Proc. Natl. Acad. Sci. USA
Vol. 84, pp. 6712–6714, October 1987
Biochemistry

The major clotting protein from guinea pig seminal vesicle contains eight repeats of a 24-amino acid domain

(protein processing/cDNA/tandem repeats/crosslinking)

JOHN T. MOORE, JAMES HAGSTROM, DANIEL J. MCCORMICK, SCOTT HARVEY, BEN MADDEN, EILEEN HOLICKY, DAVID R. STANFORD, AND ERIC D. WIEBEN*

Department of Biochemistry and Molecular Biology, Mayo Foundation, Rochester, MN 55905

Communicated by Clement L. Markert, June 17, 1987

ABSTRACT **The complete amino acid sequence of the major clotting protein from the guinea pig seminal vesicle (SVP-1) has been determined by nucleotide sequencing of cDNA clones corresponding to the 3′ terminus of an mRNA that codes for a protein precursor to SVP-1. The first 40 amino acids of the derived protein sequence are identical to those determined by N-terminal sequencing of SVP-1 isolated from the lumen of the seminal vesicle. This finding confirms that SVP-1 is cleaved from the C terminus of a larger precursor protein. The portion of the nucleotide sequence that codes for SVP-1 contains eight highly homologous but imperfect repeats of a 72-nucleotide domain. This repeated structure is also evident at the amino acid level. The consensus 24-amino acid repeat unit contains two lysine and three glutamine residues. Since the clotting of SVP-1 is known to involve the formation of γ-glutamyl-ε-lysine crosslinks, it is likely that the 24-amino acid repeating unit is the unit of function of SVP-1.**

Guinea pig seminal vesicle epithelium (GPSVE) is an androgen-dependent tissue that synthesizes four abundant secretory proteins designated SVP-1 through -4 (1). SVP-1 is a clotting protein that serves as the substrate in the formation of the copulatory plug (2, 3). Covalent clotting of this protein is catalyzed by a transglutaminase secreted by the anterior prostate (3, 4). When the crosslinking is performed *in vitro*, up to 6.6 mol of γ-glutamyl-ε-lysine are formed per mol of protein, making SVP-1 "one of the most crosslinked protein structures ever described" (3).

In the adult guinea pig, synthesis of SVP-1 represents >20% of total protein synthesis in SVE (1). We have previously presented indirect evidence that SVP-1 is cleaved from a high molecular weight precursor (5). We have now determined the N-terminal amino acid sequence of SVP-1 and the nucleotide sequence† of the portion of the pre-SVP-1 mRNA that codes for the SVP-1. Comparison of the sequences confirms the previously suspected precursor–product relationship and suggests a structural basis for the high degree of crosslinking obtained with this protein.

MATERIALS AND METHODS

Purification and Sequencing of SVP-1. SVP-1 was purified by DEAE-cellulose column chromatography as described (6). Approximately 50 μg of purified SVP-1 (2 nmol of protein) was microsequenced on an automated 470A gas-phase sequencer (Applied Biosystems, Foster City, CA). Resultant phenylthiohydantoin amino acids were identified by reversed-phase high-performance liquid chromatography on a C_{18} narrow-bore column (2.1 × 220 mm, Applied Biosystems) using a gradient of acetonitrile and sodium acetate at pH 3.6–4.8.

Isolation and Analysis of cDNA Clones. The SVP-1 cDNA clone designated M13-56B was isolated as described (5). The purified insert from this clone was used to screen a guinea pig cDNA library. The sequence of the cDNA inserts of the clone was derived by dideoxy sequence analysis (7) of both strands over the entire length of the cDNA sequence given in Fig. 1.

Homology searches of the GenBank and Protein Identification Resources databases were performed using software from DNASTAR (Madison, WI). For nucleic acid sequence searches, the NUCSCAN program was used with a hash/tuple size of 5. For amino acid sequence homology searches, the PROSCAN program was used (hash/tuple size = 2). Optimum alignments of amino acid sequence matches were performed using AALIGN, which uses the same algorithm as the Lipman and Pearson FASTP program (8).

RESULTS

cDNA clones coding for the high molecular weight precursor to SVP-1 were isolated from a GPSVE cDNA library using a short SVP-1 cDNA [pAT56B (5)] as a probe. Fig. 1 represents the compilation of DNA sequence data from five overlapping cDNA clones that encompass the 3′ half of the 1800-nucleotide (nt) precursor mRNA. The longest open reading frame of this DNA sequence codes for a basic protein with a molecular weight of 22,125. This is in good agreement with the size of mature SVP-1, as estimated by electrophoresis in the presence of $NaDodSO_4$.

The first 40 amino acids of the translated protein sequence are in perfect agreement with the N-terminal amino acid sequence of mature SVP-1, as determined by gas-phase microsequencing of the purified protein (underlined amino acid residues in Fig. 1). The isoleucine residue at position 1 is the N-terminal amino acid of mature SVP-1 from the lumen of the GPSVE. The correspondence of the sequence data identifies the translated amino acid sequence as SVP-1 and provides conclusive proof that SVP-1 is cleaved from the C terminus of a precursor protein.

The 609-nt coding portion of the nucleotide sequence that codes for SVP-1 contains eight and a half tandem repeats of a 72-nt sequence. The pattern of repeats is best illustrated by alignment of the sequence in 72-nt blocks (Fig. 2). This alignment reveals that there is sufficient homology between the repeats to derive a consensus sequence that is unambiguous in all but five positions. Twenty-six nucleotides are invariant between all of the repeats (underlined residues in

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: SVE, seminal vesicle epithelium; GPSVE, guinea pig SVE; nt, nucleotide(s).

*To whom reprint requests should be addressed.

†This sequence is being deposited in the EMBL/GenBank data base (Bolt, Beranek, and Newman Laboratories, Cambridge, MA, and Eur. Mol. Biol. Lab., Heidelberg) (accession no. J02968).

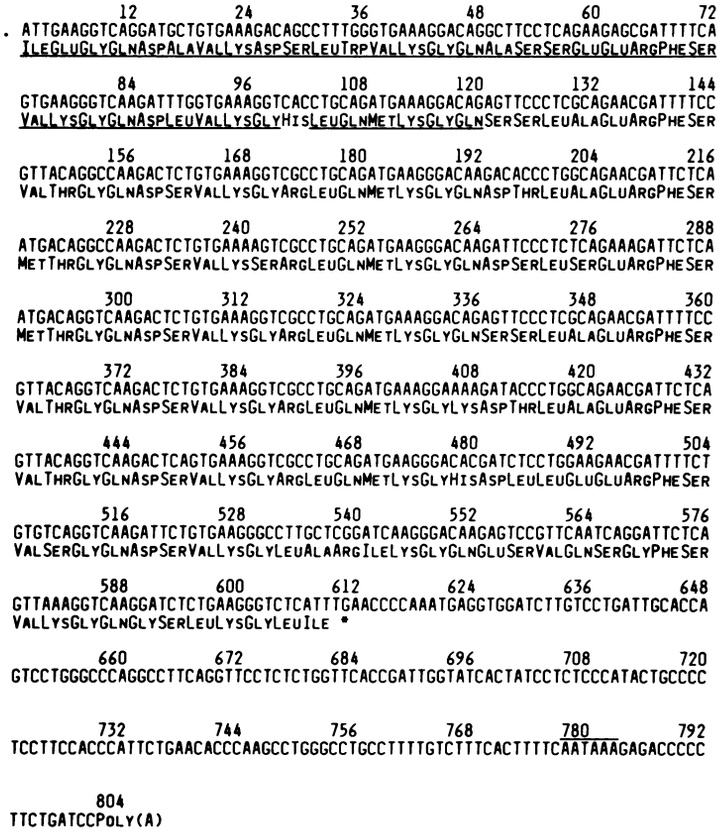

FIG. 1. Complete nucleotide and derived amino acid sequence of SVP-1. The sequence represents a compilation of data from five overlapping cDNA cones. The sequence of the first 40 amino acid residues of SVP-1 was independently determined by automated Edman degradation. (The histidine at position 34 was not identified by sequencing of the protein.) The * denotes the termination codon at the 3′ terminus of the SVP-1 coding sequence. A consensus poly(A) sequence (AATAAA) (9) is overlined.

Fig. 2). This degree of homology is not shared by subsequent 72-base blocks of the cDNA sequence from the 3′ untranslated portion of the mRNA. Thus, the conservation of the repeating nucleotide sequence may reflect selective pressure based upon the protein coding potential of this region.

As expected from the repetitive nature of the DNA sequence, the derived amino acid sequence also has significant repetitive structure (Fig. 3). Once again, there are few variations from a generally unambiguous consensus sequence. Given the known mechanism of clotting of this protein, it is particularly interesting that each repeat unit contains at least two lysines and two glutamines. Furthermore, out of 7 of the 24 amino acids in each repeat that are invariant, 2 are lysines and 1 is a glutamine. There is apparently significant selective pressure to preserve these residues, which are directly involved in the formation of crosslinks.

Inspection of the amino acid sequence reveals that there is an underlying substructure to the basic 24-amino acid repeat. Specifically, the sequence Gly-Gln-Asp-Ser appears twice within the 24-amino acid consensus sequence. However, from the alignment shown in Fig. 2, it is clear that the functional unit of selection has been the 72-nt repeat rather than a 36-nt subunit. Thus, the significance of the tetrapeptide substructure remains unclear.

DISCUSSION

Williams-Ashman *et al.* (3) managed to obtain six or seven crosslinks per mol of SVP-1 by treatment of the purified protein with a transglutaminase ("vesiculase") isolated from the prostate. Using their estimate of molecular weight (17,900), they noted that this corresponded to approximately one crosslink every 23 amino acids. These figures are given added significance by our finding that SVP-1 contains eight

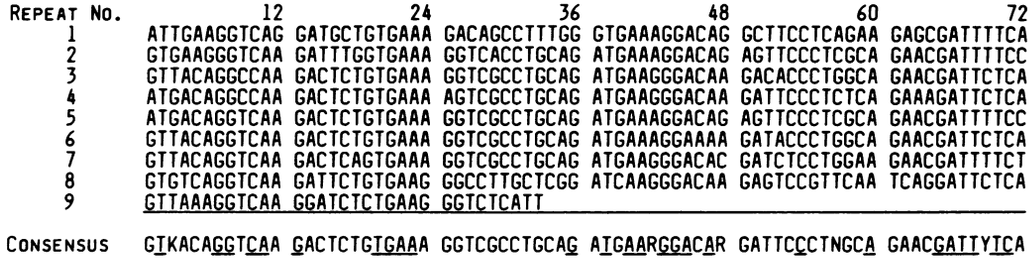

FIG. 2. Alignment of the 72-nt repeats from the SVP-1 coding sequence. The nucleotide sequence of SVP-1 has been aligned in consecutive 72-nt segments to emphasize the homology between successive 72-nt domains. A unique nucleotide was assigned to a given position in the consensus sequence if that nucleotide occurred in at least five of the eight complete repeats. Nucleotides that are invariant in all of the repeats have been underlined in the consensus sequence.

Repeat No.	Residue 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	I	E	G	Q	D	A	V	K	D	S	L	W	V	K	G	Q	A	S	S	E	E	R	F	S
2	V	K	G	Q	D	L	V	K	G	H	L	Q	M	K	G	Q	S	S	L	A	E	R	F	S
3	V	T	G	Q	D	S	V	K	G	R	L	Q	M	K	G	Q	D	T	L	A	E	R	F	S
4	M	T	G	Q	D	S	V	K	S	R	L	Q	M	K	G	Q	D	S	L	S	E	R	F	S
5	M	T	G	Q	D	S	V	K	G	R	L	Q	M	K	G	Q	S	S	L	A	E	R	F	S
6	V	T	G	Q	D	S	V	K	G	R	L	Q	M	K	G	K	D	T	L	A	E	R	F	S
7	V	T	G	Q	D	S	V	K	G	R	L	Q	M	K	G	H	D	L	L	E	E	R	F	S
8	V	S	G	Q	D	S	V	K	G	L	A	R	I	K	G	Q	E	S	V	Q	S	G	F	S
9	V	K	G	Q	G	S	L	K	G	L	I	*												
Consensus	V	T	G	Q	D	S	V	K	G	R	L	Q	M	K	G	Q	D	S	L	A	E	R	F	S

FIG. 3. Conservation of the amino acid sequence in successive 24-amino acid repeats from the SVP-1 sequence. The amino acid sequence of SVP-1 (using the one-letter abbreviations for the amino acids) has been aligned to maximize homology between successive repeats. The consensus sequence contains those residues that occur in at least four of the repeats. Residues that correspond to the consensus sequence are shown in large type. The residues in the consensus sequence that are invariant between all of the repeats have been underlined.

repeats of a 24-amino acid domain. The suggestion from these data is that, *in vitro*, one crosslink forms for every 24-amino acid repeat. If these estimations are an accurate reflection of the degree of crosslinking achieved *in vivo*, then the conserved lysine and glutamine residues within the 24-amino acid repeat stand out as the likely residues to be involved in the intermolecular interaction.

It is notable that all but 1 of the 26 glycine residues in SVP-1 are adjacent to either a glutamine or a lysine residue. Glycines also account for 2 of the 7 invariant residues in each 24-amino acid repeat. This nonrandom arrangement of glycine in the sequence of SVP-1 suggests a functional role for glycine in the clotting process. One obvious possibility is that glycine causes minimal steric interference to the formation of crosslinks between neighboring lysines and glutamines.

It is interesting in this regard that the derived amino acid sequence of SVP-1 has some sequence homology to the α_1 chain of collagen [human (10), murine (11), and bovine (12)]. Although the similarity index (9) is 7 SD above the mean of comparisons between SVP-1 and other sequences in the protein database, most of the homology is based upon the high glycine and lysine content of SVP-1. The high glycine content of collagen is thought to be directly related to the structural constraints imposed by the formation of the collagen triple helix. However, SVP-1 differs markedly from collagen in that it contains no proline or tyrosine and is crosslinked by γ-glutamyl-ε-lysine crosslinks rather than the allysine- or hydroxyallysine-based linkages found in collagen. This consideration would predict that SVP-1 might have a more meaningful homology to fibrinogen, which forms the same type of crosslinks as SVP-1 (13). Thus, it is worth noting that the best homology to the nucleic acid sequence of SVP-1 in the December 1986 update of GenBank was to the rat β-fibrinogen gene (14). Although these sequences are 39% homologous over the length of the SVP-1 sequence, the longest contiguous match is only 6 nt. Thus, there is little basis for attempting to identify functionally conserved domains between the two sequences.

The data presented here provide unequivocal confirmation of the hypothesis that SVP-1 is derived from the C terminus of a high molecular weight protein precursor (5). We cannot rule out the possibility that the production of mature SVP-1 involves the removal of several amino acids from the C terminus of the protein sequence shown in Fig. 1. However, as noted earlier, the predicted size of the complete protein given in Fig. 1 is in good agreement with our previous estimates of the size of mature SVP-1. The continued sequence conservation in the ninth repeat also suggests that this region of the precursor protein is still subject to selective pressure. Original estimates of the size of the pre-SVP-1 precursor suggested that it had a molecular weight of ≈55,000. Further analysis suggests that this precursor is actually somewhat smaller, with a molecular weight of 43,500. Since this pre-SVP-1 species also contains a 1500-dalton signal peptide (M. E. Norvitch, S.H., and E.D.W., unpublished data), it must contain no more than one copy of the SVP-1 sequence. Since the GPSVE secretes four major secretory proteins in approximately equimolar amounts (15), but synthesizes only two major mRNA species (5, 16), it is possible that one of the other three secretory proteins is derived from the N terminus of the SVP-1 precursor. Further sequencing of the three other secretory proteins will be required to test the hypothesis.

We are grateful to the late Dr. Carlo Veneziale for his preparation of the SVP-1 used in the protein sequencing experiments. This work was supported by National Institutes of Health Grant HD 9140-P3 to E.D.W. and in part by the Mayo Foundation.

1. Veneziale, C. M. (1977) *Biochem. J.* **166,** 155–166.
2. Notides, A. C. & Williams-Ashman, H. G. (1967) *Proc. Natl. Acad. Sci. USA* **58,** 1991–1995.
3. Williams-Ashman, H. G., Notides, A. C., Pabalan, S. S. & Lorand, L. (1972) *Proc. Natl. Acad. Sci. USA* **69,** 2322–2325.
4. Williams-Ashman, H. G. (1984) *Mol. Cell. Biochem.* **58,** 51–61.
5. Moore, J. T., Norvitch, M. E. & Veneziale, C. M. (1985) *J. Biol. Chem.* **260,** 3826–3832.
6. Veneziale, C. M. & Deering, N. C. (1976) *Andrologia* **8,** 73–82.
7. Sanger, F., Nicklen, S. & Coulsen, A. R. (1977) *J. Mol. Biol.* **113,** 237–251.
8. Lipman, D. J. & Pearson, W. R. (1985) *Science* **227,** 1435–1441.
9. Proudfoot, N. J. & Brownlee, G. G. (1976) *Nature (London)* **263,** 211–214.
10. Babel, W. & Glanville, R. W. (1984) *Eur. J. Biochem.* **143,** 545–556.
11. Fietzek, P. P., Allman, H., Rauterberg, J., Henkel, W., Wachter, E. & Kuhn, K. (1979) *Hoppe-Seyler's Z. Physiol. Chem.* **360,** 809–820.
12. Balian, G., Click, E. H. & Bornstein, P. (1971) *Biochemistry* **10,** 4470–4478.
13. Lorand, L. (1983) *Ann. N.Y. Acad. Sci.* **408,** 226–232.
14. Fowlkes, D. H., Mullis, N. T., Comeau, C. H. & Crabtree, G. R. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 2313–2316.
15. Veneziale, C. M. & Deering, N. C. (1976) *Andrologia* **8,** 73–82.
16. Moore, J. T., Norvitch, M. E., Wieben, E. D. & Veneziale, C. M. (1984) *J. Biol. Chem.* **259,** 14750–14756.