Chapter 28

DIFFERENTIAL GENE ACTIVITY IN COTTON EMBRYOGENESIS

Glenn A. Galau
University of Georgia
Athens, Georgia

INTRODUCTION

Attempts have been made to explain the process of cell differentiation as a consequence of the gene products which are synthesized in differentiating cells. The foundations of the Variable Gene Activity Theory of Cell Differentiation may be found in the writings of Stedman and Stedman (1950), Mirsky (1951, 1953) and Sonneborn (1950). Several beliefs are crucial to the theory. The most important is that a cell's current differentiated state is a consequence of the history of the gene products, the proteins, it has synthesized and that different states then imply different such histories. In addition, all cells contain the same complements of DNA. If there are a large number of differentiated states, then much of the gene complement in the DNA must not be expressed in any particular cell. Thus, the conclusion is that differential control of gene expression must ultimately control cell differentiation.

In the succeeding years the process of gene expression has been increasingly defined, yet in its broad outline the theory remains robust. There are now recognized to be many separate steps in the expression of a gene into the protein for which it codes. These include: (1) the transcription of a messenger RNA (mRNA) precursor from the gene sequence; (2) the processing of that primary mRNA transcript into a mature mRNA; (3) its export into the cytoplasm; and (4) the translation of the mRNA into a polypeptide. Many translation products are further assembled into functional complexes and some, as well, undergo co-translational or post-translational modifications such as cleavage or glycosylation. The direct result of gene expression is the current concentration of its protein products. A polypeptide's concentration is ultimately determined by the concentration of its cytoplasmic mRNA, the efficiency with which the mRNA is translated into the polypeptide (together these determine the synthesis rate of the polypeptide), and the intrinsic degradation rate of the polypeptide. The focus of much of molecular biology has been to understand how these processes are regulated and what are the consequences of this regulation for the cell.

As might be expected, differential gene-specific regulation of gene expression has been demonstrated to occur at every step, or level, in gene expression (Tobin,

1979; Davidson and Britten, 1979; Darnell, 1983). It is now clear that for any particular gene it is not possible to predict with any certainty at what level, if any, control in its expression may be exerted. Furthermore, control at more than one level may not be excluded *a priori*. Several general statements may be made, however. Although there are instances of changes in the translational efficiency of particular mRNAs (called translational level control), or changes in the processing or degradation rates of the polypeptide products (post-translational level control), the concentration of the polypeptide is usually intimately related to the concentration of its cytoplasmic mRNA template. The mRNA concentration in turn is determined by the rate of its entry into the cytoplasm (which is a consequence of the gene's transcription rate, the number of genes in the gene family being transcribed and the efficiency of processing and export of the transcripts) and its own intrinsic degradation rate. Here as well there are documented examples of gene-specific control being exerted at all of the levels which determine mRNA concentration, and in some cases at more than one level for a particular gene. Historically, the bias has been that the primary control of gene expression is most often probably at the transcriptional level, and, in fact, many studies with individual genes have tended to confirm this notion (Derman *et al.*, 1981; Darnell, 1983). However, there are several experimental results which suggest that changes in the processing efficiency of constitutively synthesized primary transcripts may be the basis of differential expression of many other genes (Wold *et al.*, 1978; Kamalay and Goldberg, 1984). So, unfortunately, it is still unclear to what extent transcriptional control is responsible for differential gene activity.

A rich mixture of stratagems, including in some cases limited rearrangement of particular DNA sequences, are used by organisms in order to 'remain in control' of their development. It is the task of those interested in differential gene expression then to render a comprehensive description of the important gene products, the cues responsible for their control, and how these cues are utilized. One hopes that these are ultimately responsible for, and not merely the reflection of, the events which drive development.

The growing cotton embryo has been used in several laboratories as a model system in which to study the control of gene activity as it relates to development. Earlier studies highlighted embryogenic events which were thought to be important for subsequent germination (Ihle and Dure, 1972; Dure, 1975; Choinski *et al.*, 1981). A substantial body of work has described the cotton genome and its organization and evolution (Walbot and Dure, 1976; Wilson *et al.*, 1976; Geever, 1980), the embryo mRNA content and its polyadenylation (Harris and Dure, 1978; Galau *et al.*, 1981) and the number of different genes being expressed as mRNAs during germination (Galau *et al.*, 1981). This background and the continually improving techniques of molecular cloning and protein and nucleic acid analysis have allowed a fairly detailed description of global gene activity in embryogenesis. The regulation and significance of some of these events are now being pursued. This review summarizes the rationale used in these investigations, the results obtained so far and the directions likely to be used in subsequent work.

ANALYSIS OF GLOBAL GENE ACTIVITY IN EMBRYOGENESIS

The initial approach was to plead ignorance about what happens in embryogenesis and, thus, to simply look with the widest possible field of vision and the highest resolution available. The preliminary results of this descriptive phase have been collated (Dure *et al.*, 1981; Galau and Dure, 1981), but it is essentially an ongoing project, mapping the temporal aspects of different gene activities in finer detail and better accuracy as the tools are developed. Cotyledons from several stages of normal embryo development of *G. hirsutum* L. cv. Coker 201 have received the most attention. These include embryos of about 50 mg wet weight (about 27 days postanthesis [DPA]), embryos of about 100 mg weight (about 40 DPA), mature embryos (about 55 DPA) and embryos germinated 24 hours. These stages of embryogenesis correspond to cell division and maturation respectively. In addition, an experimental developmental system has been examined. It relies on the ability of excised young embryos to precociously germinate (Ihle and Dure, 1969; Choinski *et al.*, 1981). Incubation of such embryos in abscisic acid reversibly inhibits germination (Ihle and Dure, 1970) and is thought to perhaps mimic the environment of the embryo *in ovulo* during late embryogenesis (Davis and Addicott, 1972; Rodgers, 1980b; Choinski *et al.*, 1981).

A major analytical tool to study abundant polypeptides has been 2-dimensional polyacrylamide gel electrophoresis (O'Farrell, 1975). The protein sample is first electrophoresed in a cylindrical gel containing a pH gradient where polypeptides migrate to and stop at their isoelectric point. The gel cylinder is then mounted on a slab gel and the polypeptides electrophoresed out of the cylinder and fractionated in the slab primarily on the basis of their molecular weight in sodium dodecyl sulfate (Laemmli, 1970). Both dimensions maintain the polypeptides in their denatured state, and together are capable of resolving up to 1,000 polypeptides as individual spots, depending on the concentration distribution of the polypeptides in the sample and the methods used for their detection. Only several hundred polypeptides are usually detected in cotton embryo protein samples with this system under conditions where the highest resolution is achieved (Figure 4, an example).

This technique has been applied to the analysis of the: (1) extant polypeptides, detected by Coomassie or silver stain (Oakley *et al.*, 1980); (2) polypeptides which become radioactive when excised embryos are incubated 3-6 hours in radioactive amino acids, detected by fluorography (Lasky and Mills, 1975); and (3) polypeptides which are synthesized *in vitro* from radioactive amino acids in the wheat germ translation system programmed with embryo RNA, again with detection by fluorography. The identity and concentration of the resolved extant polypeptides is a measure of the past and current gene activity, as outlined above. On the other hand, the radioactivity in each polypeptide synthesized *in vivo* approximates a measure of the synthesis rate for each, since newly synthesized

radioactive polypeptides should not turn over significantly in the short incubation time in the radioisotope. In short, it is a measure of which mRNAs are being translated by the embryos, and, if the translational efficiencies of all the mRNAs are similar *in vivo*, the concentration of these active mRNAs as well. The relative radio-specific activities of different polypeptides are a function of their relative synthesis and degradation rate constants. In the third type of analysis the identity of translatable mRNAs and their concentration are assayed directly by their translation *in vitro* and subsequent electrophoresis of their radioactive polypeptide products. In practice, shorter mRNAs appear to be preferentially translated, due to protease and nuclease activities in the preparations, and premature termination of translation may lead to artifactual spots (Dure and Galau, 1981). Direct comparisons of *in vitro*-synthesized polypeptides with those extant or synthesized *in vivo* may also be confounded if the translation products are normally processed *in vivo* to such an extent that the electrophoretic behavior is changed (as in the case of the storage proteins described below). These shortcomings are usually minor enough to allow the identification of many of the *in vitro*-synthesized polypeptides with those synthesized *in vivo*. Comparisons of the two samples can then give an estimate of the relative translational efficiencies of the mRNAs *in vivo*. This comparison asks to what extent are the available mRNAs actually being used by the embryo.

From these sorts of experiments we have no evidence for widespread gene-specific translational level control of gene expression in cotton embryogenesis. To a first approximation, if the embryo mRNA is detectable by its translation *in vitro*, the embryos are found to be translating the mRNA *in vivo*. This generalization should be tempered due to the aforementioned experimental uncertainties of the *in vitro* translation system and our limited abilities to identify unambiguously the same polypeptides in different gels containing different protein samples. However, it appears that the embryo relies on modulation of its mRNA concentration rather than differential translation of its mRNAs, at least for its abundant transcripts. How it does so, by transcriptional or post-transcriptional controls, remains unknown by these methods of analysis.

It was earlier inferred from biosynthesis inhibitor studies that the synthesis of several particular mRNAs occurred in late embryos but were not utilized until early germination (Ihle and Dure, 1972; Dure, 1975). Some of these observations have not been subsequently confirmed using similar techniques and have been critized on procedural grounds (Smith *et al.*, 1974; Radin and Trelease, 1976; Choinski *et al.*, 1981). (See Chapter 29 for additional discussion.) The enzymes encoded by these mRNAs are probably not abundant enough to be detectable on two dimensional gels, and their mRNAs have not been directly assayed at any stage, so these conclusions remain in doubt. The present results do suggest, however, that, at least for abundant mRNAs, synthesis of nontranslated mRNAs in embryogenesis is probably not of general occurrence.

From the analysis of these protein populations, seven major subsets of the

embryo mRNA complement can be operationally defined by their different temporal expression. The time course of major expression is shown diagrammatically in Figure 1. Subset 1 has at least 36 members and is by and large constitutive

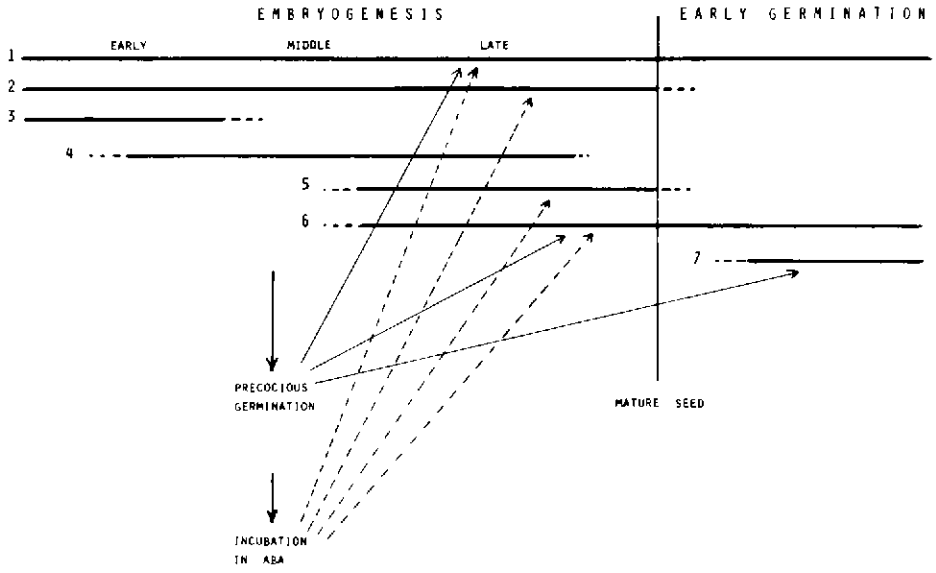

Figure 1. Representation of the time span in which the major subsets of developmentally regulated cotton mRNAs are abundant, as judged from *in vivo* and *in vitro* protein synthesis. (From Dure *et al.*, 1983b).

throughout the time period studied. At least 6 mRNA species comprise Subset 2, which is constitutive to embryogenesis only. They are found in the mRNA (by *in vitro* translation) in mature embryos, but in greatly reduced amounts at 12 hours of germination, and they fail to be detectably translated *in vivo* during the first 6 hours of germination. Thus, these mRNAs must be rapidly degraded in the first hours of germination. Subset 3 is unique to small embryos, the member mRNAs are no longer detected in embryos of about 40 mg wet weight. The major storage proteins and at least four soluble polypeptides make up Subset 4. Their mRNAs are abundant during most of embryogenesis, but not so during desiccation and germination. The mRNAs of Subset 5 and 6 both increase to detectable levels in embryos of about 100 mg wet weight. More than 14 polypeptides comprise Subset 5. Like those of Subset 2, their mRNAs are detectable in the mature seed but are degraded rapidly in the first hours of germination. Subset 6 mRNAs contain at least 5 species and are additionally present throughout early germination. Finally, there are several germination-specific mRNAs in Subset 7 which are quite abundant by 24 hours of germination.

Looking at the polypeptides and mRNAs in excised young embryos by the same techniques has been very illuminating as regards the possible cues which are responsible for the modulation in mRNA concentrations seen in normal embryogenesis. Whether precocious germination proceeds or is arrested by abscisic acid, excision appears to specifically reduce the mRNA concentrations of Subset 4 members and to induce Subset 6 member mRNAs. In addition, under conditions leading to precocious germination, Subset 2 mRNAs disappear and Subset 7 mRNAs accumulate. By this analysis of gene activity, such excised embryos completely bypass late embryogenesis and maturation. They never synthesize Subset 5 mRNAs in their development. Alternative incubation of excised embryos in abscisic acid produces quite another result. Only the mRNAs of Subset 5 are specifically modulated, increasing to very high levels. The observed induction of Subset 6 mRNAs and the decline in Subset 4 mRNAs occur just as they do in precociously germinating embryos. Subset 2 is maintained in abscisic acid, just as in normal late embryogenesis. Under these conditions the embryo appears to bypass much of its normal intermediate growth and proceeds directly to a late maturation stage and reversible dormancy. These observations of the excised embryo system have suggested that Subset 4 is maternally maintained, Subset 6 is maternally repressed and Subsets 2 and 7 are germination repressed and germination induced, respectively. The expression of Subset 5 mRNAs appears to be abscisic acid induced. These tentative designations, of course, are derived primarily from an experimental system and may not necessarily relate to the control of expression in normal embryogenesis. The experimental system does, however, allow manipulation of the expression of several mRNA subsets and should prove useful in further evaluating the control of normal expression.

Although providing a detailed picture of fundamental changes in gene expression, the foregoing experiments only looked at abundant polypeptides and their mRNAs. Furthermore, they were limited to those abundant polypeptides with isoelectric points between pH 4.2 and 8.2 and which, in addition, could be identified reliably in the series of gels containing different protein samples. A further study (Galau and Dure, 1981) was thus conducted using a technique which followed the changes in concentration of these and most of the other 25,000-30,000 different gene transcripts which are present during this time period. This involved the molecular hybridization of polyadenylated mRNAs, isolated from each of the principle stages of normal development, with complementary DNA (cDNA) synthesized from these mRNAs *in vitro*. The extent and rate of hybridization of the cDNAs with their parental mRNA populations can be used to calculate the number of different mRNA sequences and their abundance in the populations (Bishop *et al.*, 1974). By additionally hybridizing each cDNA population with the other stage RNAs (Ryffel and McCarthy, 1975; Levy and McCarthy, 1975) the abundance of many groups of mRNAs could be followed throughout development. While this study used for technical reasons only the polyadenylated mRNAs, an earlier study (Galau *et al.*, 1981) demonstrated that

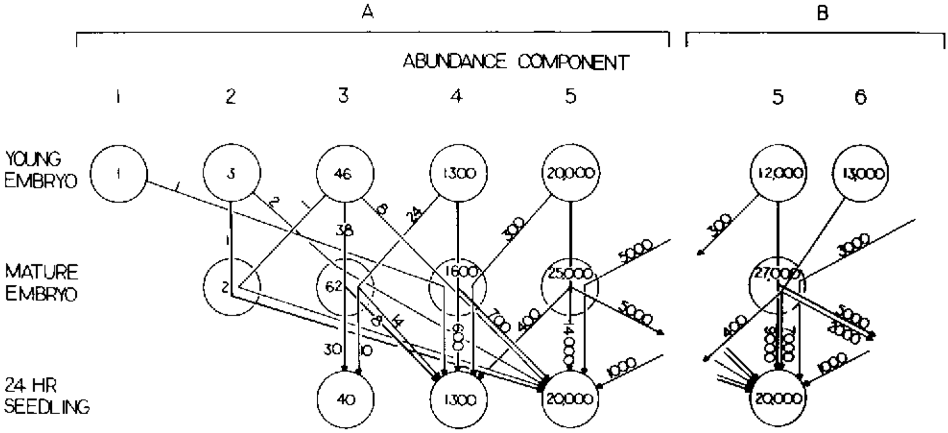

Figure 2. Changes in the concentration of major mRNA subsets with development as judged by cDNA-mRNA hybridization. The mRNA abundance classes are indicated by the circles. The number within each circle is the total number of different mRNA sequences calculated to be in each class. The change in concentration of individual groups of mRNAs are indicated by the arrows, with the associated numbers of different mRNA sequences in each group. The abundance component and (in parenthesis) the approximate percent of the total mRNA mass each sequence comprises is 1(30%), 2(4%), 3(0.4%), 4(0.02%), 5(0.002%) and 6(0.0004%). (From Galau and Dure, 1981).

polyadenylated mRNAs would be representative of all mRNA sequences, since all the sequences present in the total mRNA population are shared in the polyadenylated subset at about the same relative concentration.

A summary of the results of these studies is shown in Figure 2. At each of the three developmental stages studied in detail the frequency distribution of the individual mRNA sequences in the population is shown diagrammatically. Circles represent the different mRNAs which together comprise an abundance group, each of the different mRNA sequences in a group having about the same concentration. The numbers inside each circle indicate the number of different mRNA species in each. For example, the most abundant mRNA is detected in young embryos where it makes up 30 percent of the total mRNA, whereas the thousands of individual mRNA species in abundance component 5 each make up about 0.002 percent of the total mRNA mass. The arrows trace the detected change in concentration of particular groups of mRNAs during development and again their associated numbers indicate the number of different mRNAs in each group. For instance, the most abundant mRNA in component 1 in young embryos, the storage protein mRNA (see below), changes in concentration about 1000-fold during desiccation to form the mature embryo but changes very little

in concentration during early germination. The data were not sufficient to distinguish very well the number and concentration changes of the mRNAs in component 5. An alternative model, consistent with the data, is presented for component 5 in Panel B of the Figure.

This approach provided a very detailed picture of large scale changes in many thousands of different gene transcripts during the latter half of embryogenesis and in early germination. At least 17 groups of mRNAs were detected, based on their absolute abundance and the change in their abundance with time. If groups which share the same pattern of change (regardless of absolute abundance) are lumped together, then still at least 11 such subsets are detected. Two intermediate developmental stages (110 mg embryos and 12 hour seedlings) were also examined in less detail, but showed that the modulation is even more complex than indicated in Figure 2. Sequences in abundance components 1, 2 and 3 should have been detected in the two dimensional gel electrophoresis of the proteins in the first study (Dure *et al.*, 1981), and, in fact, the same subsets and the numbers of mRNAs in each were detected here as was predicted from that study. In addition to confirming by another technique the results of the protein analysis, the hybridization studies showed that gene specific changes in concentration occur in the less abundant mRNAs as well. Thus, it appears that there are several independent events, including developmental cues, which modulate the level of particular mRNAs. The hybridization studies did not address the question of whether or not these transcripts were actually being used in embryogenesis, though it is known that at least some of these sequences are on polysomes at 24 hr of germination (Galau *et al.*, 1981).

The events in cotton are similar to those seen in soybean embryogenesis (Goldberg *et al.*, 1981a). In fact, the number of cotton genes expressed and their modulation in concentration is not unusual when compared with gene expression in other plants (tabulated in Galau and Dure, 1981) and animals (reviewed by Lewin, 1980).

Faced with such an enormous complexity of developmental changes in gene expression, we have concentrated on a further analysis of the regulation of the mRNAs in several of the abundant subsets which were detected by protein electrophoresis and cDNA-mRNA hybridization. Hopefully, defining their regulation will be of importance in understanding gene expression in general and the larger events which occur in embryogenesis. It is expected that if the different members of each subset are in fact coordinately regulated, then these genes will share control sequences for their common regulation, be it for transcriptional or post-transcriptional regulation. These should be common to all members of the co-regulated set and found only in these members (Davidson and Britten, 1979; Davidson *et al.*, 1983). The current aims are then to isolate recombinant DNA probes for these mRNAs and to use them to define more accurately those genes which are indeed coordinately regulated. These probes will be used to isolate the genes, with subsequent sequencing to deduce putative control regions according

to this logic. Complementary work is geared towards understanding what signals regulate their expression and defining the function of these genes.

STORAGE PROTEIN mRNAs

The major high molecular weight storage protein polypeptides comprise about 30 percent of the mature embryo protein and also account for about 30 percent of the protein synthesis during much of the growth phase of embryogenesis. They were, thus, a tempting target for further study. The physical characteristics and biosynthesis of these storage protein polypeptides have been described (Dure and Chlan, 1981; Dure and Galau, 1981). Two major molecular weight (denatured) forms of 48 and 52 kd, each comprised of between 5 and 7 isoelectric variants, are observed in the mature embryo. There are in addition a large number of smaller, protein body-associated polypeptides (Chapter 27). The polypeptide composition of several of the native storage protein complexes have been examined (Cherry and Leffer, 1984). A similar set of 39, 67 and 70 kd polypeptides are observed during embryogenesis (Dure and Galau, 1981) but not in the mature embryo. Several techniques were used to work out the biosynthesis of these proteins, including exposure of excised embryos to radioactive amino acids and *in vitro* translation of the embryo mRNAs, both of which were followed by one and 2-dimensional gel electrophoresis. As tools to identify related polypeptides, polyclonal antibodies were made against each of the two major mature molecular weight forms, and a stain for carbohydrate was used to identify glycosylated proteins *in situ* after gel electrophoresis. The combination of techniques led to a suggested processing scheme presented in Dure and Galau (1981). The mature 52 kd protein is initially synthesized as a preproprotein, most likely the 60 kd form. It is presumed to co-translationally lose a signal peptide used to target the product for the endoplasmic reticulum (reviewed by Silhavy *et al.*, 1980), and at some time it is glycosylated to yield a 70 kd intermediate proprotein. This is then slowly cleaved to yield the glycosylated 52 kd mature protein plus smaller polypeptides. The rate of cleavage is very slow, such that the 70 kd proprotein is easily detectable in embryogenesis as a stainable protein. The 48 kd mature protein is likewise synthesized as a preproprotein, most likely the 69 kd form. It loses a signal peptide to yield a 67 kd proprotein and is then rapidly cleaved to form the 48 kd mature species plus smaller polypeptides. Although not worked out in *Gossypium*, the sites of processing are presumed to be the endoplasmic reticulum and protein bodies, as have been described for other storage proteins (Chrispeels *et al.*, 1982; Bollini *et al.*, 1983). All protein species of 48, 52, 60, 67, 69 and 70 kd are seen nearly equally well by antibodies made against either of the two mature 48 and 50 kd forms. In addition, the amino acid composition of the two mature forms are very similar, if not identical. Thus, they appear to be very clearly related at the amino acid sequence level.

The total number of genes coding for each of these sets of proteins is still not

directly known. One difficulty has been the inability to resolve in the first isofocusing dimension the *in vitro*-synthesized 60 and 69 kd preproproteins in order to detect the number of their isoelectric variants. Judging from the number of variants seen in the proprotein intermediates, however, a minimum of about 3 and 6 genes are believed to code for the 48 and 52 kd, species respectively. The unsupported assumption here is that each isoelectric variant is encoded by a single gene. This need not be the case at all, depending on the conservation of amino acid sequence, minor post-translational modifications, and the extent of artifactual protein modification in the separation techniques. In this as well as many other respects, the cotton storage proteins exhibit many of the features seen in other seed storage proteins (Brown *et al.*, 1982; Chlan and Dure, 1983).

A variety of observations suggested that the mRNAs for the storage proteins are highly regulated and that the expression of both of the molecular weight sets are highly coordinated. Thus, isolation of these genes would provide a collection of several functionally similar genes sharing the same control sequences, especially if they are a large multigene family dispersed throughout the genome such that each gene is transcribed independently of the others. Towards this end recombinant DNA clones containing young embryo cDNAs were made and screened for those which contained storage protein mRNA sequences (Galau *et al.*, 1983). As predicted, about 35 percent of the cDNA clones were complementary with very abundant young embryo-specific mRNAs. These turned out to code for the storage proteins. Confirmation of their identity entailed hybridization of the cloned cDNAs to the appropriately sized abundant mRNAs (60 kd and 69 kd preproproteins are synthesized from 1.9 and 2.2 kb mRNAs, respectively) and the ability of these cloned DNAs to hybridize with mRNAs, which subsequently are translatable *in vitro* into the 60 kd and 69 kd preproproteins. Surprisingly, three different mRNA sequences were discovered to code for storage proteins, not just two as had been surmised (Figure 3). The 69 kd preproprotein is encoded in a single sequence as presumed. The genes containing this sequence are probably divergent in about 15 percent of their nucleotide sequence in their mRNA transcripts. What was unexpected was that not one but two sequences encode the 60 kd preproprotein. They show very little homology with each other, but the genes within each sequence group seem to have diverged very little. Although at the protein level the 60 and 69 kd preproproteins are clearly related by antibody tests, no homologies at the nucleotide level are evident from DNA-mRNA hybridization studies.

Representative full length cDNA sequences for each of the three subfamilies have now been sequenced (Dure and Chlan, 1985) in order to deduce their amino acid sequences and their relatedness at the nucleotide level. Putative glycosylation and cleavage sites have been identified by homology with known functional sites in other proteins. If these are functional in *Gossypium*, then these data strongly suggest that both 52 kd and 48 kd mature proteins are products of the 69 kd preproprotein family. They differ principally by the presence or absence of the

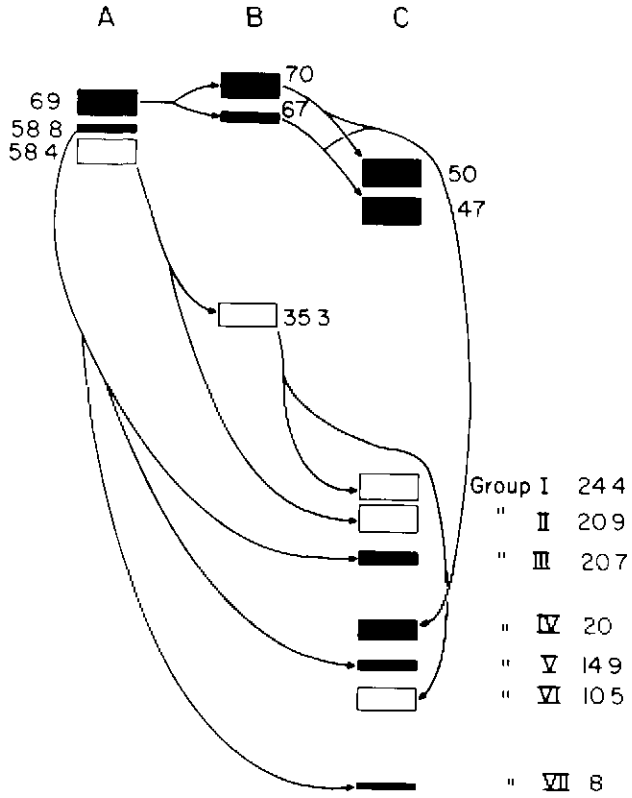

Figure 3. Biosynthesis of the major cotton storage proteins (after Dure and Chlan, 1985). The postulated pathway is based on sequences of representative cDNA clones, assuming the observed glycosylation and cleavage sites are functional. The numbers next to each protein species are their sizes, in kd, deduced from their nucleotide sequences. (A, preproproteins; B, proproteins; C, mature proteins).

glycosylation site. All three classes of preproprotein have several potential cleavage sites, suggesting that all of the many insoluble species, which together comprise 60 percent of the mature seed protein (Dure and Chlan, 1981), arise from these three families of transcripts.

The cDNAs are also being used to isolate the genes from *G. hirsutum* (Kamalay and Dure, unpublished) to discover any putative controlling sequences. In the meantime we have examined seed protein from the two A genome diploids and of five of the D genome diploids by 2-dimensional gel electrophoresis (Hughes and Galau, unpublished) using *G. hirsutum* seed proteins as internal molecular weight markers (Hughes and Galau, 1984). All species so far looked at have both

major molecular weight forms with sizes identical to those in *G. hirsutum*. Furthermore, the 48 kd protein set in all species appears to be very similar in the number of polypeptide species and in their isoelectric points. However, the A genome species synthesize only the alkaline members of the *G. hirsutum* 52 kd set while the D genome species synthesize only the acidic members. Thus, it would seem that the genus will provide genes of sufficient divergence so that sequences of regulatory importance (those expected to change very little) may be deduced by simple comparative analysis of the same gene in several selected species.

The expression of mRNAs for each of the three sets have been examined in detail (Dure *et al.*, 1983a; Galau, unpublished) by quantitative hybridization with the cloned cDNA sequences. At the stage of maximum expression, the mRNAs for the 69 kd preproproteins and one of the two 60 kd preproprotein sets are each about 15 percent of the total mRNA, while the mRNAs for the second 60 kd preproprotein set are only about 5 percent of the total mRNA. This difference in expression of the two 60 kd preproprotein sets may reflect different gene numbers or differences in the efficiency of gene expression. All three sets of genes appear to be coordinately regulated since the concentration of all three mRNA sequences change in parallel during normal development, from about 5 percent of maximum at 5 mg wet weight, with an abrupt decline to about 0.2 percent of maximum during desiccation and a small further decline during the first day of germination. They remain at about that level throughout several days of germination, and it would seem likely, though not proven yet, that these sequences must be resynthesized during germination. It is not yet known if these mRNAs are functional during germination; they are, however, the appropriate size for mature transcripts. A similar situation occurs with soybean storage protein mRNAs (and other very abundant embryo mRNA sequences) throughout a period of several weeks after germination (Goldberg *et al.*, 1981b). Since there should be no functional requirement for storage protein synthesis in embryogenesis, it seems strange that expression may not be completely repressed. One could argue that in a tissue destined for senescence, there would be no need for such a fine control. Alternatively, continued expression at 10^{-3} of the maximum level might be a necessary consequence of the control mechanisms necessary for very high, yet modulatable, expression at other times. Knowing the level of transcripts in adult tissues would be illuminating in this respect. In soybean, similar sequences are not detectable in leaves (Goldberg *et al.*, 1981b). Thus, it may be that once turned on in the cells of the embryo it is impossible to completely turn such genes off.

LATE EMBRYO-ABUNDANT (SUBSET 5) mRNAs

Recombinant DNAs containing cDNAs complementary to other subset mRNAs have been isolated as well (Galau and Dure, unpublished) and are now being studied. A major focus in our laboratory is the regulation of Subset 5

mRNAs which are normally very high in concentration only in late embryogenesis. Apparently these are the only abundant mRNAs which are specifically modulated by absicic acid in the excised young embryo system. It is hoped that analysis of these polypeptides and of their expression will speak to the proposition that the growth regulator plays a role in late embryogenesis in preventing vivipary (for reviews see Walton, 1980, 1981; Black, 1980, 1981; Khan, 1980, 1981; King, 1982).

The proteins themselves are highly soluble, not post-translationally processed *in vivo*, and fairly abundant in the mature embryo. Fifteen different cloned mRNA sequences have been identified so far as being late embryo-specific. The polypeptides they encode have been identified for 10 of these sequences (Figure 4). Eight of the cloned sequences hybridize with mRNAs that code for 2 to 6

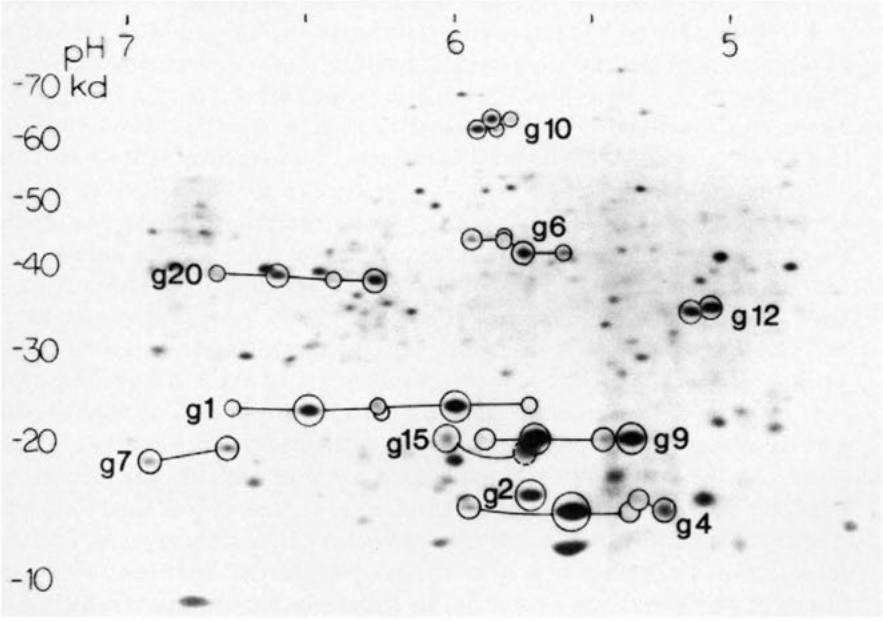

Figure 4. Late embryo-abundant (Subset 5) proteins encoded by individual cloned cDNA sequence groups. Shown are the *in vitro* translation products of mature embryo mRNA separated by two dimensional gel electrophoresis. The linked circles enclose all the polypeptides whose mRNAs hybridize with a single cDNA clone (Galau and Dure, unpublished).

different polypeptides detected on two-dimensional gels of *in vitro*-synthesized protein. Thus, each of these sequences are probably transcribed from a small multigene family. Most of them had previously been identified as Subset 5 products in the earlier global study (Dure *et al.*, 1981). Some escaped identification but can now be seen to be late embryo-specific and inducible in the excised

embryo system. Some of these mRNA sequences show distant homologies with each other, even though they encode significantly different sized polypeptides (groups 2 and 4; groups 6, 7 and 10 in Figure 4), suggesting that there is a functional relationship in the coding or noncoding regions of these mRNAs. Preliminary analysis of the seed protein in the diploid species, mentioned above, indicates that the multigene family in the allotetraploid is based in part on two smaller multigene families, one present in each of the A and D subgenomes. Using these cDNA probes, these genes are also being isolated from *G. hirsutum* (Kamalay and Dure, unpublished).

The expression of these Subset 5 genes have also been followed by *in vivo* and *in vitro* translation and by hybridization with the cloned cDNA probes (Galau, and Hughes, unpublished). For the most part the different genes within each family, for which the products are distinguishable in protein electrophoresis, are coordinately regulated in both normal and experimental embryogenesis. There are some apparent exceptions, and some translational level regulation may also occur in the expression of at least one of the gene families during abscisic acid induction. These issues will be explored further with gene-specific cDNA probes and immunoelectrophoresis to quantitate individual protein members of each family.

Like in the storage protein mRNA analysis, the current cloned cDNA probes for Subset 5 mRNAs react with all of the transcripts of a multigene family, so they measure the sum of the individual transcripts from the entire family. Fourteen of the sequence family mRNAs have now been measured in normal embryogenesis and early germination by such DNA-RNA hybridization. At least two and possibly three groups of coordinately regulated families can be discerned within this population. Although all show their highest expression during desiccation or in the mature embryo, some are modulated in early and middle embryogenesis as well, suggesting that they may be responsive to several temporally separated signals in embryogenesis. The number of different coordinately regulated sets of genes in this group of sequences is being further defined by looking at the kinetics of accumulation and decay of the mRNAs and their dose-response in the experimental induction system. In addition, possible expression will be tested in non-embryogenic tissues and in cell culture under conditions where abscisic acid is thought to be an important regulator, such as in cold stress (Rikin, 1979) and water stress (Wong and Sussex, 1980a, b). Hopefully, the inducing agent for at least some of these genes can be identified in more detail and the level determined at which the mRNA concentration is subsequently regulated.

FURTHER DIRECTIONS IN THE STUDY OF DIFFERENTIAL GENE ACTIVITY

Investigation of gene activity in cotton embryogenesis has used principally the tools of nucleic acid and protein analysis and molecular cloning to study two groups of genes which vary widely in terms of function, time of expression in

normal embryogenesis and in their regulation in excised embryos. Clearly there is a lot to be learned about gene structure and regulation. Much will be learned by comparative sequence analysis of the isolated genes, including putative control elements and the amino acid sequence of the proteins. Hopefully, at least some members of Subset 5 will prove to be inducible in normal embryogenesis with abscisic acid and perhaps play some role in the prevention of vivipary.

There are several tools, mostly biological, which should be of great help in further analysis. The limited screening of other *Gossypium* species has already proved useful. They should continue to be important in understanding gene structure and evolution. In some we may find natural variation in their expression as well. A viviparous mutant would be very useful in which to test directly the supposed connection between abscisic acid, vivipary and late embryo-abundant proteins. Lacking such mutants, we will attempt their creation in plants and in cell culture (Wong and Sussex, 1980a,b). *In ovulo* embryo culture (Stewart and Hsu, 1977a) also is attractive as an experimental system, which should be closer to normal development, so that abscisic acid-induced changes can be more easily seen independent of excision induced events. Finally, antibodies to the Subset 5 proteins would be useful in several areas, especially in their localization and evaluation of their biochemical function.

As a final point, in the absence of a wide variety of mutants which vary in the expression of the particular gene of interest, the conclusions derived from descriptive and comparative analysis of wild type genes and their expression is limited. For this reason, *in vitro* and *in vivo* expression systems are being developed in a wide variety of organisms, in order to test directly the conclusions derived from such studies. They are also of use in the isolation of functionally defined DNA sequences without having to first isolate particular genes. This area of research is moving rapidly in some other dicot plants (Caplan *et al.*, 1983). It is difficult to predict which approaches will have the best success, but expression systems in cotton will probably require improvements in cotton tissue culture and regeneration, and perhaps the modification of current delivery and transformation systems. These are long term goals but clearly are of critical importance, if we are to enter the predictive phase of the study of cotton biology.