ELSEVIER

Biochimica et Biophysica Acta 1293 (1996) 171–176

BBA Biochimica et Biophysica Acta

Short sequence-paper

Structure of a D-protein gene and amino-acid sequences of the highly repetitive D-proteins secreted by the accessory glands of the mealworm beetle [1]

Guido C. Paesen [*,2], Xu Feng [3], George M. Happ

Department of Zoology, University of Vermont, Burlington, VT 05405-0086, USA

Received 10 November 1995; revised 10 January 1996; accepted 11 January 1996

Abstract

The D-group proteins form the major component of the proteinaceous secretion of the tubular accessory glands of the yellow mealworm beetle, *Tenebrio molitor*. In a previous paper [1], we reported the sequence of two D-protein cDNAs and their inferred translation products. Both proteins contain three highly repetitive domains (A, A′ and B). In this paper, we present the cDNA-inferred sequences of 8 more D-proteins, none of which contains an A′ domain. We also present the structure of a D-protein gene. Southern analysis suggests that genes coding for an A′ domain are relatively rare. Genes with a total of 7 or 8 (A + B domain) repeats seem most common.

Keywords: D-protein; D-gene; Spermatophore; Accessory gland; (*T. molitor*)

Two pairs of accessory sex glands are part of the reproductive apparatus of the male mealworm beetle, *Tenebrio molitor*: a pair of bean-shaped accessory glands (termed BAGs) and a pair of tubular accessory glands (TAGs). The BAGs produce many of the structural proteins that are used to construct the wall of the spermatophore, the protective sac in which the male packages its semen for delivery to the female during copulation [2]. The TAG secretion contains four groups of highly abundant proteins, termed A, B, C and D [3], which seem to form part of the seminal fluid [1,4,5]. A- and B-proteins appear to be related to each other, but only the sequence of the B-proteins is known [6]. B-proteins show significant sequence resemblance to a group of pheromone- and odorant-binding proteins in insects, suggesting a function as carrier proteins for volatile substances. The C-proteins bind heparin in vitro, but their exact role remains unclear [7], as does the function of the D-group proteins.

We already reported the cDNA-inferred sequence of two D-proteins (D1 and D2) [1]. These D-proteins have a complex repetitive structure, based on repeats and subrepeats. In this paper, we present the genomic DNA for a D-protein. We also report 8 additional, cDNA-inferred, D-protein sequences.

A genomic library was constructed by insertion of *Tenebrio* DNA in between the *Bam*HI sites of Stratagene's EMBL3 vector. The DNA was purified from TAGs [8], partially digested with *Bam*HI, and size-fractionated [9]. The recombinant DNA was packaged with the Gigapack II Gold packaging extract (Stratagene) and amplified in P2392 bacteria. Screening was done by plaque hybridization [10] with digoxygenin-labeled D2-DNA. The insert from a positive clone was digested with *Sal*I and *Xho*I, resulting in a 0.8 kb fragment and a 2 kb fragment, each containing a part of the D-gene coding region and some of the flanking sequence. The fragments were inserted into pBlueskript SK(−) plasmids (Stratagene) and sequenced [11]. Sequence data were analyzed using the Genetics Computer Group [12] sequence analysis software.

* Corresponding author.

[1] The sequence data reported in this manuscript have been submitted to the Genome Sequence Data Base under the accession numbers M95697, M83942 and M83943.

[2] Present address: NERC-Institute of Virology and Environmental Microbiology, Mansfield Road, Oxford Ox1 3SR, U.K. Fax: +44 1 865 59962.

[3] Present address: Dept. of Pathology, Washington University School of Medicine, 216 South Kingshighway, St. Louis, MO 63110, USA.

PII S0167-4838(96)00009-X

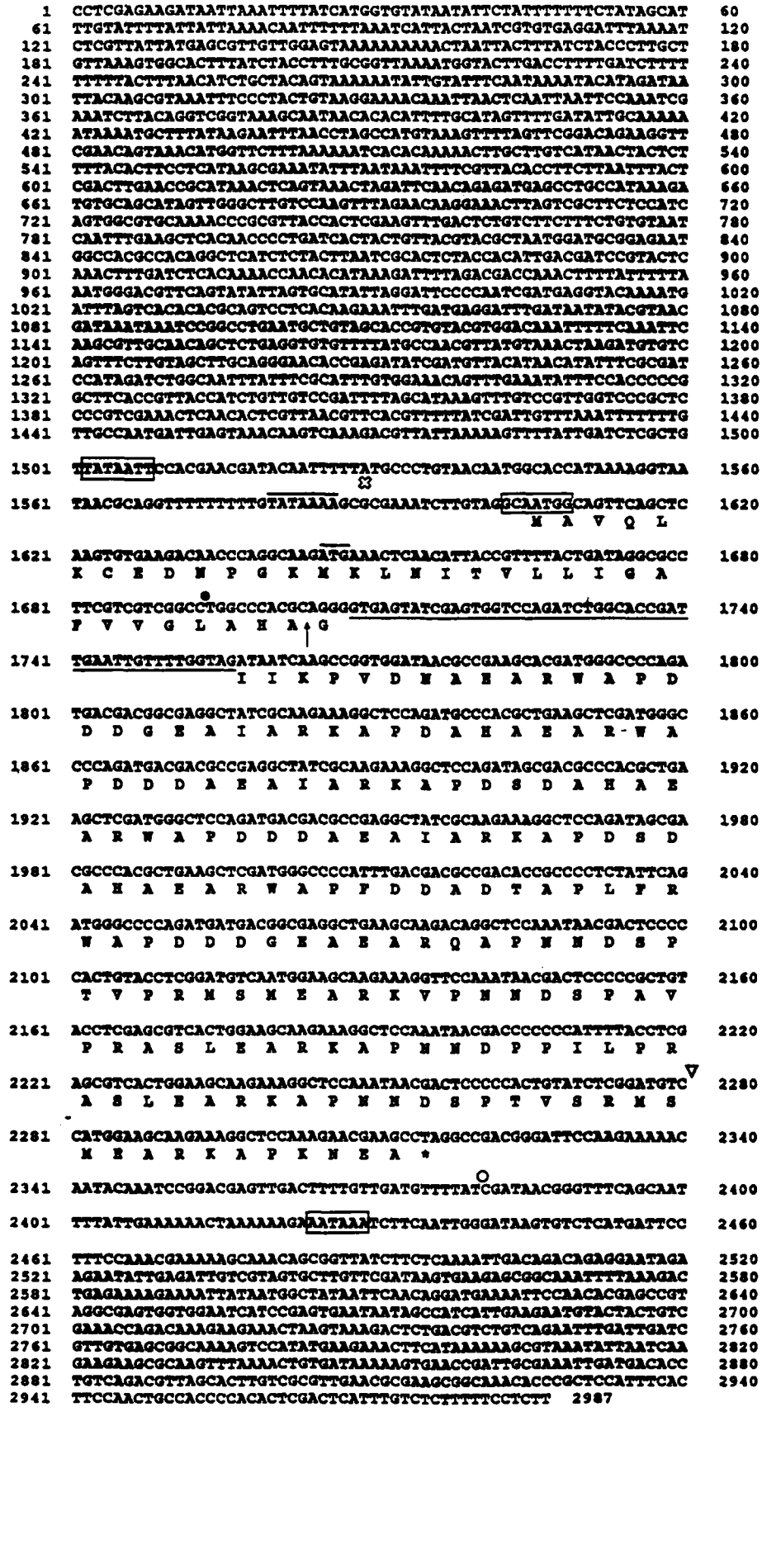

Fig. 1 shows the sequence of the D-protein gene, together with the sequence of its translation product. The open reading frame contains two possible startcodons (positions 1606–1608 and 1645–1647), the first of which corresponds with the consensus sequence for a translational initiation site [(AG)NN(AC)TGG] [13]. A possible start site for transcription is located at position 1530, which is an adenosine flanked by pyrimidines, 27 bases upstream of a putative TATA-box [14]. Alternatively, it is possible that the actual TATA-box is located in between nucleotides 1580 and 1586. If this is the case, it is unclear at which position the transcription starts, but the start site for translation would correspond with the second ATG-triplet (1645–1647), rather than with the first one. This would result in a shorter, but more typical signal sequence, i.e. one with a basic, instead of acidic, N-terminal region [15]. The signal sequence cleavage site was previously determined [1]. No relevant transcription factor sites other than the TATA-box are obvious. The D-protein gene contains a single intron, located in the region coding for the non-repetitive N-terminal domain of the mature protein (the 'unique region' in Fig. 2) [14]. The translation product of the gene is identical to one of the cDNA translation products described below ($D_{3\text{-}V}$d).

From screening of a TAG cDNA library [1] with the D2-probe and from PCR-amplification [16] with a forward D-specific primer (5'-ACGCAGGGATAATCAAGCCG-3'), constructed against the 5'-terminus of the D1- and D2-cDNAs, and a reverse, plasmid-specific primer (5'-GTTTTCCCAGTCACGACG-3'), we isolated and sequenced 8 new clones for D-proteins. The sequences obtained after PCR-amplification were all confirmed by sequencing products of additional PCR-reactions. None of the 8 cDNAs contains a startcodon. The sequences downstream of the stopcodons vary in length from 49 to 129 nucleotides, but are essentially the same as the corresponding region of the D-gene in Fig. 1.

In Fig. 2, the 8 new (cDNA-inferred) amino-acid sequences are aligned with D1 and D2. The proteins differ from one another mostly in the numbers of repeats in the A and B domains, while the actual sequences of the repeats are well conserved. We therefore named the different D-proteins according to the numbers of repeats, using Arabic numerals for the number of A and A' repeats, and Roman numerals for the number of repeats in the B domain. The previously described proteins have been renamed according to this scheme: D1 thus becomes $D_{3\text{-}3\text{-}V}$ and D2 becomes $D_{3\text{-}2\text{-}IV}$.

These two proteins provide a point of departure for comparisons among the various allelic products. The A and A' domains of $D_{3\text{-}3\text{-}V}$ and $D_{3\text{-}2\text{-}IV}$ are formed from identical repeating sequences, each of which consists of two similar halves, AEARWAPDDD(A/G)E and A(I/L)AR(K/L)APDS(E/D)AH. The A and A' domains are separated from each other, and from the B domain, by a somewhat modified repeat, termed the 'transition sequence' AEARWAPFDDADTAPLFRWAPDDD(G/A)E. The B domain repeats [(M/L/A)EAR(Q/K)(A/V)-(P/Q)NNDSP(A/T)VPR(A/V)S] appear to be derived from the same ancestor sequence as the A domain repeats and the transition sequence, but contain recognition sites for glycosylation and phosphorylation [1].

All of the newly described proteins lack the A' domain. However, the A and B domain repeats, as well as the transition sequence, are very similar to those described previously.

The four translation products of the $D_{3\text{-}V}$ group contain a B4 repeat (like $D_{3\text{-}3\text{-}V}$) and differ in only a few amino acids, mainly in the B domain. They also lack the Ser-Glu doublet in the first repeat unit of the A domain (typical for $D_{3\text{-}3\text{-}V}$ but not for $D_{3\text{-}2\text{-}IV}$). The resemblance with $D_{3\text{-}3\text{-}V}$ is especially close for $D_{3\text{-}V}$b, which not only has the same amino-acid sequence, but also an identical cDNA sequence (the A' domain left aside).

The $D_{3\text{-}IV}$a and $D_{3\text{-}IV}$b proteins can be described as $D_{3\text{-}2\text{-}IV}$ proteins without the A' domain, not only because of the numbers of repeats, but also because of the presence of the Ser-Glu doublet in the A1 repeat. $D_{3\text{-}IV}$a is identical to $D_{3\text{-}2\text{-}IV}$ in the A and B domains and differs only in the penultimate amino-acid of the transition sequence.

$D_{4\text{-}IV}$ is the only translation product that has a different number of A-domain repeats (4 instead of 3). Although it lacks the Ser-Glu doublet in the A1 repeat, it still seems more related to $D_{3\text{-}2\text{-}IV}$ than to $D_{3\text{-}3\text{-}V}$ since its B domain sequence is identical to that of $D_{3\text{-}2\text{-}IV}$.

$D_{3\text{-}III}$ is the smallest of all of the translation products, with only 3 B-domain repeats.

Although amino-acid substitutions appear to have occurred more frequently in the B than in the A domains, all of the (putative) phosphorylation and glycosylation sites (restricted to the B domain) have been conserved. Each B repeat, except the incomplete B5, has a carboxyterminal serine as a possible kinase substrate [17] and a putative N-glycosylation site [18]. The transition sequences are extremely well conserved.

To find out which D-protein genes are most common in *Tenebrio*, Southern blotting was carried out, using the digoxygenin-labeled probe [16]. Genomic DNA samples

Fig. 1. Sequence of a $D_{3\text{-}V}$ gene The signal peptidase site is marked with an arrow, the intron is underlined. The three boxes represent, from top to bottom, the putative TATA-box, the putative translational initiation signal, and the polyadenylation signal. The probable transcription start site is indicated with a cross. An alternate, putative TATA-box and a (corresponding) alternate startcodon are overlined. The asterisk marks the stopcodon. Restriction enzyme sites relevant to the Southern blots in Fig. 3, are indicated by a circle (*Bsp*106I) and a dot (*Scr*FI).

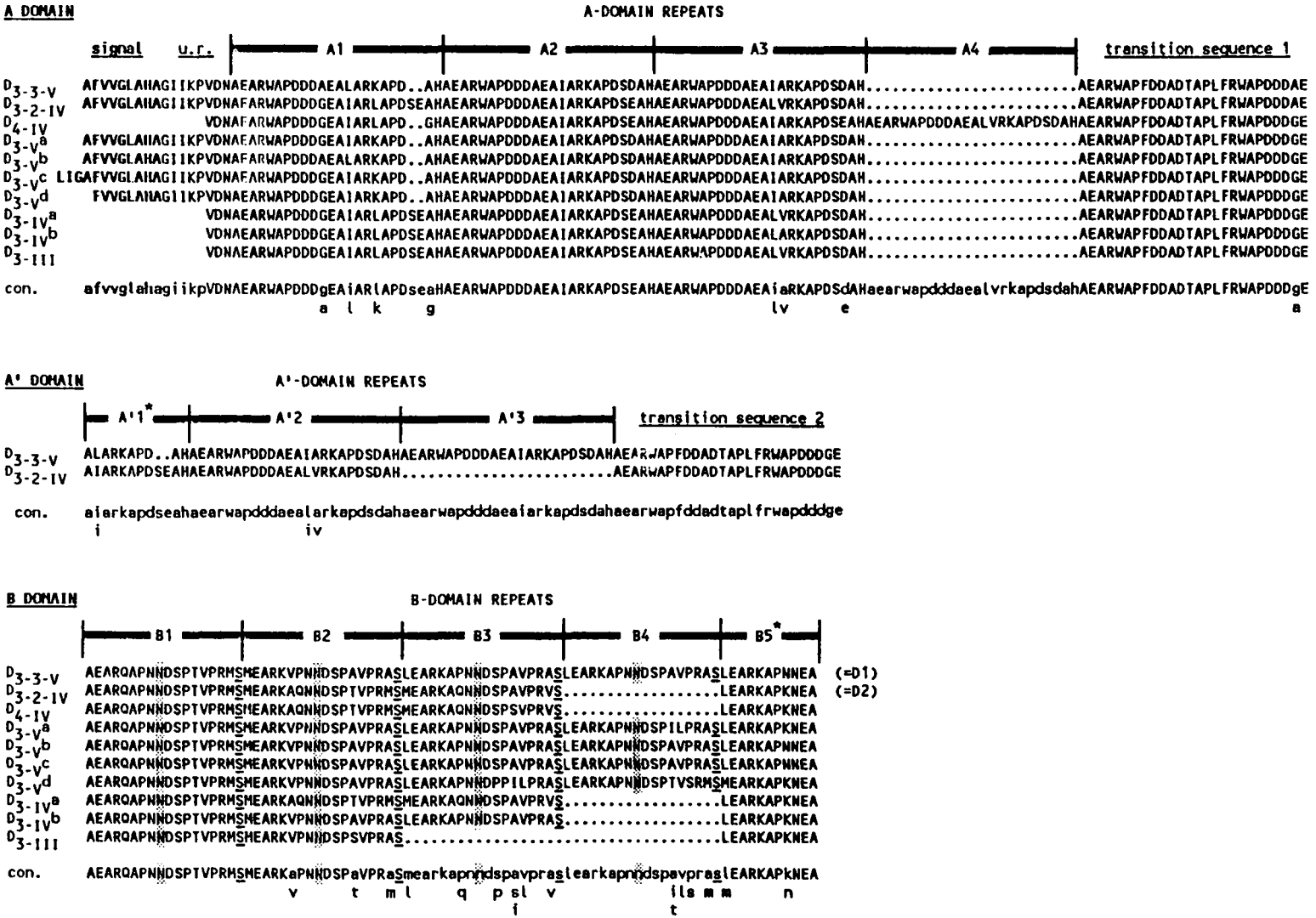

Fig. 2. cDNA-inferred sequences of D-protein variants. The repeats marked with an asterisk are shorter than the other repeats of the same domain. U.r. stands for 'unique region'. Upper case letters in the consensus sequence (con.), refer to amino acids that appear at the given position in all of the ten translation products. The putative phosphorylation sites are underlined, putative glycosylation sites are shaded.

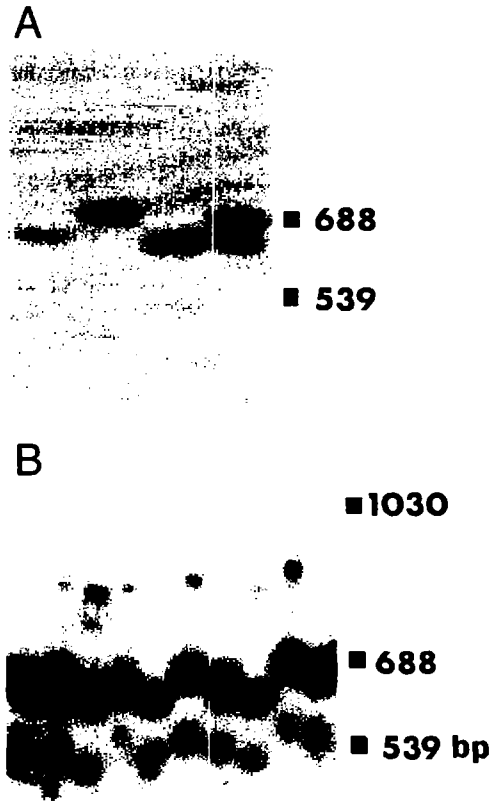

Fig. 3. Southern hybridization. (A) Each lane contains DNA from individual beetles (approx. 3 μg), double-digested with *Bsp*106I and *Scr*FI. After resolution on a 2% agarose gel, the DNA was blotted onto a nylon membrane (Hybond N+, Amersham) and hybridized with the digoxygenin-labeled probe. Activity of the anti-digoxygenin alkaline phosphatase conjugate was visualized on X-ray film (X-OMAT AR, Kodak) using the lumigen LumiPhos (Boehringer Mannheim) as a substrate. The markers were constructed from D-protein cDNAs. (B) Narrower wells were used (resulting in a higher concentration of DNA in the lanes), and the exposure time was doubled.

from individual (male) pupae were double-digested with the restriction endonucleases *Scr*FI and *Bsp*106I. *Scr*FI cuts in the region coding for the signal sequence, while *Bsp*106I cuts close to the stopcodon (note that the choice of restriction enzymes does not allow to count the number of alleles in an individual animal; it only gives an idea of the size of the D-protein genes present). Per lane (i.e., for each individual beetle), one or two bands appeared on the blot (Fig. 3A). In total, the DNA of 28 pupae was examined. In 11 pupae, only the upper band was found, in 8 other animals only the lower band was detected, and in the remaining 9 both bands appeared together. Although it is not completely clear exactly which cDNA clone(s) correspond with the two bands, the upper one corresponds with D-genes containing 8 (A + B) repeats, such as $D_{3\text{-V}}$ (of which the expected size, after cutting, is 687 bp) and/or $D_{4\text{-IV}}$ (705 bp). The lower band has approximately the size calculated for a D-gene with 7 repeats, such as $D_{3\text{-IV}}$ (639 bp). The fragments generated by the two enzymes are clearly too short to correspond with $D_{3\text{-3-V}}$ (939 bp) or $D_{3\text{-2-IV}}$ (825 bp), and too long to correspond with $D_{3\text{-III}}$ (585 bp). The Southern data thus suggest that the *Tenebrio* genome predominantly contains two (groups of) alleles, one of which has a total of 8 (A + B domain) repeats, the other 7. Individual animals have the alleles for either one, or both groups. This is in accordance with an earlier study [19], which used polyacrylamide gel electrophoresis to demonstrate that TAGs of individual beetles secrete one or two (sizes of) D-proteins. $D_{3\text{-3-V}}$, $D_{3\text{-2-IV}}$ or $D_{3\text{-III}}$ appear to be poorly represented in our *Tenebrio* population, since we never found clear bands corresponding with these alleles. Perhaps the $D_{3\text{-3-V}}$, $D_{3\text{-2-IV}}$ and $D_{3\text{-III}}$ cDNAs are artifacts (repeats might have been introduced or deleted during or after cDNA synthesis). Another possibility is that alleles containing 7 or 8 repeats are present in numerous copies, while other alleles appear in relatively low copy numbers, making them harder to detect. Indeed, when applying more concentrated DNA to the lanes, and extending the exposure to X-ray film, additional bands appeared (Fig. 3B), at positions where $D_{3\text{-3-V}}$, $D_{3\text{-2-IV}}$ and $D_{3\text{-III}}$ genes are expected to show up. However, we can not exclude the possibility that these bands are the result of incomplete digestion, or of hybridization with a fragment containing the sequence downstream of the stopcodon, of which the probe recognizes the first 40 bases. In any case, genes coding for D-proteins with an A′ domain appear to be quite rare. Similar blots were obtained with the DNA of female beetles (data not shown), indicating that the D-genes are not male-specific.

On SDS-gels, the most common D-proteins have apparent molecular masses of 27.7, 26.5 and 23.9 kDa [19]. The translation products of $D_{4\text{-IV}}$, $D_{3\text{-V}}$ and $D_{3\text{-IV}}$, seemingly the most common D-genes, have calculated masses in between 20.9 and 18.4 kDa. This discrepancy between primary and secreted proteins suggests some sort of post-translational modification, in which D-proteins are provided with covalently bound adducts, probably carbohydrates (the p*I* of the primary translation products is similar to the p*I* of the secreted proteins, suggesting that eventual adducts are uncharged). On 2D-SDS gels, the 27.7 kDa protein appears to be slightly less acidic than the two other major D-proteins. The 27.7 kDa variant may therefore be a $D_{3\text{-V}}$-protein (calculated p$I = 4.5$), while the 26.5 and 23.9 kDa proteins could be $D_{4\text{-IV}}$- and $D_{3\text{-IV}}$-products (both p$I = 4.3$). This is not necessarily in contradiction with the calculated molecular weights, which are lower for the $D_{3\text{-V}}$ translation products than for the $D_{4\text{-IV}}$-protein; $D_{3\text{-V}}$ may undergo a more extensive glycosylation, since it contains an extra B repeat (and thus an extra glycosylation site).

The function of D-proteins remains unclear. Since D-protein epitopes are found in the lumen of the spermatophore, the proteins may have a function in the storage, conservation, or maturation of the sperm cells, or they possibly affect the reproductive physiology of the female

after copulation [20]. The findings presented in this paper suggest that a fixed number of repeats is not essential for a D-protein to play its role, whatever this may be. Indeed, different individuals have proteins with different repeat compositions. Although the number of repeats may differ among the alleles, their sequence is well conserved, within each protein as well as from one protein to the other. It thus appears that each repeat is a functional entity on its own (for example a binding site).

We thank the National Institutes of Health (AI-15662) for financial support.

References

[1] Paesen, G.C., Weyda, F. and Happ, G.M. (1992) Insect Biochem. Mol. Biol. 22, 387–398.
[2] Grimnes, K.A., Bricker, C.S. and Happ, G.M. (1986) J. Exp. Zool. 240, 275–286.
[3] Happ, G.M., Yuncker, C. and Huffmire, S.A. (1977) J. Exp. Zool. 200, 223–236.
[4] Black, P.N., Landers, M.H. and Happ, G.M. (1982) Dev. Biol. 94, 106–115.
[5] Black, P.N. and Happ, G.M. (1985) Insect Biochem. 15, 639–650.
[6] Paesen, G.C. and Happ, G.M. (1995) Insect Biochem. Mol. Biol. 25, 401–408.
[7] Paesen, G.C. and Happ, G.M. (1994) Insect Biochem. Mol. Biol. 24, 21–27.
[8] Miller, S.A., Dykes, D.D. and Polesky, H.F. (1988) Nucleic Acids Res. 16, 1215.
[9] Frischauf, A.M. (1987) Methods Enzymol. 152, 556–562.
[10] Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) Molecular Cloning: A Laboratory Manual, 2nd Ed., Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
[11] Sanger, F. and Coulson, A.R. (1975) J. Mol. Biol. 94, 441–448.
[12] Devereux, J., Haeberli, P. and Smithies, O. (1984) Nucleic Acids Res. 12, 387–395.
[13] Kozak, M. (1986) Cell 44, 283–292.
[14] Breatnach, R. and Chambon, P. (1981) Annu. Rev. Biochem. 50, 349–383.
[15] Von Heijne, G. (1986) Nucleic Acids Res. 14, 4683–4690.
[16] Paesen, G.C., Schwartz, M.B., Peferoen, M., Weyda, F. and Happ, G.M. (1992) J. Biol. Chem. 267, 18852–18857.
[17] Weller, M. (1979) Protein Phosphorylation: The Nature, Function, and Metabolisms of Proteins which Contain Covalently Bound Phosphorus. Pion Limited, London.
[18] Oikawa, S., Nakazato, H. and Kosaki, G. (1987) Biochem. Biophys. Res. Commun. 142, 511–528.
[19] Grimnes, K.A. and Happ, G.M. (1985) Insect Biochem. 15, 181–188.
[20] Chen, P.S. (1984) Annu. Rev. Entomol. 29, 233–255.