Insect Biochem. Molec. Biol. Vol. 22, No. 4, pp. 387–398, 1992
Printed in Great Britain.

0965-1748/92 $5.00 + 0.00

DEDUCED SEQUENCES SHOW MULTIPLE REPEATS IN TWO D PROTEINS FROM THE TUBULAR ACCESSORY GLANDS OF *TENEBRIO MOLITOR*

GUIDO C. PAESEN, FRANTISEK WEYDA* and GEORGE M. HAPP†
Department of Zoology, University of Vermont, Burlington, VT 05405, U.S.A.

(Received 25 July 1991; revised and accepted 20 December 1991)

Abstract—The D group proteins are major secretory products of the tubular accessory glands of male mealworm beetles (*Tenebrio molitor*). They vary in apparent molecular mass between 22 and 30 kDa. In this paper we present the deduced amino acid sequence of two similar acidic D proteins, termed D1 and D2. These proteins have a structure based on peculiarly repeated sequences. Both molecules consist of three domains. Proceeding from the amino terminus, the acidic A and A′ domains are predicted to contain long alpha helical segments. The carboxy-terminal region or B domain is basic and is composed of shorter alpha helical stretches interrupted by turns. D1 and D2 differ in the number of repeats within the A′ and B domains.

Key Word Index: accessory glands; D protein; sequence; *Tenebrio molitor*; repeat

Accessory sex glands of male insects secrete salts, small sugars, lipids and amino acids, as well as a wide range of proteins. In many insects some of these proteins contribute to the formation of special structures—called spermatophores—within which sperm and seminal fluid are transferred to the female. Other proteins form part of the seminal fluid where they facilitate sperm storage, maturation or activation, or they affect physiology and behavior of the female after copulation (Leopold, 1976; Chen, 1984; Happ, 1992). Many proteins have been described on the basis of their mobility on gels and their antigenicity but very few have been isolated and characterized.

The accessory gland proteins considered in the present study are found in the male mealworm beetle, *Tenebrio molitor*. In *Tenebrio*, the male accessory gland complex consists of a pair of bean-shaped accessory glands (BAGs) and a pair of tubular glands (TAGs) (Happ, 1984). The secretions from both glands form a semisolid mass, the secretory plug, which is converted to a complex, multilayered and cylindrical spermatophore during its passage through the ejaculatory duct (Dailey *et al.*, 1980). The lumen of the completed spermatophore contains sperm and seminal fluid and is bounded by an outer wall and a central mass of secretion, the core (Gadzama and Happ, 1974). The best described BAG products are some spermatophorins (proteins that contribute to the spermatophore wall and core) (Grimnes and Happ, 1986; Grimnes *et al.*, 1986; Shinbo *et al.*, 1987) and the enzyme trehalase (Yaginuma and Happ, 1989).

The secretory epithelium of TAGs is composed of only one cell type (Gadzama *et al.*, 1977). In the fully differentiated state, the tubular glands incorporate leucine into four soluble protein groups (termed A, B, C and D). These protein groups have been defined on the basis of electrophoretic mobilities, patterns of leucine incorporation, and reactions with polyclonal antisera. Together, they account for almost 90% of the leucine incorporation in TAGs of 8 day adult males (42% incorporation in A and B proteins, 22% in C proteins and 26% in D proteins) (Black *et al.*, 1982).

There are several D proteins. On Coomassie-stained SDS-polyacrylamide gels of TAG extracts, Grimnes and Happ (1985) described five D protein bands, three of which appear to be abundant. SDS-gels of spermatophores did not reveal components with the same mobility as D proteins. Using pI-SDS two-dimensional polyacrylamide gels, Black *et al.* (1982) report that the [^{14}C]leucine-labeled D proteins with apparent molecular masses between 26 and 29 kDa were divisible into an acidic group (pI = 4.5–5.0) and a basic group (pI = 7.5–8.0) (10).

In the present paper we report the production of a cDNA expression library for adult TAGs and the purification of a major D protein by high performance liquid chromatography. With a monoclonal antibody, we screened the cDNA library and determined the sequence of two immunopositive clones. The complete sequence of two acidic D proteins was deduced.

EXPERIMENTAL PROCEDURES

Animals

Mealworms (*Tenebrio molitor* L.) were purchased from a commercial supplier. They were kept at room temperature and reared on chicken feed. Males and females were separated in the pupal stage.

*Present address: Department of Developmental Morphology, Institute of Entomology, 37005 Ceske Budejovice, Czechoslovakia.
†Author for correspondence: Marsh Life Science Bldg, Department of Zoology, University of Vermont, Burlington, VT 05405, U.S.A.

Gel electrophoresis and western blotting

Tissues were excised from pupal and adult mealworms and homogenized in distilled water. SDS-polyacrylamide electrophoresis (12% gels) followed the methods of Laemmli (1970). Immunoblotting was carried out according to Harlow and Lane (1988), using a Bio-Rad Trans-Blot Cell for protein transfer onto nitrocellulose sheets (Hoefer Scientific Instruments). The blots were developed with two monoclonal antibodies, designated T30.1 and PL17.2. The secondary antiserum consisted of anti-mouse IgG (H + L) labeled with alkaline phosphatase (Promega). 1% Bovine serum albumin (BSA) was used for blocking. The monoclonals were produced after injection of mice with homogenates of the secretory plug (PL17.2) or TAGs (T30.1), using methods described previously (Grimnes and Happ, 1986). They were screened by ELISAs against soluble TAG proteins.

Protein purification

TAGs were dissected from adult mealworms and homogenized in distilled water. The homogenate was centrifuged (12,000 *g*; 3 × 10′) and the supernatant was submitted to HPLC, using a Vydac C4 Reverse Phase Column and a 0.1% HFBA running buffer with a 10–80% acetonitrile gradient. Fractions containing D proteins were further purified by a second HPLC run (Vydac C_{18} Reverse Phase Column, 0.1% TFA, 0–100% Acetonitrile). For detection of D proteins in the fractions and estimation of their purity, aliquots of each fraction were resolved by SDS–PAGE and immunoblotted using the monoclonal antibody T30.1.

Amino acid composition and N-terminal sequence analyses were performed in the protein analytical facility of the Medical Biochemistry Department at the University of Vermont. Purified D protein samples were transferred to hydrolysis vials, supplemented with norleucine as an internal standard and treated for 24 h at 110°C in evacuated, sealed tubes containing 6N HCl with 2% thioglycolate. Separation of amino acids was by ion exchange chromatography using an Interaction AA 911 column. Amino acids were quantitated by postcolumn orthophthalaldehyde derivatization and fluorescence detection. Proline was analyzed in a separate run with postcolumn treatment with hypochlorite. Sequence analysis was performed on an Applied Biosystems 475A Protein Sequencing system employing a gas phase Edman degradation protocol with online HPLC identification of PTH derivatives of amino acids.

cDNA library construction

TAGs were excised from 5 to 8 day adult mealworms, rinsed with distilled water and collected in liquid nitrogen. For the preparation of total RNA, we used the Stratagene RNA isolation kit, which utilizes a rapid guanidium thiocyanate/phenol chloroform extraction. Spin columns containing oligo(dT)-cellulose (Pharmacia LKB) were used to purify mRNA from total RNA. cDNA was prepared with the Pharmacia LKB cDNA synthesis kit and inserted into the Lambda Zap II vector (Stratagene). The library was packaged using the Gigapack II Gold Packaging Extract (Stratagene) and amplified as described by Short *et al.* (1988), but with XL1-Blue instead of BB4 cells.

cDNA library screening

The library was screened according to Mierendorf *et al.* (1987), with the monoclonal antibody T30.1. The cDNAs were expressed in XL1-Blue cells on NZYM plates. The *p*Bluescript SK(−) phagemid of positive clones was excised *in vivo* using the R408 helper phage, as described by Short *et al.* (1988).

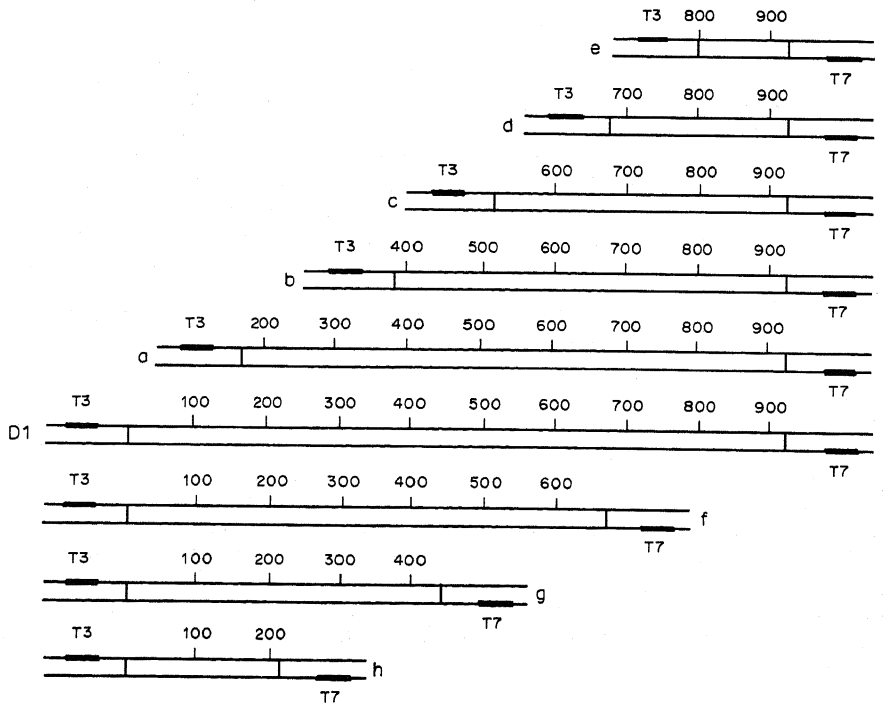

Fig. 1. Strategy for sequencing of the D1 clone. The polylinker region at the T3-side was digested with *Sac* I and *Sma* I. Progressively longer pieces of the cDNA insert were removed using exonuclease III and mung bean nuclease. The deletion products a–e, as well as the intact cDNA, were then submitted to single-stranded and double-stranded sequencing using the T3-primer. The T7-primer site makes sequencing in the other direction possible. It was used for double-stranded sequencing of the intact cDNA and the deletion products f–h, which were created after digestion of the polylinker at the T7-side with *Eco* R V and *Kpn* I. The deletion and sequencing strategy for the D2 clone was the same.

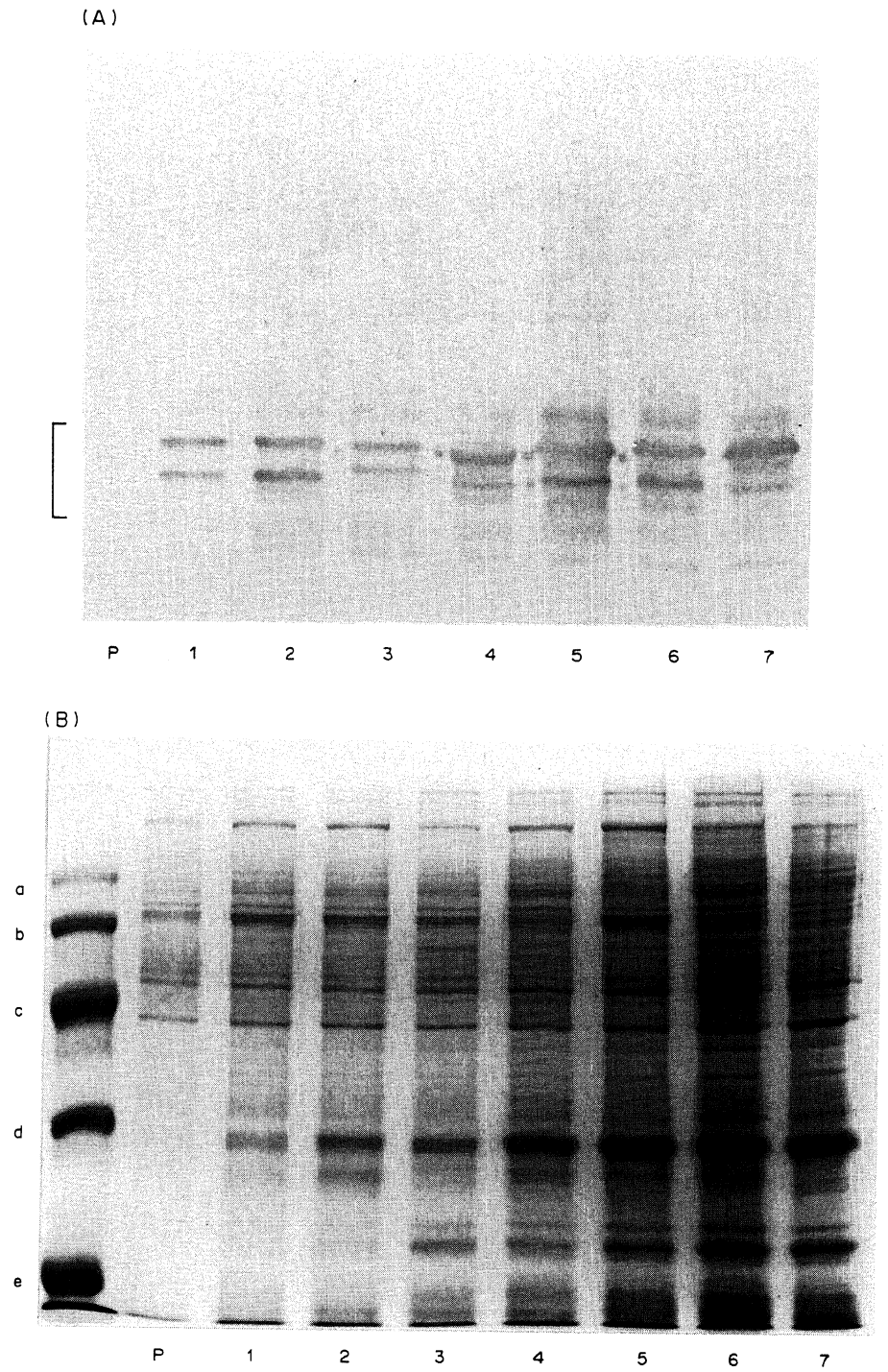

Fig. 2. D proteins separated by electrophoresis on a 12% SDS-polyacrylamide gel. Lane P contains the soluble proteins of pupal TAGs (approx. 1 day before adult ecdysis); lanes 1–7 contain TAG proteins of 1–7 day old adults, respectively. The bracket indicates the D protein region. (A) Immunoblotting and visualization of D antigens recognized by T30.1. (B) Coomassie blue stained gel. Molecular weight of the standards: a = 97,400, b = 68,000, c = 43,000, d = 29,000, e = 18,400.

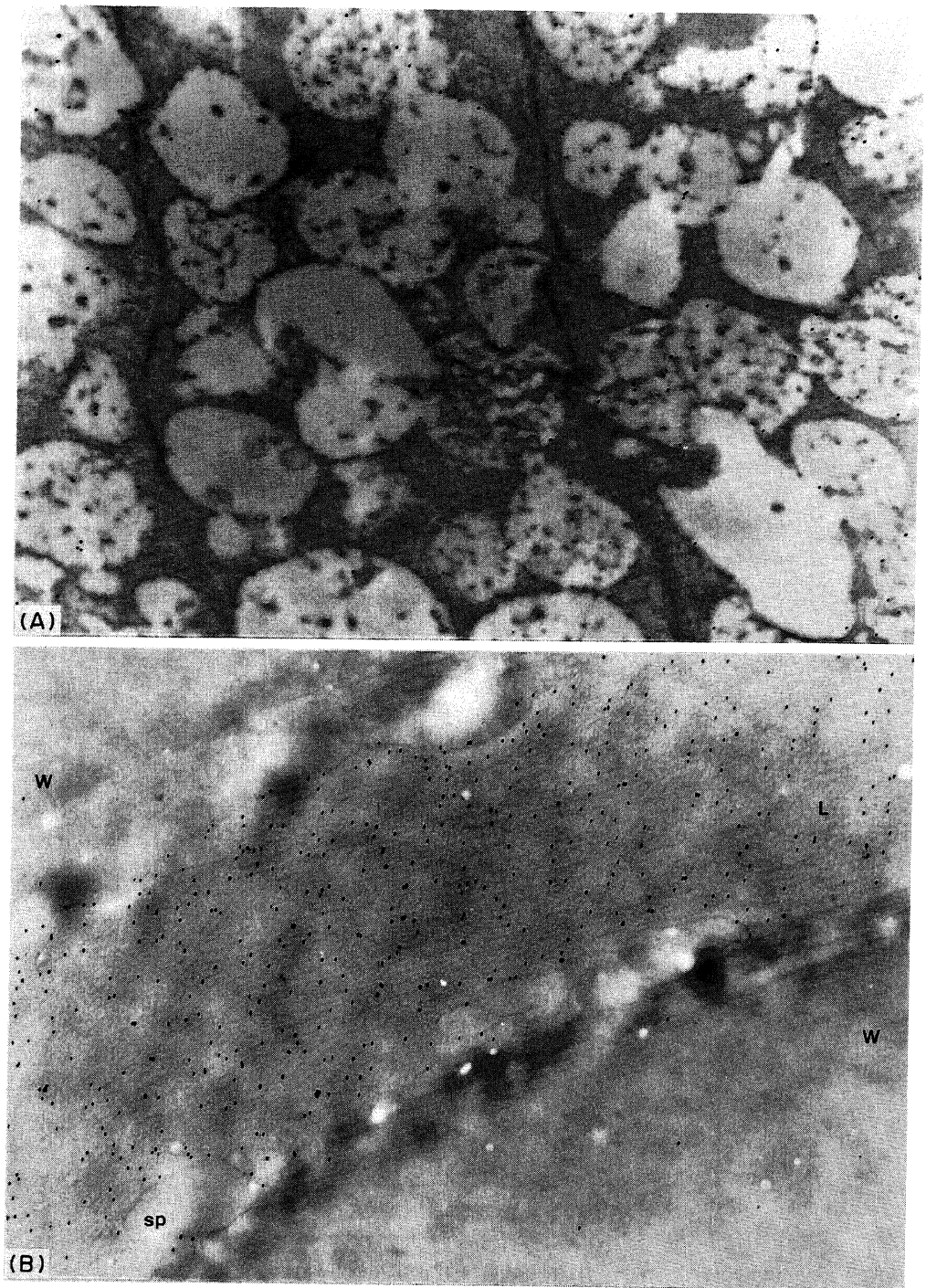

Fig. 3. Localization of PL17.2-bound epitopes in secretory vesicles of TAG cells (A) and in the lumen of the spermatophore (B) by means of immunoelectron microscopy. L = lumen, sp-sperm, W = wall.

Sequencing

Prior to sequencing, a nested set of deletions was generated, using the Stratagene Exonuclease III/Mung Bean Nuclease Deletion kit (Fig. 1). The deletion products were sequenced by the Sanger dideoxy-mediated chain termination reaction (Sanger and Coulson, 1975), using Sequenase (UCB). Double stranded as well as single stranded templates were prepared. Single stranded DNA was rescued and purified according to Short *et al.* (1988). Double stranded phagemid DNA was purified on Plasmid Quik columns (Stratagene) and alkali-denatured (Mierendorf and Pfeffer, 1987). Compressions were resolved by substituting dITP for dGTP. Sequence data were analyzed using IBI-Pustell and Genetics Computer Group (GCG) (Devereaux *et al.*, 1984) sequence analysis software.

Post-embedding immunoelectron microscopy

4% Formaldehyde (Polysciences) in 0.1 M Pipes (pH 7.2–7.3) was used for fixation of TAGs and spermatophores (2 h, room temperature). After a short dehydration in 50 and 75% ethanol, TAG and spermatophore samples were incubated in an ethanol:LR White resin mixture (1:2) and thereafter embedded in undiluted LR White resin (Ladd) (Kann and Fouquet, 1989). Sections were immunostained using the monoclonal antiserum PL17.2 and goat anti-mouse antibodies labeled with colloidal gold particles (Sigma). Other sections were stained with the T30.1 antiserum. Antibodies were dissolved in a 10 mM phosphate buffer solution (pH 7.3) containing 1% BSA. 10% goat serum (Sigma) was used as a blocking solution. Control sections underwent the same procedure, except for the incubation with primary antiserum, which was omitted. The stained sections were examined on Philips EM 201 and 300 transmission electron microscopes.

RESULTS

Specificity of the monoclonal antibodies

On western blots of TAG homogenates, both the T30.1 (Fig. 2) and the PL17.2 (data not shown) antibodies recognize bands which have the mobilities characteristic of the D proteins. Homogenates of BAGs, ejaculatory duct, vas deferens, seminal vesicle, testis, and male fat body did not contain similar antigens.

For the western blot shown in Fig. 2, we used homogenates of TAGs from males of the late pupal, young adult, and mature adult stages. The antigen for T30.1 is adult-specific. D proteins accumulate following adult ecdysis. This confirms the earlier data, which show a marked increase in leucine incorporation into D proteins of young post-ecdysial adults (Happ *et al.*, 1977).

Immunoelectron microscopy with PL17.2 revealed antigens in the secretory vesicles of TAG cells and among the sperm cells in the lumen of the spermatophore (Fig. 3). This distribution suggests that D proteins are secreted from the TAGs and form part of the seminal fluid. The same result was obtained when T30.1 was used instead of PL17.2. The T30.1 antiserum also recognized antigens in the cuticule of the ejaculatory duct. It seems unlikely that the cuticular antigen originates from the TAGs.

HPLC purification of a D protein

After two cycles of HPLC purification, a fraction was recovered which contained a single D protein band on Coomassie stained SDS-gels (apparent molecular mass of 29 kDa) and a single immunoreactive spot on western blots. The amino acid content of hydrolysates of this fraction is summarized in Table 1. The N-terminal amino acid sequence is GIIKPVDNAEARWAPDDD.

cDNA sequences of immunopositive clones

Thirteen immunopositive clones were isolated from the TAG cDNA library. The DNA sequences of clone number 1 (termed D1) and clone number 2 (termed D2) were determined. The D1 sequence is presented in Fig. 4. Translation of the open reading frames predicts products corresponding to the apparent molecular weight of D group proteins. Both the HPLC-purified D protein and the products predicted from clones D1 and D2 are rich in alanine and aspartic acid (Table 1). Neither D1 nor D2 contain cysteine.

Inspection of the open reading frames reveals no AUG startcodon. However, near the N-terminal end of the open reading frame of both clones is a sequence of amino acids (GIIKPVDNAEARWAPDDD) that corresponds exactly to the N-terminal sequence determined directly from the HPLC-purified D protein (Fig. 5). This result confirms that the isolated cDNAs indeed code for D proteins. It also indicates the existence of a "pro-piece" (ending with the amino acids AFVVGLAHA) that is removed by cotranslational modification. This pro-piece indeed has characteristics of a signal sequence; it is nonpolar and has a typical signal peptidase recognition site (AHA) at its carboxyl end (Perlman and Halvorson, 1983).

Both clones contain termination codons. The base sequences downstream of the stopcodon are almost identical in the two clones, apart from their difference in length. The D2 cDNA contains a polyadenylation signal (AATAAA). It appears that the cDNAs are not a complete representation of the mRNAs of which they were derived; startcodons and polyA-tails are absent. The clones nevertheless contain all the information to deduce the entire amino acid sequence of the mature D proteins (i.e. as they appear after cotranslational modification).

Molecular weights calculated from the inferred sequences without the putative signal sequence, are 29,122 for clone D1 and 25,344 for clone D2. Predicted isoelectric points are 4.3 and 4.2, respectively. These data are consistent with the electrophoretic migration pattern of D proteins on 2-D gels (Black *et al.*, 1982).

Protein structure of the D1 protein

Leaving aside the first 8 amino acids which are presumed to be part of the signal sequence, the mature protein can be divided into three domains, which we termed A, A′ and B (Fig. 5). The A′ domain (from amino acid 105 to 188) is in fact a repeat of the A domain (from amino acid 18 to 104). The primary sequence of the B domain (from amino acid 198 to 280) differs substantially from A and A′. A and A′ domains are rich in aspartic and glutamic acid, and are consequently very acidic (pIs about 4.0). B domains, on the contrary, are extremely basic (pI = 9.8), due to lower aspartic acid and higher lysine and arginine contents. The similarity between A and A′,

```
1 GC GCC TTC GTC GTC GGC CTG GCC CAC GCA GGG ATA ATC AAG CCG GTG GAT AAC GCC GAA GCA CGA TGG GCC CCA GAT GAC GAC GCC GAG GCT
    Ala Phe Val Val Gly Leu Ala His Ala Gly Ile Ile Lys Pro Val Asp Asn Ala Glu Ala Arg Trp Ala Pro Asp Asp Asp Ala Glu Ala>

93 TTG GCA AGA AAG GCT CCA GAT GCC CAC GCT GAA GCT CGA TGG GCC CCA GAT GAC GAC GCC GAG GCT ATC GCA AGA AAG GCT CCA GAT AGC
    Leu Ala Arg Lys Ala Pro Asp Ala His Ala Glu Ala Arg Trp Ala Pro Asp Asp Asp Ala Glu Ala Ile Ala Arg Lys Ala Pro Asp Ser>

183 GAC GCC CAC GCT GAA GCT CGA TGG GCT CCA GAT GAC GAC GCC GAG GCT ATC GCA AGA AAG GCT CCA GAT AGC GAC GCC CAC GCT GAA GCT
    Asp Ala His Ala Glu Ala Arg Trp Ala Pro Asp Asp Asp Ala Glu Ala Ile Ala Arg Lys Ala Pro Asp Ser Asp Ala His Ala Glu Ala>

273 CGA TGG GCC CCA TTT GAC GAC GCC GAC ACC GCC CCT CTA TTC AGA TGG GCC CCA GAT GAC GAC GCC GAG GCT TTG GCA AGA AAG GCT CCA
    Arg Trp Ala Pro Phe Asp Asp Ala Asp Thr Ala Pro Leu Phe Arg Trp Ala Pro Asp Asp Asp Ala Glu Ala Leu Ala Arg Lys Ala Pro>

363 GAT GCC CAC GCT GAA GCT CGA TGG GCC CCA GAT GAC GAC GCC GAG GCT ATC GCA AGA AAG GCT CCA GAT AGC GAC GCC CAC GCT GAA GCT
    Asp Ala His Ala Glu Ala Arg Trp Ala Pro Asp Asp Asp Ala Glu Ala Ile Ala Arg Lys Ala Pro Asp Ser Asp Ala His Ala Glu Ala>

453 CGA TGG GCT CCA GAT GAC GAC GCC GAG GCT ATC GCA AGA AAG GCT CCA GAT AGC GAC GCC CAC GCT GAA GCT CGA TGG GCC CCA TTT GAC
    Arg Trp Ala Pro Asp Asp Asp Ala Glu Ala Ile Ala Arg Lys Ala Pro Asp Ser Asp Ala His Ala Glu Ala Arg Trp Ala Pro Phe Asp>

543 GAC GCC GAC ACC GCC CCT CTA TTC AGA TGG GCC CCA GAT GAT GAC GGC GAG GCT GAA GCA AGA CAG GCT CCA AAT AAC GAC TCC CCC ACT
    Asp Ala Asp Thr Ala Pro Leu Phe Arg Trp Ala Pro Asp Asp Asp Gly Glu Ala Glu Ala Arg Gln Ala Pro Asn Asn Asp Ser Pro Thr>

633 GTA CCT CGG ATG TCA ATG GAA GCA AGA AAG GTT CCA AAT AAC GAC TCC CCC GCT GTA CCT CGA GCG TCA CTG GAA GCA AGA AAG GCT CCA
    Val Pro Arg Met Ser Met Glu Ala Arg Lys Val Pro Asn Asn Asp Ser Pro Ala Val Pro Arg Ala Ser Leu Glu Ala Arg Lys Ala Pro>

723 AAT AAC GAC TCC CCC GCT GTA CCT CGA GCG TCA CTG GAA GCA AGA AAG GCT CCA AAT AAC GAC TCC CCC GCT GTT CCT CGA GCG TCA CTG
    Asn Asn Asp Ser Pro Ala Val Pro Arg Ala Ser Leu Glu Ala Arg Lys Ala Pro Asn Asn Asp Ser Pro Ala Val Pro Arg Ala Ser Leu>

813 GAA GCA AGA AAG GCT CCA AAT AAC GAA GCC TAG GCC GAC GAG ATT CTA AGA AAA ACA ATA CAA ATC CGG ACG AGT TGA CTT TTG TTG ATG
    Glu Ala Arg Lys Ala Pro Asn Asn Glu Ala End

903 TTT TAT CGA TAA CGG GTT TCA G
```

Fig. 4. Nucleotide sequence and deduced amino-acid sequence of D1. The (putative) peptidase recognition

and the difference between these domains and B, is also reflected in the predicted hydrophilicity, surface probability, flexibility, antigenic index and secondary structure (Fig. 6). The B domain seems to be more flexible and obtains significantly higher antigenic index and surface probability scores. This is consistent with the predicted secondary structure. Chou–Fasman (Chou and Fasman, 1978) as well as Garnier–Osguthorpe–Robson (Garnier *et al.*, 1978) analyses indicate long alpha helical stretches and few turns in domains A and A′, while shorter alpha helix pieces and more turns are predicted in domain B. Putative phosphorylation and glycosylation sites are restricted to the B domain, adding to the contrast with A and A′ (Fig. 7).

The most prominent feature of the D1 protein is its complex repetitive structure. Not only are A and A′ repeats of each other, the domains themselves are composed from subrepeats, which, in their turn, consist of two very similar halves ("sub-subrepeats"). The B domain is also repetitive. Its subsequence, moreover, is similar to the subunits of the A domains.

The A domain can be seen, *grosso modo*, as a 3.5 times repeat of the major subsequence AEARWAPDDDAEAIARKAPDSDAH, followed by the "unique" sequence DTAPLF and ending on RWAPDDDAE, which is a truncated version of the major subsequence. The first repeat (amino acids 18–39) differs a bit from the following two in that it lacks 2 amino acids (after position 37). Notice the two very similar halves (AEARWAPDDDAE and AIARKAPDSDAH) of which the major subsequence is composed.

The A′ domain is almost identical to the A domain, except that the first 3 amino acids of A are not found in A′ and that amino acid 196 is a glycine in A′ and an alanine in A.

The B domain of the D1 protein is an almost 5-fold, slightly degenerate repeat of the subsequence LEARKAPNNNSPAVPRAS. The first four repeats contain a putative *N*-glycosylation (NDS; Oikawa

Table 1. Amino acid composition of D proteins

Amino acid	D1	D2	D_{HPLC}
Alanine	27.68	28.63	20.7
Aspartic acid/asparagine	19.56	19.65	16.0
Proline	11.44	10.68	ND
Arginine	8.49	8.12	7.9
Glutamic acid/glutamine	7.75	8.12	9.7
Serine	4.43	3.42	5.0
Lysine	4.06	3.85	3.4
Tryptophan	3.32	3.85	ND
Leucine	2.58	2.14	5.9
Isoleucine	2.21	2.56	2.4
Valine	2.21	1.71	3.5
Histidine	2.21	2.56	1.9
Phenylalanine	1.48	1.71	1.4
Threonine	1.11	1.28	1.9
Glycine	0.74	0.85	3.5
Methionine	0.74	0.85	1.5
Tyrosine	0.00	0.00	0.5
Cysteine	0.00	0.00	ND

Amino-acid (AA) composition of the D1 and D2 proteins as derived from their cDNA sequences, are compared to the AA-composition of the HPLC-purified D protein (D_{HPLC}). D1 and D2 compositions are given in percentage of total amino acid content. The D_{HPLC} composition is presented in peak area percentages, multiplied by 0.85 to correct for amino acids that were not determined (ND).

et al., 1987) and a putative serine phosphorylation site (RMS or RAS; Weller, 1979). The fifth repeat is 5 amino acids short. The first part of the repeat (LEARKAPNND) is very similar to parts of the major subsequence of the A domains (AEARWAPDDD and AIARKAPDSD).

The entire D1 protein can be viewed as a 20-fold repeat of a consensus sequence (L/A)(E/I)AR(W/K)AP(D/N)(D/N)D, supplemented with some other amino acid groups. The divergence from this consensus sequence as well as the arrangement of the repeats and the nature of the added amino acids determines the difference between the A and B domains.

Protein structure of D2

The protein coded for by clone D2 is very similar to the D1 protein (Figs 4 and 5). The main difference is that D2 lacks one repeat in region A′, and one in region B. As a consequence it has only 3 putative glycosylation and 3 putative phosphorylation sites.

DISCUSSION

The number of D proteins and their molecular weights

At this point we do not know the exact, native molecular weights of the proteins belonging to the D group. Black *et al.* (1982) used SDS-gel electrophoresis to resolve 3 major (radiolabeled) D protein bands in homogenates of pooled TAGs, corresponding with molecular weights of 29,100, 27,500 and 26,400. Autoradiographs of two-dimensional gels revealed basic as well as acidic D proteins. Grimnes and Happ (1985) ran individual TAGs on SDS-gels and stained the (unlabeled) proteins with Coomassie blue. They reported four D protein bands and molecular masses of 27,730, 26,300, 25,090 and 23,900 Da. Rarely, an additional band with an estimated molecular mass of about 22 kDa was found. On two-dimensional gels, the 27,700, 26,300 and 23,900 variants were found to be acidic. The other (more rare) variants were never seen on Coomassie stained 2D gels.

In this study (Fig. 2), we find at least five T30.1-positive bands, one of which shows a molecular weight somewhat higher than 28.5 kDa (the apparent molecular weight of carbonic anhydrase on SDS-gels). All the others have a slightly higher electro-

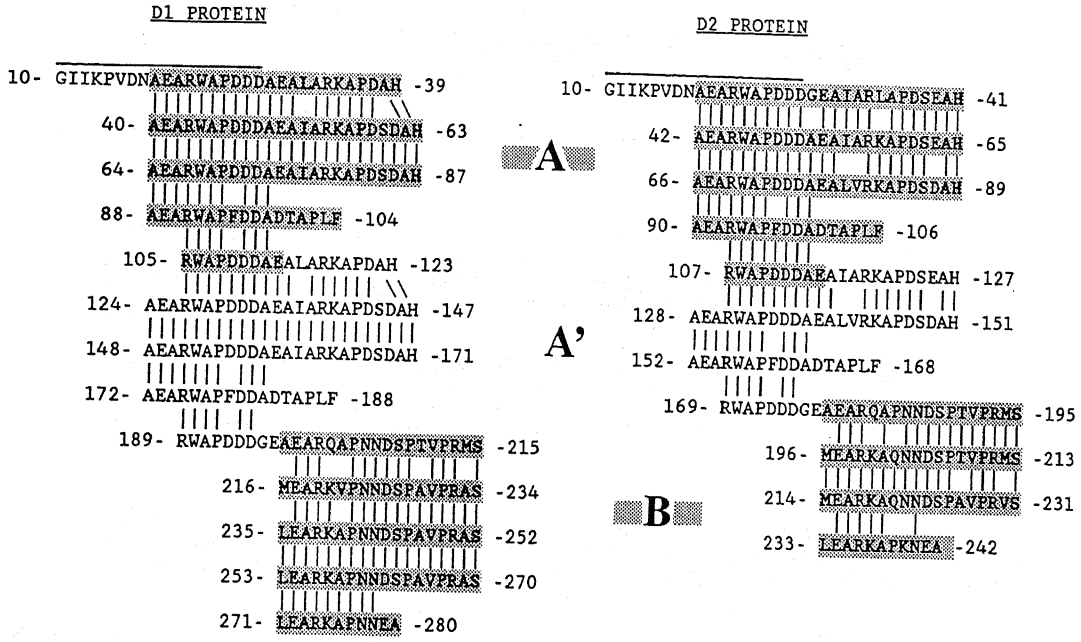

Fig. 5. The amino acid sequence of D1 and D2. After alignment of the repeats, three domains (A, A′, and B) can be distinguished. The first 9 amino acids, presumed to belong to the signal sequence, are not shown in the figure. The overlined region corresponds to the N-terminal sequence obtained by direct sequencing of the HPLC-purified protein. The numbers indicate the positions of the amino acids in the protein.

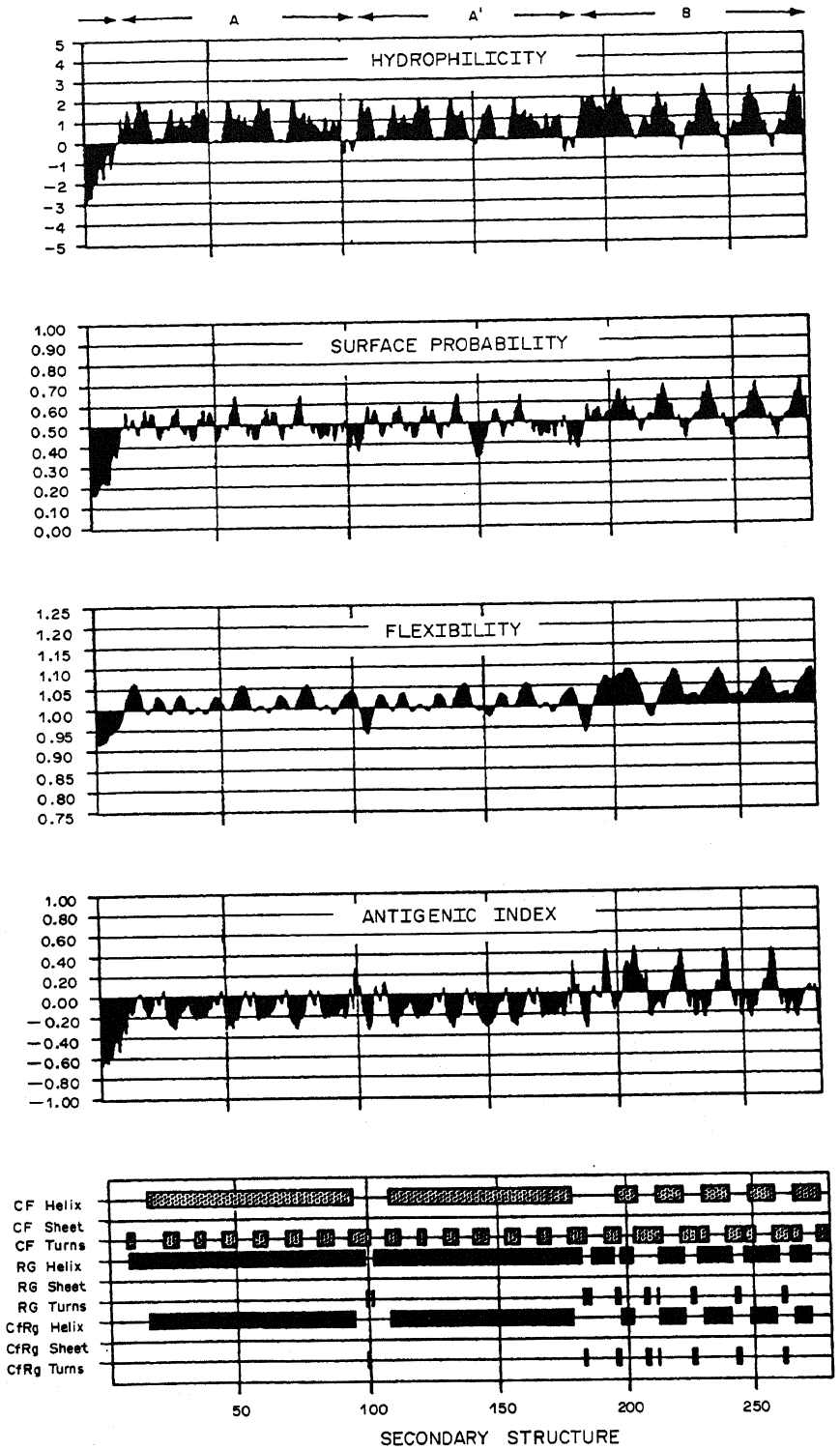

Fig. 6. Schematic representation of hydrophilicity, surface probability, flexibility, antigenic index and secondary structure predictions (IBI, Kodak) for the D1 protein. CF = Chou–Fasman prediction; RG = Robson–Garnier prediction, CfRg = Chou–Fasman–Robson–Garnier prediction consensus.

phoretic mobility than this standard. The band with the second highest molecular weight is stained most intensely and may correspond with the common 27,700-Da variant described by Grimnes and Happ (1985).

Clone D1 is the longest of the 13 examined cDNAs and from the calculated molecular mass of the mature D1 protein (29,036 Da), we suspect that this clone codes for the largest of the D proteins that are usually visible on SDS-gels. If glycosylation occurs, it does

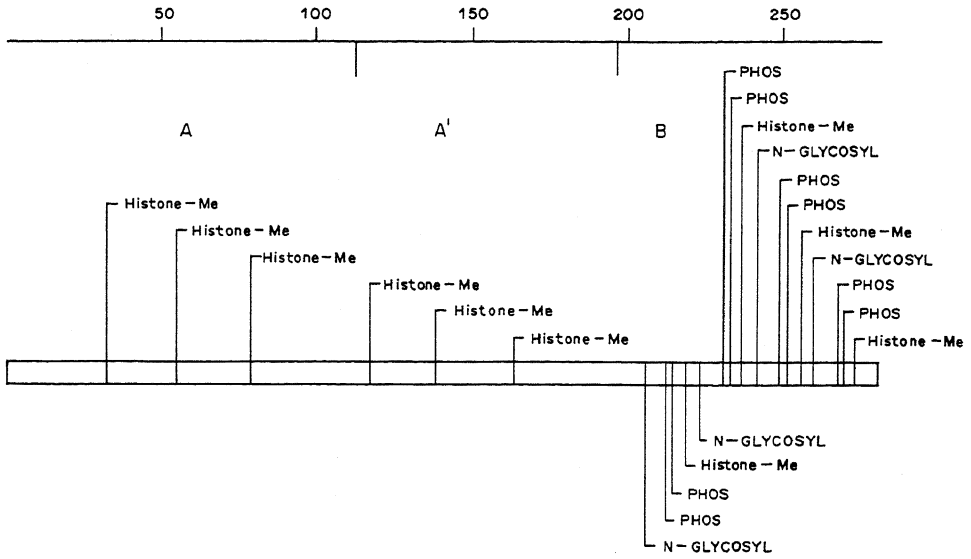

Fig. 7. Putative phosphorylation (PHOS), methylation (Histone-Me) and glycosylation (N-GLYCOSYL) sites in D1.

not seem to affect the electrophoretic mobility. Supposing that this also applies to the other D proteins, we suggest that D2 (calculated molecular weight = 25,172) is seen on gels as the 25,090-Da variant.

What are the function and fate of D proteins?

Several products of the TAGs, including the A/B proteins (10) and an antigen recognized by the monoclonal antibody PL15.2, are found in the lumen of the fully-formed spermatophore. Immunoelectron microscopy with PL17.2 and T30.1 localized D epitopes in the secretory vesicles and lumen of the TAGs and mixed with the sperm in the lumen of the spermatophore (Fig. 3). Western blotting and staining with Coomassie blue confirm the presence of D proteins in the adult tubular accessory glands and their absence in the bean shaped glands. However, homogenates of the spermatophores yielded no Coomassie-stained or immunoreactive D bands. It appears that either D proteins become insoluble, or the epitopes are masked by binding to other molecules in the spermatophore. The fact that the secretory plug can be used to raise a monoclonal antibody (PL17.2) against D proteins, suggests that the latter indeed become part of the spermatophore.

Many roles are played by seminal fluid components, including sperm activation (Osanai *et al.*, 1987), acceleration of egg maturation in the female (Boggs and Gilbert, 1979), initiation of oviposition (Friedel and Gillot, 1976), reduction of mating receptivity (Bauman *et al.*, 1975), inhibition of reinsemination (Adams *et al.*, 1972) or induction of contractions in the female tract to aid sperm movement (Davey, 1959). At this point in the investigation, we do not know what the function of the highly abundant D proteins is.

The role of the D proteins is problematic. We searched the GenEMBL and NBRF databases with the nucleotide and amino acid sequences of D1 and D2 using both GCG and IBI software. There were no highly significant homologies with previously reported sequences. Since no homologies were found with active site regions of enzymes, we think it unlikely that the D proteins are enzymes. Their high abundance also argues against a role as a signal acting on the reproductive physiology of the female. Searches of the databases did reveal some resemblance with the carboxy-terminal regions of the mid-size neurofilament of the chicken (Zopf *et al.*, 1987) and the large neurofilament of the mouse (Julien *et al.*, 1988), as shown in Fig. 8. These regions of the neurofilament proteins form cross-links among neurofilaments and their surrounding structures (Julien and Mushynski, 1983; Hirokawa *et al.*, 1984). It is possible that the D proteins link with one another to form a network that surrounds the sperm bundles and that gives the seminal fluid its viscous character. Another potential similarity is found with

```
D protein B domain repeat          - LEARKAPNNND -
D protein A domain subrepeat 1     - AIARKAPDSDAH -
D protein A domain subrepeat 2     - AEARWAPDDDAE -
Plasmodium S-antigen repeat        - AEARKSDE -
Chicken neurofilament repeat       - AEA KSP -
```

Fig. 8. Alignment of the D protein repeats with the repeats that compose the *Plasmodium* S-antigen and the C-terminal domain of the heavy neurofilament of the chicken. The *Plasmodium* sequence is 20 times repeated. The C-terminus of the neurofilament contains 18 repeats of the sequence AEAKSP and 33 repeats of similar sequences (such as GEAKSP and AEVKSP).

the S-antigen of the NF7 isolate of the malarial parasite *Plasmodium falciparum* (Cowan *et al.*, 1985). The function of the S-antigen is not understood. The repetitive nature of the D molecules, and also their resemblance to the neurofilament carboxy-terminal domains, suggest a structural function.

The repetitive character of the D proteins is also consistent with their roles as binding or transport molecules. If the repeats are binding sites, each D molecule would be an efficient carrier for ions or other small molecules such as polyamines. D proteins share some features with the prostatic spermine-binding protein of the rat (Chang *et al.*, 1987). The acidic part of the rat protein and the D proteins contain multiple aspartic-acid patches, mostly triplets. Low molecular weight polyamines (among them spermine) and small basic proteins are reported to induce mating refusal behavior in virgin female house flies (*Musca domestica*) (Adams *et al.*, 1972).

D proteins also share some structural characteristics with proteins belonging to the calmodulin family (calmodulin, proctolin, parvalbulmin, calcinuerin). The secondary structure of both groups of proteins consists of alpha helical structures alternating with aspartic acid-rich regions; the latter may be analagous to the calcium-binding sites in calmodulin (Wylie and Varnum, 1988). Moreover, the sequence duplications that gave rise to the repeated domains in D proteins and in calmodulin and related proteins may have taken place in a similar manner (*vide infra*).

The evolutionary relationships of the D protein genes

The most prominent characteristic of the D genes is their complex structure, based on repeats, subrepeats and even sub-subrepeats. Comparable features are described for proteins of the Balbiani ring genes of the dipteran *Chironomus tentans* (Pustell *et al.*, 1984), which comprise the larval feeding tube. Mouse (Biro *et al.*, 1975) and guinea pig satellite DNAs (Southern, 1970) also have internal repeats. The special structure of these satellites and of the Balbiani ring genes made them favorable subjects for molecular evolution studies. Repeats are generated by duplication of an ancestor sequence by many different mechanisms (Smith, 1976; Jeffreys *et al.*, 1985; Kornberg *et al.*, 1964; Botchan, 1974). Deletion and addition, as well as simple base substitution contributed to the divergence of the repetitive DNA (Efstratiadis, 1980; Weldon and Kafatos, 1982; Konkel *et al.*, 1979). We believe that parallel events occurred in the genes for D proteins.

We can deduce a probable order of evolutionary events. The DNA sequences coding for the two halves of the major subsequence of the A domains (AEARWAPDDDAE and AIARKAPDSDAH) obviously are the result of a duplication of an ancestor sequence, which was followed by a divergence (by base substitution) between the two daughter sequences. Most of this divergence very probably happened before new duplications copied the combined daughter sequences (corresponding with the future code for AEARWAPDDDAEAIARKAPDSDAH). Indeed, the same substitutions are found in each subsequence of the A domain. The DNA of the A domain ancestor, as a whole, seems to be duplicated only after it was somewhat modified (among others by the "insertion" of the TAPLF sequence). It is not clear whether the A domain gained one extra after this "domain-duplication" or whether the A′ domain lost a repeat. The disappearance as well as the insertion of an extra repeat could be the result of a slipped mispairing during DNA replication, as described by Efstratiadis (1980), or of gene conversion. The DNA sequence corresponding with the B domain repeat (LEARKAPNND...) probably evolved from the same ancestor sequence as the A domain repeats, but must have replicated independently, since it is only found in the tail of the gene. A simultaneous duplication would have resulted in a gene with alternating A and B subsequences.

The fact that repeats are easily lost or gained during evolution is a probable explanation for the existence of more than one D protein. The major subsequence of the A domain has a calculated molecular mass of about 2500 Da, while that of the B domain repeat is 2000 Da. A D2 protein with an extra A (or A′) repeat would have a (calculated) molecular weight of 27,670, close to the molecular weight of 27,700-Da variant, observed by Grimnes and Happ (1985). The 26,530-Da variant, seen on gels, has the same molecular mass as a D1 protein lacking one A repeat (26,500). The 23,900-Da variant is of the same size as a D1 protein that lacks two A repeats (24,000).

We expect that an analysis of the genomic DNA of *T. molitor* will permit further deductions about the evolution of this group of closely related proteins.

Acknowledgements—We wish to thank Christine Yuncker Happ and Kathy Cahill for, respectively, HPLC purifications and immunoblotting. We thank Kay Grimnes for monoclonal antiserum production, Robert Kelm for determination of amino acid composition and Xavier Villarreal for amino-terminal sequence analysis. Partial financial support was provided by grants from the National Institutes of Health (AI-15662) and the USDA (87 CRCR-1-2406).

REFERENCES

Adams T. S., Holman G. M., and Nelson D. R. (1972) Amines that induce monocoitic behavior in the housefly, *Musca domestica*. *Chemosphere* **1**, 39–42.

Baumann H., Wilson K. J., Chen P. S., Humbel R. E. (1975) The amino acid sequence of a peptide (PS-1) from *Drosophila funebris*: A paragonial peptide from males which reduces the receptivity of the female. *Eur. J. Biochem.* **52**, 521–529.

Biro P. A., Carr-Brown A., Southern E. M. and Walker P. M. B. (1975) Partial sequence analysis of mouse satellite DNA: Evidence for short range periodicities. *J. mol. Biol.* **94**, 71–86.

Black P. N., Landers M. H. and Happ G. M. (1982) Cytodifferentiation in the accessory glands of *Tenebrio molitor*. VIII. Crossed immunoelectrophoretic analysis in the postecdysial tubular accessory glands. *Dev. Biol.* **94**, 106–115.

Boggs C. L. and Gilbert L. E. (1979) Male contribution to egg production in butterflies: Evidence for transfer of nutrients at mating. *Science* **206**, 83–84.

Botchan M. R. (1974) Bovine satellite DNA consists of repetitive units 1400 base pairs in length. *Nature* **251**, 288–292.

Chang C., Saltzman A. G., Hiipakka R. A., Huang, I-Y. and Liao S. (1987) Prostatic spermine-binding protein. Cloning and nucleotide sequence of cDNA, amino acid sequence, and androgenic control of mRNA level. *J. biol. Chem.* **262,** 2826–2831.

Chen P. S. (1984) The functional morphology and biochemistry of insect male accessory glands and their secretions. *A. Rev. Ent.* **29,** 233–255.

Chou P. Y. and Fasman G. D. (1978) Prediction of secondary structure of proteins from their amino acid sequence. *Adv. Enzymol.* **47,** 45–147.

Cowan A. F., Saint R. B., Coppel R. L., Brown G. V., Anders R. F. and Kemp D. J. (1985) Conserved sequences flank variable tandem repeats in two S-antigen genes of *Plasmodium falciparum. Cell* **40,** 775–783.

Dailey P. J., Gadzama N. M. and Happ G. M. (1980) Cytodifferentiation in the accessory glands of *Tenebrio molitor*. VII. A congruent map of cells and their secretions in the layered elastic product of the male bean-shaped gland. *J. Morphol.* **178,** 139–154.

Davey K. G. (1959) Spermatophore production in *Rhodnius prolixus. Q. J. microsc. Sci.* **100,** 221–230.

Devereux J., Haeberli P. and Smithies O. (1984) A comprehensive set of sequence analysis programs for the VAX. *Nucl. acids Res.* **12,** 387–395.

Efstratiadis A., Posansky J. W., Maniatis T., Lawn R. M., O'Connell C., Spritz R. A., DeReil J. K., Forget B. G., Weissman S. M., Slightom J. L., Blechl A. E., Smithies O., Baralle F. E., Shoulders C. C. and Proudfoot N. J. (1980) The structure and evolution of the human β-globin gene family. *Cell* **21,** 653–668.

Friedel T. and Gillott C. (1976) Contributions of male-produced proteins to vitellogenesis in *Melanoplus sanguinipes. J. Insect Physiol.* **22,** 489–495.

Gadzama N. M. and Happ G. M. (1974) The structure and evacuation of the spermatophore of *Tenebrio molitor* L. (Coleoptera: Tenebrionidae). *Tiss. Cell* **6,** 95–108.

Gadzama N. M., Happ C. M. and Happ G. M. (1977) Cytodifferentiation in the accessory glands of *Tenebrio molitor*. I. Ultrastructure of the tubular gland in the post-ecdysial adult male *J. exp. Zool.* **200,** 211–222.

Garnier J., Osguthorpe D. J. and Robson B. (1978) Analysis of the accuracy and implications of simple methods for predicting secondary structure of globular proteins. *J. molec. Biol.* **120,** 97–120.

Grimnes K. A. and Happ G. M. (1985) Partial characterization of D group proteins of the tubular accessory glands of *Tenebrio molitor. Insect Biochem.* **15,** 181–188.

Grimnes K. A. and Happ G. M. (1986) A monoclonal antibody against a structural protein in the spermatophore of *Tenebrio molitor* (Coleoptera). *Insect Biochem.* **16,** 635–643.

Grimnes K. A., Bricker C. S. and Happ G. M. (1986) Ordered flow of secretion from accessory glands to specific layers of the spermatophore of mealworm beetles: Demonstration with a monoclonal antibody. *J. exp. Zool.* **240,** 275–286.

Happ G. M. (1984) Structure and development of male accessory glands in insects. In *Insect Ultrastructure* (Edited by King R. C. and Akai H.), Vol. 2, pp. 365–396. Plenum Press, New York.

Happ G. M. (1992) Maturation of male reproductive system and its endocrine regulation. *A. Rev. Ent.* **37,** 303–320.

Happ G. M., Yuncker C. and Huffmire S. A. (1977) Cytodifferentiation in the accessory glands of *Tenebrio molitor*. I. *J. exp. Zool.* **200,** 223.

Harlow E. and Lane D. (1988) *Antibodies.* A Laboratory Manual, Cold Spring Harbor Laboratory, U.S.A.

Hirokawa N., Glicksman M. A. and Willard M. B. (1984) Organization of mammalian neurofilament polypeptides within the neuronal cytoskeleton. *J. cell Biol.* **98,** 1523–1536.

Jeffreys A. J., Wilson V. and Thein S. L. (1985) Hypervariable "minisatellite" regions in human DNA. *Nature* **314,** 67–73.

Julien J.-P. and Mushynski W. E. (1983) The distribution of phosphorylation sites among identified proteolytic fragments of mammalian neurofilaments. *J. biol. Chem.* **258,** 4019–4025.

Julien J.-P., Cote F., Beaudet L., Sidky M., Flavell D., Grosveld F. and Mushynski W. (1988) Sequence and structure of the mouse gene coding for the largest neurofilament subunit. *Gene* **68,** 307–314.

Kann M.-L. and Fouquet J.-P. (1989) Comparison of LR White resin, Lowicryl K4M and Epon post-embedding procedures for immunogold staining of actin in the testis. *Histochemistry* **91,** 221–226.

Konkel D. A., Maizel J. V. Jr and Leder P. (1979) The evolution and sequence comparison of two recently diverged mouse chromosomal β-globin genes. *Cell* **18,** 865–873.

Kornberg A., Bertsch L. L., Jackson J. F. and Khorana H. G. (1964) Enzymatic synthesis of deoxyribonucleic acid. XVI. Oligonucleotides as templates and the mechanism of their replication. *Proc. natn Acad. Sci., U.S.A.* **51,** 315–323.

Laemmli V. K. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **277,** 680–685.

Leopold R. A. (1976) The role male accessory glands in insect reproduction. *A. Rev. Ent.* **21,** 199–221.

Mierendorf R. C. and Pfeffer D. (1987) Direct sequencing of denatured plasmid DNA. *Meth. Enzymol.* **152,** 556–562.

Mierendorf R. C., Percy C. and Young R. A. (1987) Gene isolation by screening λgtll libraries with antibodies. *Meth. Enzymol.* **152,** 458–469.

Oikawa S., Nakazato H. and Kosaki G. (1987) Primary structure of human carcinoembryonic antigen (CEA) deduced from cDNA sequence. *Biochem. biophys. Res. Commun.* **142,** 511–528.

Osanai M., Aigaki T. and Kasuga H. (1987) Energy metabolism in the spermatophore of the silkmoth, *Bombyx mori*, associated with the accumulation of alanine derived from arginine. *Insect Biochem.* **17,** 71–75.

Perlman D. and Halvorson H. O. (1983) A putative signal peptidase recognition site and sequence in eucaryotic and procaryotic signal peptides (1983) *J. molec. Biol.* **167,** 391–409.

Pustell J., Kafatos F. C., Wobus U. and Baeumlein H. (1984) Balbiani ring DNA: sequence comparisons and evolutionary history of a family of hierarchically derived repetitive protein-coding genes. *J. molec. Evol.* **20,** 281–295.

Sanger F., and Coulson A. R. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. molec. Biol.* **94,** 441–448.

Shinbo H., Yaginuma T. and Happ G. M. (1987) Purification and characterization of a proline-rich secretory protein that is a precursor to a structural protein of an insect spermatophore. *J. biol. Chem.* **262,** 4794–4799.

Short J. M., Fernandez J. M., Sorge J. A. and Huse W. D. (1988) λ ZAP: a bacteriophage λ expression vector with *in vivo* excision properties. *Nucl. acids Res.* **16,** 7583–7600.

Smith G. P. (1976) Evolution of repeated DNA sequences by unequal crossover. *Science* **191,** 528–535.

Southern E. M. (1970) Base sequence and evolution of guinea pig α-satellite DNA. *Nature* **227,** 794–798.

Weldon J. C. and Kafatos F. C. (1982) Accepted mutations in a gene family: evolutionary diversification of duplicated DNA *J. molec. Evol.* **19,** 87–103.

Weller M. (1979) *Protein Phosphorylation: The Nature, Function, and Metabolism of Proteins which Contain Covalently Bound Phosphorus.* Pion, London.

Wylie D. C. and Vanamam T. C. (1988) Structure and evolution of the calmodulin family of calcium regulatory proteins. In *Molecular Aspects of Cellular Regulation: Calmodulin* (Edited by Cohen P. and Klee C. B.), Vol. 5, pp. 1–15. Elsevier, Amsterdam.

Yaginuma T. and Happ G. M. (1989) Trehalase from the bean-shaped accessory glands and the spermatophore of the male mealworm beetle, *Tenebrio molitor*. *Gen. comp. Endocr.* **73,** 173–185.

Zopf D., Hermans-Borgmeyer I., Gundelfinger E. D. and Betz H. (1987) Identification of gene products expressed in the developing chick visual system—characterization of a middle-molecular-weight neurofilament. *Genes Dev.* **1,** 699–708.