et al.* (1983a).
[h]Bell *et al.* (1983b).
[i]Lopez *et al.* (1983).

specifies the small hydrophobic amino acid glycine. Hence, a second enzyme may be concerned in cleavage of the connector from the sector that precedes it. As indicated by the asterisks in rows 2 and 5, the neurophycin sequences of the rat are interrupted by two introns, placed at corresponding points in the genes both for oxytocin and vasopressin, evidence in addition to the high level of homology throughout all their parts that the cistrons for the two hormones have been derived from a common ancestor in relatively recent times.

The Glucagon Family of Genes. Glucagon is a member of a family of peptide hormones that also includes secretin, vasoactive intestinal peptide, gastric inhibitory peptide, and growth hormone-releasing hormones. However, it is the only component whose gene structure has been sufficiently documented to be reported. Currently the sequence of the cistron for this substance has been established from five sources, all of which are included in Table 7.19.

Under the control of blood levels of glucose, various amino acids, and several hormones (Heinrich *et al.*, 1984b), glucagon is secreted by the A cells of the islets of the pancreas. Its chief target organ is the liver, where it plays an important role in protein and carbohydrate metabolism. Chiefly it is concerned with glucose metabolism, through actions that inhibit glycogen synthesis, accelerate glycogen breakdown, and stimulate formation of glucose. In the genes that encode this 29-amino acid peptide, a presequence but no prosequence is found, the 5′-terminal region encoding a protein called glicentin, a peptide with much the same activity as glucagon itself (Figure 7.2D). This 30-codon section is separated by a pair of codons specifying lysine and arginine from the 29-codon-long ultimate gene for the glucagon. Thus here the protease (or proteases) acts at each end of this combination of basic amino acids.

After the glucagon gene, there is a connector of ten codons, including a dual combination for lysine and arginine at each end; this is followed by a stretch of 37 triplets, the ultimate coding sector for what is referred to as "glucagonlike peptide I." Then, in mammalian cistrons, but not those from the anglerfish, a connector of 17 codons follows, provided with coding signals for two arginines at the 5′ end and for two or three of the same at the other terminus. The latter serves as the apparent cleavage site from which a 33-codon-long region encoding "glucagonlike peptide II" is removed following translation. Finally there is a short telosequence of either two lysine codons or one arginine and one lysine. In the anglerfish, the glucagonlike peptide I coding region is followed by a short telosequence of three codons, beginning with two for arginine. Recently it has been proposed that the early cleavages of processing act at three points to separate glicentin, glucagon, and the intact 3′ half containing the two glucagonlike peptides; however, the remaining steps in activation remain undetermined (Patzelt and Schiltz, 1984). Other genes that fall into this category are still more complexly structured than the examples cited, but current information is not sufficiently extensive to provide for a precise detailed discussion. One is of the particular note, that for a precursor known as pro-opiomelacortin, for it encodes three products, β-endorphin, melanocyte-stimulating hormone, and corticotropin (ACTH) (Oates and Herbert, 1984).

Thus the glucagon cryptomorphic gene consists of seven or eight regions, a presequence, three or four ultimate genes (one each for glicentin and glucagon, and either one or two for glucagonlike peptides), two connecting sectors (one in fish), and a telosequence. The obvious distinctions of this arrangement from that of the oxytocin provide

Table 7.20
Presequences of Cryptomorphic Genes of Subclass II[a]

Row 1	
Rat oxytocin[b]	*ATG GCC* TGC *CCC* AGT --- *CTC* --- *GCT* TGC TGC *CTG*
Bovine vasopressin[c]	*ATG* --- *CCC* GAC *GCC* ACA *CTG CCC GCC* TGC *TTC CTC*
Rat vasopressin[b,d]	*ATG ATG CTC* AAC ACT ACG *CTC* TCT *GCT* TGC *TTC CTG*
Bovine glucagon[e]	*ATG* AAA AGC *CTT* TAC *TTT GTG GCT GGA TTG TTT GTA*
Rat glucagon[f]	*ATG* AAG ACC *GTT* TAC *ATC GTG GCT GGA TTG TTT GTA*
Hamster glucagon[g]	*ATG* AAG AAC *ATT* TAC *ATT GTG GCT GGA TTT TTT* TGT
Anglerfish glucagon[h]	*ATG* AAA CGC *ATC* CAC TCC *CTG GCT GGT ATC CTT CTG*
Row 2	
Rat oxytocin	*CTT GGC CTA* --- --- --- --- *CTG* GCT ↓ *CTG* ACC
Bovine vasopressin	--- AGC *CTG* --- --- --- --- *CTG* GCC *TTC* ACC
Rat vasopressin	--- AGC *CTG* --- --- --- --- *CTG* GCC *CTC* ACC
Bovine glucagon	*ATG CTG GTA* CAA *GGC* --- AGC TGG CAA CGT TCC
Rat glucagon	*ATG CTG GTA* CAA *GGC* --- AGC TGG CAG CAT *GCC*
Hamster glucagon	*GGT GCT GGT* CAA *GGC* --- AGC TGG CAG CAT TCC
Anglerfish glucagon	*GTG CTT GGT TTA ATC* CAG AGC AGC TGC CGG *GTT*

[a]Codons for charged amino acids are underscored and those for hydrophobic ones are italicized. The arrow marks the cleavage site.
[b]Ivell and Richter (1984a).
[c]Land *et al.* (1982).
[d]Majzoub *et al.* (1984).
[e]Lopez *et al.* (1983).
[f]Heinrich *et al.* (1984b).
[g]Bell *et al.* (1983a).
[h]Lund *et al.* (1983).

firm evidence that subclass II as here presented is not a natural grouping but merely one of convenience for immediate needs.

Subclass II Cryptomorph Presequences. The presequences of the two families examined here as subclass II of cryptomorphic genes display far more shared characteristics than do their mature coding regions (Table 7.20). In both cases these sectors are relatively short, consisting of 15–20 codons, at most one of which per sequence encodes a charged amino acid. They differ strongly, however, in the distribution of triplets for hydrophobic amino acids. Among the oxytocin family members, that type forms the larger part of the structure, 11 of the 15 of oxytocin and 10 or 11 of vasopressin falling in this category. On the other hand, the presequences of the glucagon family are less heavily equipped with such codons, but they have them grouped centrally, where ten occur without interruption, neither singly nor paired as in the rest. At the cleavage site there is

no recognizable common system to signal the protease, although each family shows a high level of conservation here. The first family has two apolar codons preceding the terminus, one for a moderate-sized amino acid, the other for a small one, whereas the second group encodes the hydrophilic amino acids tryptophan and glutamine, both of rather large size. On the 3′ side of the cleavage site, the three members of the oxytocin family have codons for either leucine or phenylalanine, followed by threonine, while the glucagon cistrons vary, specifying either arginine or histidine at the first site and serine or valine at the second. Only the last two can be rated as small amino acids. Consequently, we remain without a clue as to the nature of the signal for the cleaving enzyme.

Evolutionary Notes. The availability of a glucagon gene sequence from a lower vertebrate opens an avenue permitting speculation as to the possible origin of its compound nature. Since this more primitive cistron lacks the nucleotides encoding a second glucagonlike peptide, it is obvious that the latter arose by a duplicative process in some higher form. Whether this event occurred in some amphibian or reptile or only in the lower mammals cannot be determined until its primary structure has been established from an avian source. This proposal, besides being self-evident, is along standard lines in current literature, but the rat and human genes provide data that seem to carry farther the process of development of compound structures such as the present one. In these two, the DNA, not just the mRNA, provided the basis for sequencing, so the several introns that exist have been detected. One such is found in row 5 of Table 7.19 at a point corresponding to the penultimate site in the anglerfish telosequence. Consequently, it may be that in some instances such as here an intron is associated with the point of duplication of a gene segment, although in what capacity remains unclear. That this may be the case is further suggested by the presence of an inserted region close after the 3′ end of the section encoding glucagon itself. Thus glucagonlike peptide I may be the result of a prepiscine duplication of the glucagon sequence, with a later duplication producing the coding section for glucagonlike peptide II. There is no evidence as to the origin of the insert located near the middle of the glicentin region. Establishment of the glucagon gene structure from agnathans or elasmobranchs probably will be necessary to disclose the actual steps in the phylogeny of this unusual complex gene.

7.5.4. Cryptomorphic Genes of Subclass III

Since in the glucagon family of subclass II genes just analyzed, the region encoding glicentin was actually a prosequence, it appears logical to place in the ensuing group, subclass III, other but different cistron structures that share this same feature. Here the distinctive characteristics are the existence of a single ultimate gene, which displaces the prosequence of a much larger DNA coding frame, most of whose product is without known function. In short, a relatively small ultimate functional gene is hidden in the 5′ section of a large structure, the product not becoming active until the major 3′ part is removed enzymatically. Only three types of hormones of vertebrates are currently known to represent this subclass, luteinizing hormone-releasing hormone, angiotensinogen, and ubiquitin (Lund *et al.*, 1985).

Luteinizing Hormone-Releasing Hormone. Luteinizing hormone-releasing hormone (LHRH), also called gonadotropin-releasing hormone, is an important element in the control of reproduction in vertebrates. Produced by hypothalamic neurons, it is se-

Table 7.21
Gene Structures of Subclass III Cryptomorphs[a]

Row 1	Ultimate gene ↓
Human LHRH[b]	CAG CAC TGG TCC TAT GGA CTG CGC CCT GGA AAG AGA GAT GCC GAA AAT TTG ATT GAT TCT TTC CAA GAG*
Rat angiotensin[c]	GAC CGC GTA TAC ATC CAC CCC TTT CAT CTC CTC TAC TAC AGC AAG AGC ACC TGC GCC CAG CTG GAG AAC
Human angiotensin[d]	GAC CGG GTG TAC ATA CAC CCC TTC CAC CTC GTC ATC CAC AAT GAG AGT ACC TGT GAG CAG CTG GCA AAG
Row 2	
Human LHRH	ATA GTC AAA GAG GTT GGT CAA CTG GCA GAA ACC CAA CGC TTC GAA TGC ACC ACG CAC CAG CCA CGT TCT
Rat angiotensin	CCC AGT GTG GAG ACG CTC CCA GAG CCA ACC TTT GAG CCT GTG CCC ATT CAG GCC AAG ACC TCC CCC GTG
Human angiotensin	GCC AAT GCC GGG AAG CCC AAA GAC CCC ACC TTC ATA CCT GCT CCA ATT CAG GCC AAG ACA TCC CCT GTG
Row 3	↓ Telosequence
Human LHRH	CCC CTC CGA GAC CTG AAA GGA GCT CTG* GAA AGT CTG ATT GAA GAG GAA ACT GGG CAG AAG AAG ATT TAA
Rat angiotensin	GAT GAG AAG ACC CTG CGA GAT AAG CTC GTG CTG GCC ACT GAG AAG CTA GAG GCT GAG GAT CGG CAG CGA
Human angiotensin	GAT GAA AAG GCC CTA CAG GAC CAG CTG GTG CTA GTC GCT GCA AAA CTT GAC ACC GAA GAC AAG TTG AGG
Row 4	
Rat angiotensin	GCT GCC CAG GTC GCG ATG ATT GCC AAC TTC ATG GGT TTC CGC ATG TAC AAG ATG CTG AGT GAG GCA AGA
Human angiotensin	GCC GCA ATG GTC GGG ATG CTG GCC AAC TTC TTG GGC TTC CGT ATA TAT GGC ATG CAC AGT GAG CTA TGG
Row 5	
Rat angiotensin	GGT GTA GCC AGT GGG GCC --- GTC CTC TCT CCA CCG GCC CTC TTT GGC ACC CTG GTC TCT TTC TAC CTT
Human angiotensin	GGC GTG GTC CAT GGG GCC ACC GTC CTC TCC CCA ACG GCT GTC TTT GGC ACC CTG GCC TCT CTC TAT CTG

Row 6

Rat angiotensin	GGA TCG TTG GAT CCC ACG GCC AGC CAG CTG CAG GTG CTG CTG GGC GTC CCT GTG AAG GAG GGA GAC TGC
Human angiotensin	GGA GCC TTG GAC CAC ACA GCT GAC AGG CTA CAG GCA ATC CTG GGT GTT CCT TGG AAG GAC AAG AAC TGC

Row 7

Rat angiotensin	ACC TCC CGG CTG GAC GGA CAT AAG GTC CTC ACT GCC CTG CAG GCT GTT CAG GGC TTG CTG GTC ACC CAG
Human angiotensin	ACC TCC CGG CTG GAT GCG CAC AAG GTC CTG TCT GCC CTG CAG GCT GTA CAG GGC CTG CTA GTG GCC CAG

Row 8

Rat angiotensin	GGT GGA AGC AGC AGC CAG ACA CCC CTG CTA CAG TCC ACC GTG GTG GGC CTC TTC ACT GCC CCA GGC TTG
Human angiotensin	GGC AGG GCT GAT AGC CAG GCC CAG CTG CTG CTG TCC ACG GTG GTG GGC GTG TTC ACA GCC CCA GGC CTG

Row 9

Rat angiotensin	CGC CTA AAA CAG CCA TTT GTT GAG AGC TTG GGT CCC TTC ACC CCC GCC ATC TTC CCT CGC TCT CTG GAC
Human angiotensin	CAC CTG AAG CAG CCG TTT GTG CAG GGC CTG GCT CTC TAT ACC CCT GTG GTC CTC CCA CGC TCT CTG GAC

Row 10

Rat angiotensin	TTA TCC ACT GAC CCA GTT CTT GCT GCC CAG AAA ATC AAC AGG TTT GTG CAG GCT GTG ACA GGG TGG AAG
Human angiotensin	TTC --- ACA GAA CTG GAT GTT GCT GCT GAG AAG ATT GAC AGG TTC ATG CAG GCT GTG ACA GGA TGG AAG

Row 11

Rat angiotensin	ATG AAC TTG CCA CTA GAG GGG GTC AGC ACG GAC AGC ACC CTA TTT TTC AAC ACC TAC GTT CAC TTC CAA
Human angiotensin	ACT GGC TGC TCC CTG ATG GGA GCC AGT GTG GAC AGC ACC CTG GCT TTC AAC ACC TAC GTC CAC TTC CAA

(*continued*)

Table 7.21 (Continued)

Row 12	
Rat angiotensin	GGG AAG ATG AGA GGC TTC TCC CAG CTG ACT GGG CTC CAT GAG TTC TGG GTG GAC AAC AGC ACC TCA GTG
Human angiotensin	GGG AAG ATG AAG GGC TTC TCC CTG CTG GCC GAG CCC CAG GAG TTC TGG GTG GAC AAC AGC ACC TCA GTG
Row 13	
Rat angiotensin	TCT GTG CCC ATG CTC TCG GGC ACT GGC AAC TTC CAG CAC TGG AGT GAC GCC CAG AAC AAC TTC TCC GTG
Human angiotensin	TCT GTT CCC ATG CTC TCT GGC ATG GGC ACC TTC CAG CAC TGG AGT GAC ATC CAG GAC AAC TTC TCG GTG
Row 14	
Rat angiotensin	ACA CGC GTG CCC CTG GGT GAG AGT GTC ACC CTG CTG CTG ATC CAG CCC CAG TGC GCC TCA GAT CTC GAC
Human angiotensin	ACT CAA GTG CCC TTC ACT GAG AGC GCC TGC CTG CTG CTG ATC CAG CCT CAC TAT GCC TCT GAC CTG GAC
Row 15	
Rat angiotensin	AGG GTG GAG GTC CTC GTC TTC CAG CAC GAC TTC CTG ACT TGG ATA AAG AAC CCG CCT CCT CGG GCC ATC
Human angiotensin	AAG GTG GAG GGT CTC ACT TTC CAG CAA AAC TCC CTC AAC TGG ATG AAG AAA CTG TCT CCC CGG ACC ATC
Row 16	
Rat angiotensin	CGT CTG ACC CTG CCG CAG CTG GAA ATT CGG GGA TCC TAC AAC CTG CAG GAC CTG CTG GCT CAG GCC AAG
Human angiotensin	CAC CTG ACC ATG CCC CAA CTG GTG CTG CAA GGA TCT TAT GAC CTG CAG GAC CTG CTC GCC CAG GCT GAG
Row 17	
Rat angiotensin	CTG TCT ACC CTT TTG GGT GCT GAG GCA AAT CTG GGC AAG ATG GGT GAC ACC AAC CCC CGA GTG GGA GAG
Human angiotensin	CTG CCC GCC ATT CTG CAC ACC GAG CTG AAC CTG CAA AAA TTG AGC AAT GAC CGC ATC AGG GTG GGG GAG

Row 18	
Rat angiotensin	GTT CTC AAC AGC ATC CTC CTT GAA CTC CAA GCA GGC GAG GAG GAG CAG CCC ACA GAG TCT GCC CAG CAG
Human angiotensin	GTG CTG AAC AGC ATT TTT TTT GAG CTT GAA GCG --- GAT GAG AGA GAG CCC ACA GAG TCT ACC CAA CAG
Row 19	
Rat angiotensin	CCT GGC TCA CCC GAG GTG CTG GAC GTG ACC CTG AGC AGT CCG TTC CTG TTC GCC ATC TAC GAG CGG GAC
Human angiotensin	CTT AAC AAG CCT GAG GTC TTG GAG GTG ACC CTG AAC CGC CCA TTC CTG TTT GCT GTG TAT GAT CAA AGC
Row 20	
Rat angiotensin	TCA GGT GCG CTG CAC TTT CTG GGC AGA GTG GAT AAC CCC CAA AAT GTG GTG TGA
Human angiotensin	GCC ACT GCC CTG CAC TTC CTG GGC CGC GTG GCC AAC CCG CTG AGC ACA GCA TGA

[a]Codons for basic amino acids are underscored. Arrows indicate cleavage sites. Asterisks indicate location of introns.
[b]Luteinizing hormone-releasing hormone; Seeburg and Adelman (1984).
[c]Ohkuba *et al.* (1983).
[d]Kageyama *et al.* (1984).

creted into capillaries to induce the release of luteinizing and follicle-stimulating hormones from the anterior pituitary. A small protein, it consists of only nine amino acids and is encoded by a region of 30 nucleotide residues at the very outset of the mature coding sector of its gene, that is, immediately following the presequence (Table 7.21). If the remainder of the 72-triplet-long gene encodes a useful product, its function remains unelucidated. That it may have some physiological activity is suggested by the presence of a possible telosequence having much the same structure as those of subclass II cryptomorphs in consisting of two codons for lysine, followed by one for an uncharged amino acid. Two introns interrupt the coding region for the unknown substance, but their significance similarly remains obscure.

The ultimate gene section is followed immediately by a combination of lysine and arginine codons typical of many vertebrate gene cleavage signals. In this case, however, it is preceded by a triplet for glycine, which amino acid is employed in amidation of the carboxyl end of the active hormone (Seeburg and Adelman, 1984), so that the peptide actually consists of only nine monomeric units, not ten as sometimes stated in the literature.

The Angiotensinogen Family of Genes. Angiotensinogen, when secreted by the liver into the bloodstream, is a molecule consisting of 453 amino acids, plus a transit peptide of 24 such residues (Figure 7.2E). Although this precursor is thus quite large, when cleaved by renin the angiotensin I that is released is only ten residues in length. This must be processed further by another protease, called dipeptidyl carboxylpeptidase, during which activity it loses two residues at the carboxyl end to produce the functional octapeptide angiotensin II (Ohkuba *et al.*, 1983). When thus mature, the hormone in the bloodstream induces arteriolar constriction and stimulates the adrenal cortex to release aldosterone. In addition, angiotensin is synthesized in the brain by like processes, where it causes thirst and is active in the control of vasopressin and corticotropin release. Briefly stated, it thereby plays an important role in control of blood pressure and water balance in the body.

As in luteinizing hormone-releasing hormone, the ultimate gene is located at the 5′ end of the mature cistron, immediately after the presequence. Also as there, this region is short, containing just 30 nucleotide residues, exactly as in the other member of this subclass. Unlike its predecessor, however, this region is not followed by codons for either lysine or arginine, nor in fact by any recognizable signaling combination. Thus, at present the processes of recognition of the cleavage site by renin are totally unknown. Whether the mature gene section is broken by introns also is unknown, since both the rat and human sequences were established from DNA complementary to the processed messengers (Ohkuba *et al.*, 1983; Kageyama *et al.*, 1984).

Whether calcitonin, a thyroid-produced enzyme associated with calcium metabolism, should be considered a member of the present or next subclass could not be established, since the precise limits of the presequence are not determined (Le Moullec *et al.*, 1984). The single ultimate gene is followed by a 23-codon-long sector that is proteolytically removed posttranslationally, but the unusually long presequence (84 codons) suggests that that part may actually be a prosequence, at least in part. In addition, the gene for gastrin-releasing hormone, a 27-amino acid peptide released by the stomach and upper intestine of mammals and the proventriculus of birds, belongs in this subclass (Spindel *et al.*, 1984).

7.5.5. Subclass IV Cryptomorphic Genes

The members of subclass IV cryptomorphic genes grade into the diplomorphic variety and could justifiably be considered extreme examples of those that bear both pre- and prosequences. As a matter of observation, one family has been referred to as preproproteins in the literature (Montminy *et al.*, 1984). The chief reason for classifying the components as cryptomorphs lies in the exceptionally great proportions of what otherwise would be considered a prosequence. Here the ultimate genes are short, as in subclass III, but lie at or toward the 3′-terminal region rather than at the extreme 5′ end. As shown shortly, the large functionless section greatly exceeds the region encoding the ultimate product. Since the small active part is thus hidden within a huge precursorial form, the several examples currently known appear better to be treated as cryptomorphic genes.

Kininogen Genes. Despite their pharmacological and medical importance, the kinins have been relatively poorly explored at the molecular level. They are small peptides of 9–11 amino acids, in this respect resembling a number of other hormonal products of cryptomorphic genes. Of the two more abundant members of the family, bradykinin and kallidin, the sequence of the former alone has been established and that only from human and bovine liver (Kitamura *et al.*, 1983, 1985; Nawa *et al.*, 1983). In the genomes of these mammals, two cistrons appear to exist which encode quite similar kininogens, as the precursorial molecules are known. In addition, two forms of precursor occur, of high and low molecular weight, the first representing the complete translational product, the other the amino acid half, the carboxyl portion being removed proteolytically at the point represented by a gap in row 18 (Table 7.22). Since the ultimate gene region just precedes this point, in one case it is located at the middle and in the other near the 3′ terminus of the mature gene. At activation, the bradykinin is released by action of the kallikreins examined earlier in this chapter (Section 7.4.1).

The gene, which is of unusually great length, is unique among cryptomorphic forms in lacking a presequence (Figure 7.2H). However, the 5′ end of the mature coding region contains a high percentage of codons for hydrophobic amino acids, 11 of the first 15 triplets encoding that type of monomer; its typical presequencelike structure is further enhanced by the presence of a codon for a basic amino acid located in the second site (Table 7.22). Thus this part may serve in transit through the cell membrane during the secretory processes in lieu of a removable transit peptide. But the actual procedures employed in penetrating the cell membrane are still to be established, a statement equally applicable to all secreted products studied in this newly opened field of investigation. Toward the 3′ end is a peculiarity, the significance of which is presently unknown. Beginning in row 18 there are numerous codons for histidine (CAC, CAT) that continue at a high level of frequency into row 22, where they abruptly cease. At some areas these codons occur at every two or three sites, with the CAT combination greatly preferred over the other. This region, it would seem, should offer investigators a unique opportunity for study of the particular qualities of the amino acid in a naturally occurring product. *In vivo* the activated bradykinin has functions similar to oxytocin, vasopressin, angiotensin, and the other vasoactive hormones, for it produces smooth muscle contraction, induces dilation of blood vessels (resulting in lowering of blood pressure and increased blood flow), and influences the emigration of granulocytic leukocytes.

Table 7.22
Sequences of Subclass IV Cryptomorphic Genes[a]

Row 1

Bovine kininogen[b]	ATG AAA TTA ATC ACC ATC CTT TT C CTT *TGT* TCC AGG CTG --- CTA CCA AGT TTA ACC CAA GAG TCC TCT CAA
Human somatocrinin[c]	CCA CCT CCC CCT TTG ACC CTC AG*G ATG CGG CGG TAT GCA GAT GCC ATC TTC ACC AAC AGC TAC CGG AAG GTG
Human somatostatin[d]	GCT CCC TCG GAC --- CCC AGA CT C CGT CAC TTT --- CTG --- CAG AAG TCC CTG GCT GCT GCC GCG GGG AAG
Rat somatostatin[e]	GCG CCC TCG GAC --- CCC AGA CT C CGT CAG TTT --- CTG --- CAG AAG TCT CTG GCG GCT GCC ACC GGG AAA
Anglerfish somatostatin I[f]	GGA CAG AGA GAC --- TCC AAA CT C CGC --- CTG CTG CTG --- CAC CGG --- TAC CCG CTG CAG GGC TCC AAA
Anglerfish somatostatin II[f]	GAG CAG AGC GAC AAC CAG GAC CT G GAC CTG GAG CTG CGT --- CAG CAC --- TGG CTG CTG GAG AGA GCC CGG[j]
Anglerfish somatostatin III[g]	GGA CAG AGA GAC --- TCC AAA CT C CGC --- CTG CTG CTG --- CAC CGG --- TAC CCG CTG CAG GGC TCC GGG
Catfish somatostatin 22[h]	GGC --- CGG CCT --- CAT GTG GT T TTG AAT TCG --- --- --- --- --- --- --- --- --- --- GCT CTG GAG
Catfish somatostatin 14[i]	GCG CCG TCT GAT --- GCC AAA CT G CGC CAG TTC --- CTC --- CAG AGG TCG ATT CTT GCA CCG TCT GTA AAA

Row 2

Bovine kininogen	GAA ATC GAC TGC AAC GAC CAG GAT GTA TTT AAA GCT GTG GAC GCT GCT CTG ACA AA A TAC AAC AGT GAA
Human somatocrinin	CTG GGC CAG CTG TCC GCC CGC AAG CTG CTC CAG GAC ATC ATG AGC AGG CAG CAG GG*A GAG AGC AAC CAA
Human somatostatin	CAG*GAA CTG GCC AAG TAC TTC TTG GCA GAG CTG --- CTG TCT GAA CCC AAC CAG AC G GAG --- AAT GAT
Rat somatostatin	CAG*GAA CTG GCC AAG TAC TTC TTG GCA GAA CTG --- CTG TCT GAG CCC AAC CAG AC A GAG --- AAC GAT
Anglerfish somatostatin I	CAG GAC ATG ACT CGC TCC GCC TTG GCC GAG CTG CTC CTG TCG GAC CTC CTG CAG GG G GAG --- AAC GAG
Anglerfish somatostatin II	CAG GAG TGG AGT AAA CGG GCG CTG GAG GAG CTG --- CTG GCT CAG ATG TCT CTG CC A GAG GCC ACG TTC
Anglerfish somatostatin III	CAG GAC ATG ACT CGC TCC GCC TTG GCC GAG CTG CTC CTG TCG GAC CTC CTG CAG GG G GAG --- AAC GAG
Catfish somatostatin 22	GAG GCT CGC AAC GTG CCG TTT GGA GAA GAG --- GTA CCA GAG AGA CTC ACT CTT CC T GAG CTC CAG TGG
Catfish somatostatin 14	CAG GAG CTC ACC AGG TAC ACG CTC CCA GAG CTG CTC GCA GAG CTC GCG CAA GCC GA A AAC GAG --- ---

Row 3

Bovine kininogen	AAC AAG AGT GGC AAC CAG TTT GTA TTG TAC CGC ATA ACC GAG GTC GCC AGA ATG GAT AAT CCT GAC ACA
Human somatocrinin	GAG CGA GGA GCA AGG GCA CGG CTT GGT CGT CAG GTA GAC AGC ATG TGG GCA GAA CAA AAG CAA ATG GAA
Human somatostatin	GCC CTG GAA CCT --- GAA GAT CTG TCC CAG GCT GCT GAG CAG GAT GAA ATG AGG CTT GAG CTG CAG AGA
Rat somatostatin	GCC CTG GAG CCT --- GAG GAT TTG CCC CAG GCA GCT GAG CAG GAC GAG ATG AGG CTG GAG CTG CAG AGG
Anglerfish somatostatin I	GCT CTG GAG GAG GAG AAC TTC CCT CTG GCC GAA GGA GGA CCC GAG GAC GCC CAC GCC GAC CTA GAG CGG
Anglerfish somatostatin II	CAG CGG GAG GCG GAG GAC GCG TCC ATG GCA ACA GAA GGA CGG --- --- --- --- ATG AAC CTA GAG CGG
Anglerfish somatostatin III	GCT CTG GAG GAG GAG AAC TTC CCT CAG GCC AGA AGG GGA ACC CGA GGA CGC CCA CGC CGA CCT AGA GCG
Catfish somatostatin 22	ATG CTC AGT GAG AAC AAC GAG CTC ACG --- CCC GTT CAG GTG GAA GAA --- --- GCC --- --- --- ---
Catfish somatostatin 14	GTG CTG GAC TCG GAC GAG GTG TCT CGC GCC GCC GAA AGC GAG GGC GCG --- CGC CTG GAG ATG GAG CGA

Row 4

Bovine kininogen	TTT TAT TCC TTG AAG TAC CAA ATC AAG GAG GGC GAC *TG T* CCT TTT CAA AGT AAC AAA ACT TGG CAG GAC
Human somatocrinin	TTG GAG AGC ATC CTG GTG GCC CTG CTG CAG AAG CAC AG*C AGG AAC TCC CAG GGA TGA
Human somatostatin	TCT GCT AAC TCA AAC CCG GCT ATG GCA CCC CGA GAA CG C AAA GCT GGC *TGC* AAG AAT TTC TTC TGG AAG
Rat somatostatin	TCT GCC AAC TCG AAC CCA GCC ATG GCA CCC CGG GAA CG C AAA GCT GGC *TGC* AAG AAC TTC TTC TGG AAG
Anglerfish somatostatin I	GCC GCC AGC GGG GGG CCT CTG CTC GCC CCC CGG GAG AG A AAA GCC GGC *TGC* AAG AAC TTC TTC TGG AAA
Anglerfish somatostatin II	TCC GTG GAC TCT ACC AAC AAC CTA CCC CCT CGT GAG CG T AAA GCT GGC *TGT* AAG AAC TTC TAT TGG AAG
Anglerfish somatostatin III	GGC CGC CAG CGG GGG CCT CTG CTC GCC CCC CGG GAG AG A AAG GCC GGT *TGC* AAG AAC TTC TTC TGG AAA
Catfish somatostatin 22	CCT CGC AGC AGG --- --- --- --- CTG GAG CTG GTC AG G AGA GAC AAC*K**TGC* ATG AAC TAC TTC TGG AAG
Catfish somatostatin 14	GCC GCC GGT --- --- CCC ATG CTG GCT CCC CCC GAG CG C AAA GCC GGC *TGC* AAG AAT TTC TTC TGG AAA

(*continued*)

Table 7.22 (Continued)

Row 5

Bovine kininogen	*TGT* GAC TAC AAG GAC TCT GCA CAA GCT GCC ACA GGA GAG *TGC* ACA GCG ACC GTG GCC AAG AGA GGG AAT
Human somatostatin	ACT TTC ACA TCC *TGT* TAG
Rat somatostatin	ACA TTC ACA TCC *TGT* TAG
Anglerfish somatostatin I	ACC TTC ACC TCC *TGC* TGA
Anglerfish somatostatin II	GGC TTC ACT TCC *TGT* TAA
Anglerfish somatostatin III	ACC TTC ACC TCC *TGC* TGA
Catfish somatostatin 22	TCC AGG ACA GCA *TGC* TGA
Catfish somatostatin 14	ACT TTC ACG TCG *TGT* TAA

Row 6

Bovine kininogen	ATG AAG TTC TCC GTG GCT ATC CAG ACC *TGC* CTG ATC ACT CCA GCC GAG GGC CCC GTG GTG ACA GCC CAG

Row 7

Bovine kininogen	TAT GAG *TGC* CTT GGC *TGT* GTG CAT CCC ATA TCT ACC AAG AGC CCC GAC TTG GAG CCT GTT CTG AGA TAT

Row 8

Bovine kininogen	GCC ATC CAA TAT TTT AAC AAC AAC ACC AGT CAT TCC CAC CTC TTT GAT CTG AAA GAA GTA AAA AGA GCC

Row 9

Bovine kininogen	CAA AGA CAG GTG GTG TCT GGA TGG AAC TAT GAA GTT AAT TAC TCA ATT GCA CAA ACT AAT *TGT* TCC AAG

Row 10

Bovine kininogen GAG GAA TTT TCA TTC TTA ACT CCA GAC *TGC* AAG TCC CTT TCA AGT GGT GAT ACT GGT GAA *TGT* ACA GAT

Row 11

Bovine kininogen AAA GCA CAT GTA GAT GTC AAG CTA AGA ATT TCT TCC TTC TCG CAG AAA TGT GAC CTT TAT CCA GTG AAG

Row 12

Bovine kininogen GAT TTT GTA CAA CCA CCC ACC AGG CTT *TGT* GCC GGC *TGC* CCC AAA CCT ATA CCT GTT GAC AGC CCA GAC

Row 13

Bovine kininogen CTG GAG GAG CCT CTG AGC CAT TCC ATC GCA AAG CTT AAT GCA GAG CAT GAT GGA GCC TTC TAT TTC AAG

Row 14

Bovine kininogen ATT GAC ACT GTG AAA AAA GCA ACA GTA CAG GTG GTA GCT GGA TTG AAG TAT TCT ATT GTG TTC ATA GCA

Row 15

Bovine kininogen AGG GAA ACC ACA *TGT* TCT AAG GGA AGT AAT GAA GAG CTG ACC AAG AGT *TGT* GAG ATC AAT ATA CAT GGT

Row 16

Bovine kininogen CAA ATT CTA CAC *TGT* GAT GCT AAT GTC TAT GTG GTG CCT TGG GAG GAA AAA GTT TAC CCT ACT GTC AAC

Row 17

Bradykinin

Bovine kininogen *TGT* CAA CCA CTT GGA CAG ACC TCA CTC ATG AAA AGG CCT CCG GGT TTT TCA CCT TTC CGA TCA GTT CAA

(*continued*)

Table 7.22 (Continued)

Row 18	
Bovine kininogen	GTG ATG AAA ACT GAA GGA AGC ACA ACT ↓ GTA AGT CTA CCC CAC TCT GCC ATG TCA CCT GTA CAA GAT
Row 19	
Bovine kininogen	GAA GAG CGG GAT TCA GGA AAA GAA CAA GGA CCC ACT CAT GGG CAT GGC TGG GAC CAT GGA AAG CAA ATA
Row 20	
Bovine kininogen	AAA TTA CAT GGC CTT GGC CTT GGC CAT AAA CAT AAG CAT GAC CAA GGT CAT GGG CAC CAT GGA AGT CAT
Row 21	
Bovine kininogen	GGT CTT GGC CAT GGA CAT CAA AAG CAA CAT GGT CTT GGC CAT GGA CAT AAG CAT GGT CAT GGC CAC GGA
Row 22	
Bovine kininogen	AAA CAT AAA AAC AAA GGA AAA AAC AAT GGA AAG CAT TAT GAT TGG AGG ACA CCC TAT TTG GCA AGT TCT
Row 23	
Bovine kininogen	TAT GAA GAT AGC ACT ACA TCC TCT GCA CAG ACG CAA GAG AAG ACA GAA GAG ACA ACA CTC TCT TCC CTA
Row 24	
Bovine kininogen	GCC CAG CCA GGT GTA GCC ATT ACC TTT CCT GAC TTT CAG GAC TCA GAT CTC ATT GCA ACT GTG ATG CCT
Row 25	
Bovine kininogen	AAT ACA CTA CCA CCT CAC ACA GAG AGT GAT GAT GAC TGG ATC CCT GAC ATC CAG ACA GAG CCA AAT AGC

Row 26	
Bovine kininogen	CTT GCA TTT AAA TTG ATT TCA GAC TTT CCA GAA ACA ACC TCC CCC AAA *TGT* CCT AGT CGC CCC TGG AAG
Row 27	
Bovine kininogen	CCA GTT AAT GGA GTG AAT CCA ACT GTG GAA ATG AAA GAG TCT CAT GAT TTT GAT CTT GTT GAT GCT CTT
Row 28	
Bovine kininogen	CTT TAA

[a]Codons for basic amino acids are underscored, those for cysteine are italicized, and those for tyrosine are stippled. Asterisk indicates the point of insertion of an intron. Mature genes are partially or completely boxed.
[b]Kitamura *et al.* (1983); Nawa *et al.* (1983).
[c]Gubler *et al.* (1983); Mayo *et al.* (1985).
[d]Shen *et al.* (1982); Shen and Rutter (1984).
[e]Montminy *et al.* (1984); Tavianini *et al.* (1984).
[f]Hobart *et al.* (1980).
[g]Goodman *et al.* (1980).
[h]Magazin *et al.* (1982).
[i]Minth *et al.* (1982).
[j]Insert AGC GCC GGA CTC CTG TCC.
[k]Insert ACG GTG ACG AGC AAA CCA CTC AAC.

Somatocrinin Genes. Since the somatocrinins have only recently been recognized as a distinct class of hormones, it is not surprising that but a single gene sequence has been determined, in this case, from human sources. However, it has been established twice (Gubler *et al.*, 1983; Mayo *et al.*, 1985). Structurally it is more typical than the preceding, bearing a presequence in orthodox fashion. To a degree, this example provides an intermediate step between subclasses III and IV, for the ultimate gene region is located only a short distance from the 5′ terminus (Figure 7.2F). However, since the active product consists of 44 amino acid residues, it actually occupies the middle region of the precursor, with 11 sites preceding and 33 following. The ultimate coding region is preceded by the standard pair of codons for basic amino acids (in this case, both are for arginine), but the presence within the active part of two comparable pairs raises questions as to what other criteria are involved in recognition by the protease during cleavage. Immediately following activation, the somatocrinin molecule is amidized through modification of the glycine residue encoded by the GGT that follows the ultimate gene region proper. The gene is unique and is located in man on chromosome 20 (Mayo *et al.*, 1985).

The Somatostatin Gene Family. The somatostatin gene family is well on its way to becoming one of the most thoroughly explored of cryptomorphic types, for two cistrons from mammalian sources and five from fish have been determined. When additional ones from avian, reptilian, and lower vertebrates have been added to this list, a well-rounded picture of their phylogeny will emerge. Even now some details are clearly discernible, but with the report of the occurrence of several varieties of somatostatinlike products in *Bacillus subtilis* (LeRoith *et al.*, 1985), a need for investigation into possible ancient origins for this hormone is clear. However, here, before those evolutionary aspects can be understood, description of the gene structure and function is an essential prelude. Although principally in the pancreas, stomach, and small intestine, somatostatin occurs abundantly also in the nervous system, particularly the hypothalamus, and may be a neurotransmitting agent. Chiefly its effects are inhibitory, retarding the secretion of other hormones, including somatocrinin, insulin, glucagon, gastrin, and growth hormone (Shen *et al.*, 1982). Characteristically, the primary transcript of the gene has understandably been considered a "prepro"protein, with the inhibitory (pro)sequence exceptionally long (Figure 7.2G) (Shen *et al.*, 1982; Funckes *et al.*, 1983; Montminy *et al.*, 1984; Shen and Rutter, 1984; Tavianini *et al.*, 1984). At least two major forms of the ultimate product are known—somatostatin I, whose sequence is given in six of the seven provided in Table 7.22, and somatostatin II, represented by the second of the anglerfish structures (Hobart *et al.*, 1980). The latter appears to inhibit insulin secretion, but not that of glucagon (Shen *et al.*, 1982). Among its peculiarities of structure is the presence of a six-codon insertion located at the end of row 1.

These two forms of the 14-amino acid peptides just precede the 3′ termini of the gene structures in rows 4 and 5 of the table. However, they encode only the more familiar representatives of this hormone. Often along with this short type, a peptide of 28 amino acid residues may be found in the bloodstream, at least of mammals. This variety is indicated in Table 7.22 as extending to the 5′ end of row 4, and appears to undergo cleavage to produce the shorter type. But whether this duplex type is merely an intermediate product of processing or has activity in its own right has not been sufficiently investigated, although both species have proven to be generated by the same convertase (Gluschankof *et al.*, 1985). Nor is it established whether the double molecule occurs in

anglerfish or other piscine subjects. The catfish somatostatin gene presents additional problems, for its ultimate cistron encodes a product eight residues longer than the rest, consisting of 22 amino acids instead of just 14. Although the last 12 codons of this ultimate gene may be noted to be largely homologous to the remaining examples, there are a number of differences that may affect its activity. Hence, further investigations of its functions in this fish are greatly needed.

The remainder of the mature gene of somatostatin 22 may be readily seen also to differ widely from the rest, especially in length, much of rows 1, 3, and 4 being unrepresented. With this limited material, it is not possible to attempt to establish homology on a site-to-site basis, the arrangement in Table 7.22 being entirely preliminary and suggestive only. Even its 5′ end is not firmly determined, for in the original description the entire sequence, including that of the transit peptide down to the ultimate gene, was considered as a propeptide (Magazin *et al.*, 1982). Obviously extensive phylogenetic changes in structure of this gene have occurred within what is often considered to be a monophyletic class of vertebrates, in spite of evidence to the contrary (see Berg, 1940). While any deductions regarding the evolutionary history are now necessarily tentative, two trends may be noted in the material at hand. First, the entire gene, including the presequence as described shortly, appears to have undergone lengthening with phylogenetic advancement. On this basis, the gene may be expected to be still shorter in the Cyclostomata, if not the Elasmobranchia as well. Hence, among the lower vertebrates, the cistron may be diplomorphic rather than cryptomorphic. The second trend apparently is in the opposite direction, involving a reduction in size of the ultimate product, with hagfishes and relatives having a somatostatin of perhaps 30 amino acid residues.

Among the obvious additions that need to be made to this subclass when nucleotide sequences are more adequately determined is the gastrin gene, whose product stimulates the secretion of the gastric juices (Yoo *et al.*, 1982; Kato *et al.*, 1983). Additional knowledge will most likely indicate their placement here; this includes the cholecystokinins (Gubler *et al.*, 1984; Kuwano *et al.*, 1984; Deschenes *et al.*, 1985) and natriuretic factor synthesized in the cardiac atrium of mammals (Maki *et al.*, 1984; Sonnenberg and Veress, 1984; Zivin *et al.*, 1984), together with cardiodilatin (Kennedy *et al.*, 1984).

7.5.6. Presequences of Subclass IV Cryptomorphs

The presequences of subclass IV cryptomorphic genes add to the impression that has been growing increasingly firm, that other than the presence of an abundance of codons for hydrophobic amino acids, they lack general characteristics of a distinctive nature. Only half of the eight examples of Table 7.23 show a triplet encoding a basic amino acid near the 5′ end. An additional sequence, along with one of the four just mentioned, displays two for arginine more centrally located, and another, that of anglerfish II, possesses a like codon at the extreme 3′ terminus. Although all are, as already indicated, rich in triplets for apolar monomers, the distribution of those codons varies broadly from one sequence to another. In the human somatocrinin representative, all are situated in the 5′ half, whereas in that of the rat, the opposite end is more densely supplied. More frequently, however, the midsection contains the major part, but even the three species from the anglerfish fail to be entirely consistent in this matter.

Their cleavage sites show a similar lack of constancy of structure. Five of the final

Table 7.23
Presequences of Subclass IV Cryptomorphic Genes[a]

Row 1	
Human somatocrinin[b]	*ATG CCA CTC* TGG *GTG TTC TTC TTT GTG ATC CTC* ACC *CTC* --- --- ---
Human somatostatin[c]	*ATG CTG* TCC TGC CGC *CTC* CAG TGC *GCG CTG GCT GCG CTG* --- --- ---
Rat somatostatin[d]	*ATG CTG* TCC TGC CGT *CTC* CAG TGC *GCG CTG GCC GCG CTC* --- --- ---
Anglerfish somatostatin I[e]	*ATG* AAG *ATG GTC* TCC TCC TCG CGC *CTC* CGC TGC *CTC CTC GTG CTC CTG*
Anglerfish somatostatin II[e]	*ATG* CAG TGT *ATC* CGT TGT *CCC GCC ATC TTG GCT CTC CTG GCG TTG GTT*
Anglerfish somatostatin III[f]	*ATG* --- --- *GTC* TCC TCC TCG CGC *CTC* CGC TGC *CTC CTC GTG CTC CTG*
Catfish somatostatin 22[g]	*ATG* TCG TCT --- TCA *CCA CTC* CGT *CTC* --- *GCT CTT GCC CTC ATG* TGC
Catfish somatostatin[h]	*ATG CCC* TCC ACG CGG *ATC* CAG TGC *GCC CTG GCT CTC CTG GCC GTC* ---
Row 2	↓
Human somatocrinin	--- AGC AAC AGC --- --- TCC CAC --- --- TGC TCC *CCA CCT*
Human somatostatin	--- TCC *ATC GTC CTG GCC CTG GGC* TGT *GTC* ACC *GGC GCT CCC*
Rat somatostatin	--- TGC *ATC GTC CTG GCT TTG GGC GGT GTC ACC GGG GCG CCC*
Anglerfish somatostatin I	*CTG* TCC *CTG* ACC *GCC* TCC *ATC* AGC TGC TCC *TTC GCC GGA* CAG
Anglerfish somatostatin II	*CTG* TGC *GGC CCA* AGT *GTT* TCC TCC CAG *CTC* GAC AGA GAG CAG
Anglerfish somatostatin III	*CTG* TCC *CTG* ACC *GTC* TCC *ATC* AGC TGC TCC *TTC GCC GGA* CAG
Catfish somatostatin 22	*CTG* --- *GTC* TCA *GCC GTC* --- *GGT GTC ATA* TCG TGC *GGC* CGG
Catfish somatostatin 14	--- --- *GCG CTC* TCC *GTC* TGC AGC --- *GTC* TCA *GGC GCG CCG*

[a]Codons for apolar (hydrophobic) amino acids are italicized and those for basic ones are underscored. Arrow indicates cleavage site.
[b]Gubler *et al.* (1983); Mayo *et al.* (1985).
[c]Shen *et al.* (1982).
[d]Montminy *et al.* (1984); Tavianini *et al.* (1984).
[e]Hobart *et al.* (1980).
[f]Goodman *et al.* (1980).
[g]Magazin *et al.* (1982).
[h]Minth *et al.* (1982).

codons bordering that point are for the small apolar amino acids glycine and alanine, one is for a small polar uncharged unit (serine), another encodes the large and charged arginine, and the last one specifies the large, sulfur-bearing cysteine. Moreover, the penultimate codons have equally wide ranges in the amino acids they encode. On the downstream side of the cleavage line much greater consistency is displayed, for all the triplets designate apolar types, except that of the distinctive structure of anglerfish somatostatin II (Hobart *et al.*, 1980), which codes for aspartic acid. Signals for proline are most frequently found in the second site of the mature coding sector, although the catfish somatostatin 22 gene has one for arginine, and all three of the anglerfish encode glutamine. Thus, as before, the combination of amino acids actually recognized by the enzyme that removes the transit peptide remains obscure.

7.6. STRUCTURE OF SUBCLASS V CRYPTOMORPHIC GENES

The members of the fifth and last subclass of cryptomorphic genes are structurally strongly in contrast to all the others. Instead of the mature coding region being extensive but encoding only one or more ultimate products of small size, here the final proteins occupy nearly the entire precursor. That is, a precursorial transcript, usually bearing a transit peptide, is produced, which typically is cleaved into two or three "subunits" that are the actual functional proteins. Thus the only unproductive regions of the mature gene are the short segments that bear the signals for the cleaving enzymes. Because the ultimate genes consequently are large molecules, whose complete sequences contribute little to a basic understanding of gene structure and nature, only the portions with a greater contribution to make toward fuller appreciation of the fundamentals of subclass V cistrons are given, along with diagrams wherever these are helpful toward this same goal.

7.6.1. Genes of Certain Types of Complement

In vertebrates and other metazoans is a large group of genes known as the major histocompatibility complex, whose products are involved in immune reactions of various sorts (Steinmetz and Hood, 1983; Kaufman *et al.*, 1984). Of the three classes into which the members are grouped, those of class III, which encode ingredients of "complement," are most frequently representatives of subclass V cryptomorphs, although several species from class II may also prove to be. In addition, a small number from classes I and II show indications of the type of complexity of organization that is treated in the next chapter, which deals with "assembled" genes.

The proteins of complement are secreted into the bloodstream, where they enter into an intricate cascade of interactions, which ultimately result in lysis of an invading cell (Dillon, 1983, pp. 382–384). In the principal chain of processes, called the classical pathway, nine major substances are involved, referred to as C1–C9; to this number two additional types, B and D, must be added, which are active in the second, or alternative, pathway. To judge from the manner in which the molecules are fractured during the interactions, C1–C5, C9, and B are encoded by genes that belong to the present subclass of cryptomorphs, but not all of these have been fully sequenced. Another deficiency in many relevant reports is the frequent failure to show the presequence that theoretically should be present on these secreted products. Nor are the other articles consistently clear as to the precise location of the cleavage points, timing of the processing, or functions of the resulting fractions. What has been adequately documented, nevertheless, provides sufficiently deep insight into the structure and activities of this more than usually interesting set of cistrons. Two that have been explored most thoroughly, C3 and C4, are presented first to set the stage for others also available. The gene for C5 has now had its sequence established, but it is not included, since it adds little to the total picture presented here (Lundwall *et al.*, 1985).

The Mature Gene of Factor C3. Complement component C3 is undoubtedly the best known of the entire interacting chain, for, besides being the most abundant member, it plays critical roles in both the classic and alternative pathways. The gene consists of two primary parts, a presequence of 72 nucleotide residues and a mature coding sequence of ~2940 (Figure 7.3); immediately after translation the protein encoded by the latter is

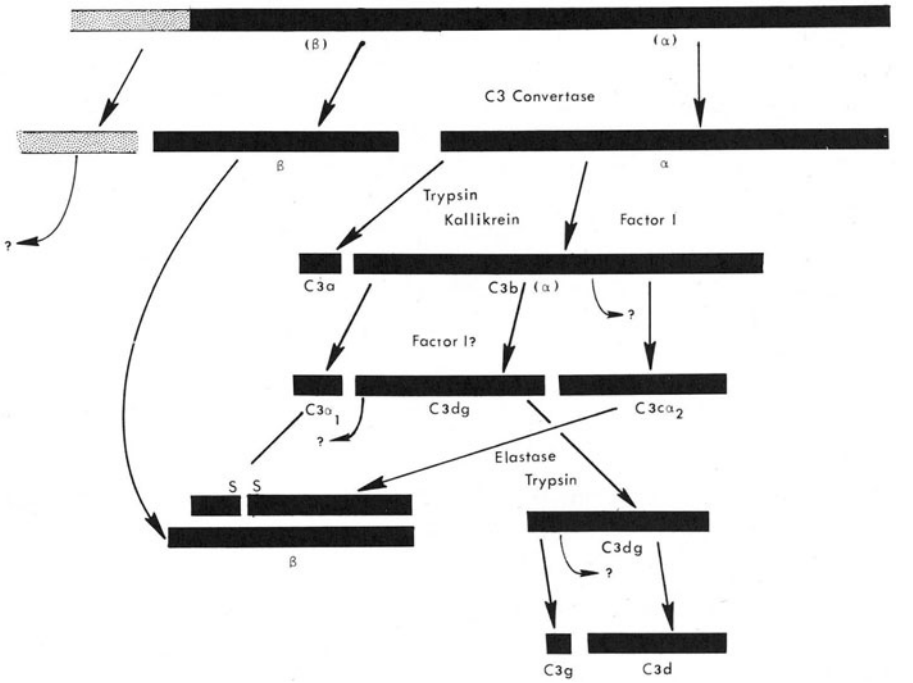


Figure 7.3. The intricately interacting behavior of complement C3. The gene structure is represented by the topmost diagram. (Based on de Bruijn and Fey, 1985.)

inert, becoming activated in the presence of foreign cells by action of either of two enzymes called convertases, one in each major pathway (de Bruijn and Fey, 1985). Following cleavage and removal of the transit peptide enzymatically into β and α chains, the latter of these ultimate products (Fey *et al.*, 1984; Wetzel *et al.*, 1984) undergoes multiple fractioning to carry out various distinctive functions, while the β remains intact (Figure 7.3). The first protease, C3 convertase, splits off a small segment, C3a, leaving the greater part as C3b, which is highly reactive. In one set of reactions, it unites with a dual combination of complement constituents C2a and C4b to form C2a·C4b·C3b. This product then may continue through the classical cascade of complement activity. Or, in the alternative pathway, C3b may unite with a different double molecule to produce pro-Bb·D·C3b to participate in the eventual lysis of the target cell, largely by acting on C5. Furthermore, it is capable of opsonizing bacteria, that is, making those organisms more susceptible to destruction by phagocytes. The small C3a sector of 78 amino acids split off by the convertase serves as an anaphylatoxin that principally binds to receptors on mast cells (macrophages) to induce the release of histamine and other factors stored in the granules of those cells. But C3b has other immune-related activities (Weigle *et al.*, 1983). The downstream portion of that chain may be acted upon by trypsin to release a particle called $C3c\alpha_1$ (or kallikrein may carry out a similar action nearby), which remains attached to the B subunit. Still farther downstream, the molecule may be cleaved by factor I at two

nearly adjacent points to produce both a median sector known as C3dg and a carboxyl terminal portion $C3c\alpha_2$, which is similar to $C3c\alpha_1$ in bearing a carbohydrate. The latter product is joined by a disulfide bridge to C3dg and joins it in forming a dimer with the β chain. In addition, some of these particles may be united by covalent bonds to C3b to carry out discrete immune reactions. Subsequently C3dg is further reduced by twin actions of elastase and trypsin to produce a short peptide C3g and a longer one C3d (Figure 7.3). Thus the seemingly simple apparent bipartite mature coding region is actually multifold, encoding a diversity of ultimate products.

The Mature Gene of Factor C4. The fourth component of complement, C4, is encoded in man by two separate but closely linked loci on chromosome 6 (Carroll *et al.*, 1984); polymorphism is rank, since gene *C4A* has 13 known alleles and *C4B* has 22 (Mauff *et al.*, 1983). It is synthesized in macrophages as well as in the liver as a single chain of molecular weight 200,000. Although the actual start sites of translation and transcription alike do not seem to have been established, so that the full length of the presequence is unknown, one is present nevertheless (Belt *et al.*, 1984). The mature coding region encodes three subunits, β, α, and γ, in that order $5' \rightarrow 3'$ (Schreffler *et al.*, 1984). These combine into the mature but inert protein as a simple αβγ trimer within the cell, the mechanism for its passage through the cytoplasmic membrane into the bloodstream remaining unexplored.

Activation is induced by a subparticle of $C\bar{1}\bar{s}$, which releases the peptide C4a from the NH_2 end of the α chain, a substance apparently serving as an anaphylotoxin along with C3a. Later, factor I may further cleave this same chain, releasing the peptide C4d from the carboxyl half, an activity which results in the inactivation of the remaining C4 protein.

Both C3 and C4 show a limited degree of homology with α_2-macroglobulin, one of the plasma proteins, so it has been claimed that the three have had a common evolutionary origin (Sottrup-Jensen *et al.*, 1985). About two-thirds of C3 was shown to be similar in structure to the macroglobulin at a level between 19 and 31% homology, whereas C4 had only a comparatively short segment that displayed such a relationship. Since the overall level of identity thus was low, it may be that, rather than having common origin, the three have been exposed to similar genetic influences at the molecular level, as proposed for a certain pea nuclear protein in the preceding chapter.

7.6.2. Miscellaneous Representatives of Subclass V Cryptomorphs

Currently a small number of important proteins from a diversity of sources are known that constitute this subclass V of cryptomorphic genes, but the variety that these few established types displays strongly intimates that eventually it will prove to be a large group embracing numerous proteins of exceptional functional significance. As a whole, the structural complexities of those whose gene sequences have been determined have not been fully appreciated, for certain of the representatives undergo multifold stages of processing reminiscent of those of complement factors C3 and C4. Why they do so can scarcely be imagined at the moment. Why, for a case in point, should something as seemingly inert as a seed-storage protein need to be exposed to multistage processing? If nothing else, their structural ramifications certainly indicate that deeper investigations into their functions would doubtlessly be most profitable. But first continuity is best served if

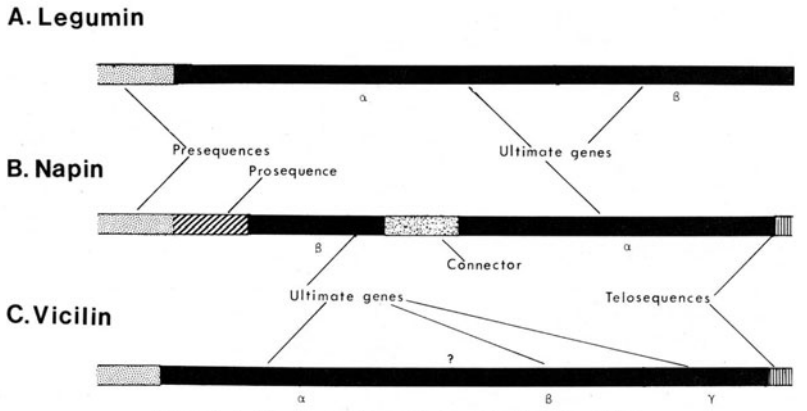


Figure 7.4. Cryptomorphic genes for seed-storage proteins.

some blood factors of vertebrates are examined, since their activities are in a cascading fashion much like that of complement.

Blood Factor X. Only a single factor from the blood-clotting cascade seems to have had its gene sequence established—and even that is incomplete at the 5′ end (Leytus *et al.*, 1984). This solitary representative, for factor X, then must be taken to exemplify others, such as the blood factor IX, whose processing events are similar. This protein is encoded by a gene whose mature coding section is 1338 nucleotide residues in length. Whether that portion is preceded by a presequence has not been established, but a strong likelihood exists that it is. As shown in Figure 7.4, the mature gene provides for two subunits, referred to as light and heavy chains, that are separated by three codons in the series arginine, lysine, arginine, which signal the cleavage point. After translation and cleavage by an undetermined enzyme, the two subunits unite to form the mature inert protein, which is an αβ dimer. Activation of the factor involves cleavage by either factor IXa or VIIa, the proteolytic action removing the NH_2 end of the heavy chain, a region of about 50 amino acids (Leytus *et al.*, 1984). This reduced dimer, then known as factor Xa, converts prothrombin to thrombin in the presence of various ingredients to begin active blood-clot formation. In addition to the several cleavages by proteases, the mature product is complicated by addition of two carbohydrate moieties to the NH_2-terminal piece of the heavy chain and by conversion of 11 and 12 glutamate residues of the light and heavy chains, respectively, to γ-carboxyglutamic acid, in which process vitamin K plays an important role.

Among other blood-clotting factors whose gene structure may fall into this category is bovine protein C (Long *et al.*, 1984b), but not all vertebrate members of the subclass are of that nature. The important hormone insulin secreted by pancreatic β cells is the product of a gene whose mature coding region encodes B, C, and A chains, the C component being of unknown function. The completed hormone is in the form of the hexamer A_3B_3 (Hahn *et al.*, 1983). By coincidence, the gene encoding the insulin-receptor protein also is a member of the present subclass (Ebina *et al.*, 1985).

Seed-Storage Proteins. Unlike the simple diplomorphic examples seen at the beginning of this chapter, several types of seed-storage proteins are encoded by subclass V cryptomorphic genes and thus undergo varying degrees of posttranslational cleavages. By far the most straightforward representatives of the group are the several legumin genes from the pea, two examples of which (*legA* and *legD*) have been fully sequenced (Lycett *et al.*, 1984; Bown *et al.*, 1985). After a fairly typical presequence of 21-codon length, there is a long mature coding region, interrupted by three introns. This consists of a 996-nucleotide sequence for the α subunit at the 5′ end and one of 756 nucleotides encoding the β subunit, without any connecting elements between them, nor does there seem to be a telosequence (Figure 7.4A). Nothing has been reported as to the processing of the nascent product of translation. Homologs also are present in wheat and other grains (Robert *et al.*, 1985).

The legumin gene is unusual in having the larger subunit encoded in the 5′ portion; however, another member of this class of proteins, napin from rape (*Brassica napus*), has the more frequent arrangement. As in the foregoing, it begins with a presequence of 21 codons, but it differs in having a prosequence that codes for 17 amino acids (Crouch *et al.*, 1983). The first ultimate gene that ensues, for the β subunit, is only 36 codons in length; it is separated from the coding sequence for the α subunit by a connector of 20 codons (Figure 7.4B). At neither end of this region is there a recognizable cleavage signal, the 5′ terminus bearing triplets for proline, asparagine, and tryptophan, and the 3′ having those for proline, glutamine, and glycine, but proline may prove to be the principal element. The ultimate coding sector for the α subunit consists of 81 codons, followed by a telosequence consisting of triplets for proline, serine, and tyrosine, before the TAG translation-terminal signal. Napins constitute ~20% of the total rape seed protein and are broken down rapidly during germination. The holoenzyme, an αβ dimer of molecular weight 13,000, is highly basic, largely as the result of the high percentage (~25%) of glutamine residues it contains.

In a third member of the group, the level of complexity of the above example is equalled, but by a different means; a fourth, for ricin of castor bean, whose sequence has also been established recently (Lamb *et al.*, 1985), does not merit additional attention. Another current addition is that for glycinin, a complex product of soybean (Momma *et al.*, 1985a,b). Here in a gene for vicilin, 11 copies of which exist in the pea genome (Domoney and Casey, 1985), no inhibitory sequence intervenes between the transit peptide and mature coding sectors, the latter following the former directly (Lycett *et al.*, 1983). Moreover, the mature portion resembles that of the legumin gene in having the large subunit (α) region at the 5′ end (Figure 7.4C). This sector then leads into the β portion without any connector element, a condition repeated at its close, where the γ-coding part begins. This sector of 94 triplets is adjoined by a telosequence of unusual length, since it consists of 12 condons, including only one each for basic and acidic amino acids. However, the real complexity of the present structure lies in the apparent multifold posttranslational processing. Although details are not firmly established, it is reported that cleavage may in some cases not occur at the α–β juncture, and in others may not take place at any site, except perhaps the γ-telosequence point of contact (Lycett *et al.*, 1983). As a consequence, it is not clear whether vicilin is ever in the form of an αβγ trimeric protein, nor is it established how the precursor is processed.

7.7. COMPARISONS OF THE SEVERAL CLASSES OF COMPLEX GENES

Because of the numerous subdivisions needed to embrace the various types of diplomorphic and cryptomorphic genes and their products, a synopsis of their principal features seems desirable. Three major types, simple, compound, and complex, are thus readily compared and their subdivisions more easily comprehended.

1. *Simple* genes encode single products that are ready for employment immediately following processing or translation.
2. *Compound* genes, such as those of many viruses (Chapter 10), encode two or more products that are useful in the cell after processing has separated them, that is, no latent period exists.
3. *Complex* genes encode two or more products, which remain intact during a latent period of varying length. Two principal classes are recognized:

   A. *Diplomorphic* genes encode either two or three different peptides, one of which is a transit (signal) sequence useful in passage through membranes. Two varieties exist:
      Simple diplomorphs, in addition to the active product, encode a transit peptide (presequence) which is removed from the product upon transport through a membrane.
      Complex diplomorphs also encode a transit peptide, but over and above provide for an inhibitor sequence (prosequence) which after being translated remains attached to the principal gene product until the latter is needed in the cell.
   B. *Cryptomorphic* genes similarly encode a presequence, but usually lack a propeptide coding region. The ultimate gene product or products are coded within sectors from which they must be cleaved before they can be active. A highly varied assembly, they fall into at least the five following subclasses:
      Subclass I. In this category is placed those genes that encode multiple copies of the same or virtually the same protein, such as those that provide for the mating factors (α-pheromones) and glucoamylase (Yamashita *et al.*, 1985) in yeast or the enkephalins in vertebrates.
      Subclass II. This subclass embraces genes that contain ultimate coding regions for two or more different products which remain intact within a primary translational product for a more or less prolonged latent period.
      Subclass III. Besides the presequence, subclass III cryptomorphs contain an ultimate gene for a single product, often arranged as a prosequence, but always in the 5′ position of the original translational coding area.
      Subclass IV. The member genes resemble those of subclass III in encoding a single ultimate product, but they differ in having the coding sector for the principal protein in the 3′ portion. Often a sequence for an inhibitor peptide may be present, so that they approach the complex diplomorphs structurally.
      Subclass V. The genes that constitute this subclass each encodes a precursorial translational product, typically bearing a transit peptide, comprised of two or more so-called subunits (actually functional proteins) so

that most of the coding region results in active products. Complement components of the major histocompatibility complex provide the clearest examples of this category.

It is to be anticipated that additional subclasses of cryptomorphic genes will come to light as explorations into the coding structures of the green plants, seaweeds, invertebrates, and protistans continue, for the hormones and enzymic secretions of all living things need the protective control that this device affords. In this class, as among the simpler diplomorphs, control is exercised by successive cleavages of large molecules into smaller parts. In the next chapter, the opposite convention is followed, lesser components being combined into larger ones, providing for diversity in the product rather than protection.