Structural analysis of sulphated glycoprotein 2 from amino acid sequence

Relationship to clusterin and serum protein 40,40

James K. TSURUTA,* Kathy WONG,† Irving B. FRITZ† and Michael D. GRISWOLD*‡

*Biochemistry/Biophysics Program, Washington State University, Pullman, WA 99164-4660, U.S.A.,
and †Banting and Best Department of Medical Research, University of Toronto, 112 College Street, Toronto, Ont.,
Canada M5G 1L6

Sulphated glycoprotein 2 (SGP-2) is the major secreted protein product of rat Sertoli cells; likewise, clusterin is a major constituent of ram rete testis fluid. Isolation and sequencing of the intact subunits and peptides derived from clusterin show that it is the ram homologue of rat SGP-2. Human serum protein 40,40 (SP-40,40), a component of the SC5b-9 complex of complement, has recently been reported to be the human homologue of rat SGP-2. Analysis of the amino acid sequences of rat SGP-2 and human SP-40,40 show that both of these proteins have a significant relationship to the heavy chain of myosin. The regions of highest sequence similarity correspond to the major amphipathic domains in SGP-2/SP-40,40 and the long α-helical-tail domain of myosin, which forms a rod-like structure. SGP-2 has anomalous sedimentation behaviour which indicates that it probably exists in an extended conformation. A putative dinucleotide-binding structure has been identified in the longest stretch of identity between SGP-2 and SP-40,40. Elucidation of these features of SGP-2 and SP-40,40 may help to direct future studies into the role of these proteins in the reproductive and complement systems.

INTRODUCTION

The process of spermatogenesis, the development of spermatozoa from less differentiated germinal cells, occurs in the seminiferous tubules of the testes. In the seminiferous tubule, germinal cells undergo various stages of cytodifferentiation within an architectural structure generated primarily by Sertoli cells [for reviews, see Fawcett (1975) and Clermont (1972)]. As a consequence of the close physical association between Sertoli cells and developing germinal cells, it is postulated that Sertoli cells, which are secretory in nature, provide essential physical and biochemical support for the process of spermatogenesis [for reviews, see Fritz (1978) and Griswold (1988)].

Sulphated glycoprotein 2 (SGP-2), previously called 'dimeric acidic glycoprotein', has been identified as the major glycoprotein secreted by rat Sertoli cells in culture, comprising over 40 % of all protein released (Sylvester *et al.*, 1984). It is a heterodimer, composed of two disulphide-linked subunits of approx. 34 and 47 kDa respectively, as judged by mobility on SDS/PAGE (Sylvester *et al.*, 1984). The complete amino acid sequence of SGP-2 has been determined from the base sequence of a cDNA for the complete mRNA transcript, and also from *N*-terminal sequencing of purified subunits and fragments of SGP-2 (Collard & Griswold, 1987). SGP-2 contains 23.7 % (w/w) *N*-linked carbohydrate, which is heavily sulphated (Griswold *et al.*, 1982). Immunofluorescent staining demonstrated that SGP-2 is secreted into the lumen of the tubule and that it associates with the surface of spermatozoa (Sylvester *et al.*, 1984).

Clusterin, a glycoprotein isolated from ram rete testis fluid (RTF), which elicits the aggregation of a wide variety of cells (Fritz *et al.*, 1983, 1984), is synthesized by Sertoli cells (Blaschuk *et al.*, 1983; Blaschuk & Fritz, 1984; Rosenior *et al.*, 1987) and by rete testis epithelial cells in culture (Rosenior *et al.*, 1987).

As previously noted, the rat Sertoli-cell product, SGP-2, and the ram Sertoli-cell product, clusterin, share many properties, including subunit molecular masses, extent and type of glycosylation, tendency to aggregate and distribution in the reproductive tract (Griswold *et al.*, 1982; Blaschuk *et al.*, 1983; Blaschuk & Fritz, 1984; Sylvester *et al.*, 1984; Tung & Fritz, 1985; Rosenior *et al.*, 1987). These similar properties prompted our investigation into the degree of similarity between the primary structure of rat SGP-2 and that of ram clusterin. The results of *N*-terminal sequencing of purified subunits and proteolytic fragments of clusterin are presented and demonstrate that SGP-2 and clusterin are protein homologues found in the male reproductive systems of the rat and the ram respectively.

A normal human serum glycoprotein, SP-40,40, is a disulphide-linked heterodimer with subunit molecular masses of 40 kDa each. SP-40,40 has been established to be a member of the human complement system by the direct demonstration of its presence within the SC5b-9 complex of complement (Murphy *et al.*, 1988). SP-40,40 may be a multifunctional protein (Kirszbaum *et al.*, 1989). Its complete amino acid sequence has been deduced from sequencing of a cDNA coding for its full-length mRNA transcript and from protein sequencing of the *N*-termini of its purified subunits as well as of a number of proteolytic fragments; SP-40,40 has 76.8 % amino acid sequence identity with rat SGP-2 (Kirszbaum *et al.*, 1989); thus SP-40,40 is the human counterpart of rat SGP-2 and ram clusterin.

The availability of the full-length protein sequences of both rat SGP-2 and human SP-40,40 prompted a more extensive scrutiny of known protein sequences than had been performed in the past.

Abbreviations used: ABI, Applied Biosystems Inc.; LDH, lactate dehydrogenase; MWKW, myosin heavy chain from *Caenorhabditis elegans*; NBRF, National Biomedical Research Foundation; PIR, Protein Identification Resource; PTH, phenylthiohydantoin; RTF, rete testis fluid; SGP-2, sulphated glycoprotein 2; SP-40,40, serum protein 40,40; TFA, trifluoroacetic acid; UWGCG, University of Wisconsin Genetics Computer Group; VADMS, Visualization, Analysis and Design in the Molecular Sciences; V8 proteinase, *Staphylococcus aureus* V8 proteinase.

‡ To whom correspondence should be sent.

This search of known protein sequences was done in hope of elucidating possible structural or functional domains common to both SGP-2 and SP-40,40. The results to be presented indicate that a significant distant relationship exists between these two proteins and the heavy chain of myosin from the slime mould *Dictyostelium discoideum*. A putative nucleotide-binding sequence has been tentatively identified in the sequences of both SGP-2 and SP-40,40, as well as four myosin-like domains that are predicted to form amphipathic helices.

METHODS

Immunoaffinity purification of ram clusterin

Clusterin was isolated from ram rete testis fluid by immunoaffinity chromatography, using slight modifications of procedures previously described (Blaschuk *et al.*, 1983). Briefly, 11 mg of purified monoclonal IgG against clusterin, prepared from ascites fluid of mice inoculated with a hybridoma called HCn-17 (Blaschuk *et al.*, 1983), was allowed to react with 3.5 ml of CNBr–Sepharose 4B (Pharmacia) in coupling buffer (0.10 M-$NaHCO_3$, pH 8.3, in 0.5 M-NaCl). After the IgG had been covalently linked, the gel mixture was incubated with coupling buffer containing 0.2 M-diethanolamine for 2 h at room temperature. The mixture was then washed alternately for three cycles with 50 mM-sodium acetate, pH 4.0, containing 0.5 M-NaCl, and 50 mM-Tris, pH 7.5, containing 0.5 M-NaCl. The IgG-conjugated gel was packed into a column and conditioned by applying 3 mg of purified BSA (Boehringer-Mannheim), followed by elution and washing with two cycles of buffers at pH 7.5, 4.0 and 2.5. RTF, kindly provided by Dr. M. Courot, Nouzilly, France, was mixed end-over-end with the gel overnight at 4 °C. For each run, 4 ml of RTF (0.7 mg of protein/ml) was added to 4 ml of gel, with approx. 3 mg of IgG conjugated/ml of gel. The gel was packed into a column, and washed with 50 ml of Tris at pH 7.5; clusterin was eluted with 20 ml of 50 mM-acetate buffer, pH 4.0, containing 0.5 M-NaCl, followed by 50 ml of 50 mM-glycine buffer at pH 2.5, containing 0.5 M-NåCl. The eluant was dialysed against 50 mM-Tris buffer, pH 7.5, then freeze-dried. The material was reconstituted in water at a concentration of 2 mg/ml or higher. Samples were checked for potency in the aggregation assay (Fritz *et al.*, 1983) and subjected to SDS/PAGE in the presence or absence of mercaptoethanol. After staining, the gel run under non-reducing conditions revealed a single silver-stained peak at approx. 80 kDa, and the gel run under reducing conditions showed doublet peaks at approx. 40 kDa. The immuno-affinity-purified material was finally subjected to h.p.l.c. fractionation, using a C3 reverse-phase column in a Beckman apparatus. A single sharp peak was collected, using an acetonitrile/TFA gradient (Griswold *et al.*, 1982), dialysed against 20 mM-Tris, pH 7.5, and freeze-dried.

Reduction/carboxymethylation and h.p.l.c. purification of clusterin subunits

Approx. 0.75 mg of the h.p.l.c.-fractionated, immuno-affinity-purified clusterin was reduced and carboxymethylated to allow the separation of the disulphide-linked subunits. Briefly, a 1.5 mg/ml solution of ram clusterin was adjusted to 6 M-guanidinium chloride/200 mM-Tris/HCl/2 mM-EDTA, pH 8.0, in a total volume of 0.5 ml. This solution was made 117 mM in dithiothreitol (Calbiochem), purged with argon, and reduction was allowed to occur with gentle rocking at room temperature for 18 h; subsequently the reaction temperature was raised to 37 °C for an additional 2 h. The reduced clusterin solution was cooled on ice and sufficient iodoacetic acid was added in the dark to provide a 10 mM excess of iodoacetic acid over the original dithiothreitol concentration. Carboxymethylation was allowed to proceed for 1 h in the dark at 0 °C with gentle rocking.

The carboxymethylated clusterin subunits were purified for gas-phase Edman degradation with the use of a Vydak C4 (214TP; 10 μm particle size; column dimensions 250 mm × 4.6 mm) reverse-phase column in a Beckman apparatus. The carboxymethylation reaction mixture was adjusted to 24 % acetonitrile and was loaded on to the C4 column at 24 % B [solvent A is 0.1 % (v/v) TFA in water; solvent B is 0.1 % (v/v) TFA in acetonitrile] using a flow rate of 1 ml/min. The carboxymethylated clusterin subunits were eluted with a gradient of 24–55 % solvent B over a period of 50 min at a flow rate of 1 ml/min. Two major protein peaks were eluted at 42 and at 48 % B; they were designated 'Cln-N' and 'Cln-C' respectively. The eluant fractions corresponding to these peaks in the A_{214} profile were pooled and freeze-dried in a Savant Speed-Vac concentrator. Before freeze-drying portions were removed from the eluant pools for analysis by SDS/PAGE; under reducing conditions both Cln-N and Cln-C appeared as single silver-stained bands with an approximate molecular mass of 40 kDa. Protein samples were stored at −20 °C until gas-phase sequencing was initiated.

Preparation and purification of *Staphylococcus aureus*-V8-proteinase-derived peptides

V8 proteinase (Boehringer-Mannheim) was incubated with carboxymethylated Cln-N at an enzyme/substrate ratio of 1:20 (w/w). Digestion was carried out according to the manufacturer's protocols for specific cleavage at the carboxy side of glutamate residues. Peptides were purified using a Vydak C18 column (218TP; 10 μm, 250 mm × 4.6 mm) with the above-mentioned solvent system.

Gas-phase protein sequencing

Standard protocols from the manufacturer were utilized for all aspects of gas-phase sequencing. Gas-phase Edman chemistry was performed in an Applied Biosystems 470-A protein sequencer; on-line analysis of PTH amino acid derivatives was performed with an Applied Biosystems 120-A PTH analyser utilizing a Spheri-5 PTH column (5 μm; 220 mm × 2.1 mm); instrument control, data acquisition and data analysis were performed with an Applied Biosystems 900-A controller/data-aquisition module.

Computer sequence analysis

All computing services, software and consultation were provided and managed by VADMS laboratory, a molecular-sciences computer resource facility located at Washington State University supported by the Washington State University Graduate School, Academic Computing Services and the National Institutes of Health. Software was run on a VAX 11/785 computer (Digital Equipment Corp.); the GAP, COMPARE and DOTPLOT programs are contained in the University of Wisconsin Genetics Computer Group Sequence Analysis Software Package (UWGCG) (Devereux *et al.*, 1984); the RELATE, FASTP and SEARCH programs are contained in the Protein Identification Resource (PIR), which utilized version 18.0 of the NBRF/PIR database (George *et al.*, 1986); Macromodel software was created by Professor Clark Still, Department of Chemistry, Columbia University, New York, NY 10027, U.S.A.

RESULTS

Reversed-phase h.p.l.c. purification of carboxymethylated clusterin subunits

H.p.l.c.-fractionated immuno-purified ram clusterin was de-

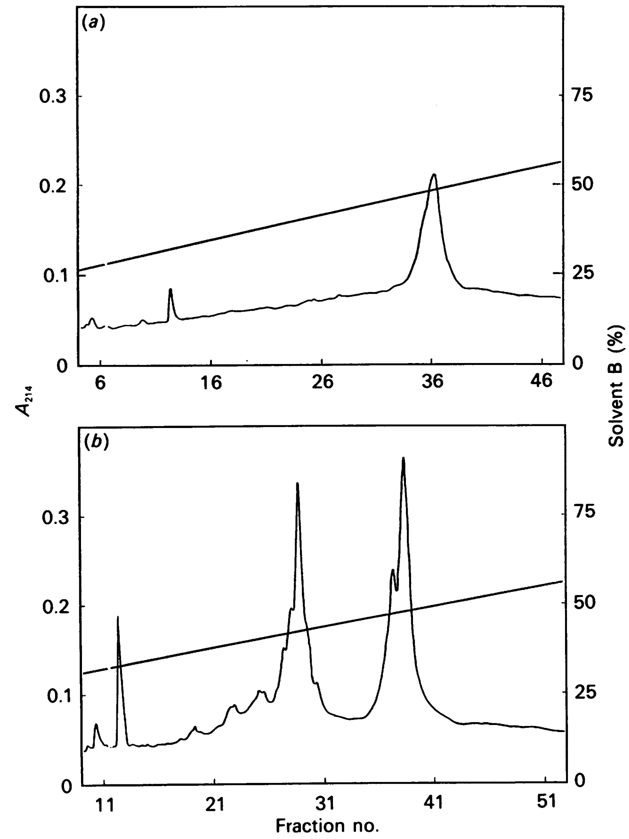


Fig. 1. Elution profile from reverse-phase h.p.l.c. separation of C3 fractionated/immuno-purified ram clusterin and carboxymethylated ram clusterin subunits

(*a*) Before reductive carboxymethylation of the C3-fractionated/immunopurified clusterin, a small sample of this protein was solubilized in 200 μl of 24 % solvent B. This solution was loaded on to a Vydak C4 reversed-phase column (214TP, 10 μm, 250 mm × 4.6 mm) at 24 % solvent B at a flow rate of 1 ml/min. The protein was eluted using a linear gradient of 24–55 % solvent B over a period of 50 min at a flow rate of 1 ml/min and a fraction size of 1 min/fraction. (*b*) After reductive carboxymethylation of ram clusterin the reaction mixture was adjusted to 24 % solvent B. Column-loading and protein-elution conditions were identical with those described above.

termined to be of sufficient purity to allow its use in the present study without further purification. A small portion of h.p.l.c.-fractionated immunopurified clusterin was subjected to reversed-phase h.p.l.c. analysis utilizing a Vydak C4 reversed-phase column (see the legend Fig. 1 for column loading and elution conditions). The A_{214} profile of this analysis is shown in Fig. 1(*a*); it can be seen that the vast majority (> 93 %) of the eluted protein is contained within a single symmetrical peak centred at approx. 47 % solvent B.

The extensive reduction and subsequent carboxymethylation of h.p.l.c.-fractionated immunopurified clusterin was undertaken to disrupt the disulphide bonds that aid in the maintenance of the subunit structure of native clusterin. The two subunits of clusterin are readily separated using reversed-phase h.p.l.c. over an analytical Vydak C4 column and have been designated Cln-N and Cln-C (see Fig. 1*b*). Cln-N is eluted at approx. 42 % solvent B, and Cln-C is eluted at approx. 48 % solvent B. The shoulder preceding Cln-C in the elution profile may represent unreduced clusterin. This would agree with the results of Collard & Griswold (1987), which indicate that rat SGP-2 is very difficult to reduce completely into separate subunits.

N-Terminal protein sequence of purified Cln-N, Cln-C and V8-proteinase-derived peptides

Gas-phase Edman degradation of the carboxymethylated clusterin subunit designated as Cln-N yielded the following protein sequence:

ISGKELQEMSTEGSKYVNKEIKN

A number of the V8-proteinase-generated peptides from carboxymethylated Cln-N were also subjected to gas-phase Edman degradation. The resulting sequences are identified in Fig. 2 by the presence of an overbar. These peptide sequences allowed the *N*-terminal sequence of Cln-N to be extended to a total length of 36 residues. An additional internal fragment with the sequence:

QGREQSSVM

was sequenced from Cln-N; it shows 60.0 % and 72.7 % identity with SP-40,40 and SGP-2 respectively. Cln-C yielded the following sequence:

NVMPFPLLEPLNFHDVFQPFY

The sequences derived from Cln-N and Cln-C were compared with the sequences contained in the NBRF/PIR database (release 18.0; George *et al.*, 1986) with the use of the algorithms FASTP (Lipman & Pearson, 1985) and SEARCH (Dayhoff *et al.*, 1983).

```
                                         ↓
SGP-2N     22  EQEFSDNELQELSTQGSRYVNKEIQNAVQGVKHIKTLIE
                  :|: ||||:||:||:|||||| ||: :| :|| ::|
Cln-N      1      ISGKELQEMSTEGSKYVNKEIKNALKEVLQIKLVME
                  :|: |||||| :||||||||| ||: :| ||| ::|
SP-40,40   1   DQTVSDNELQEMSNQGSKYVNKEIQNAVNGVKQIKTLIE
```

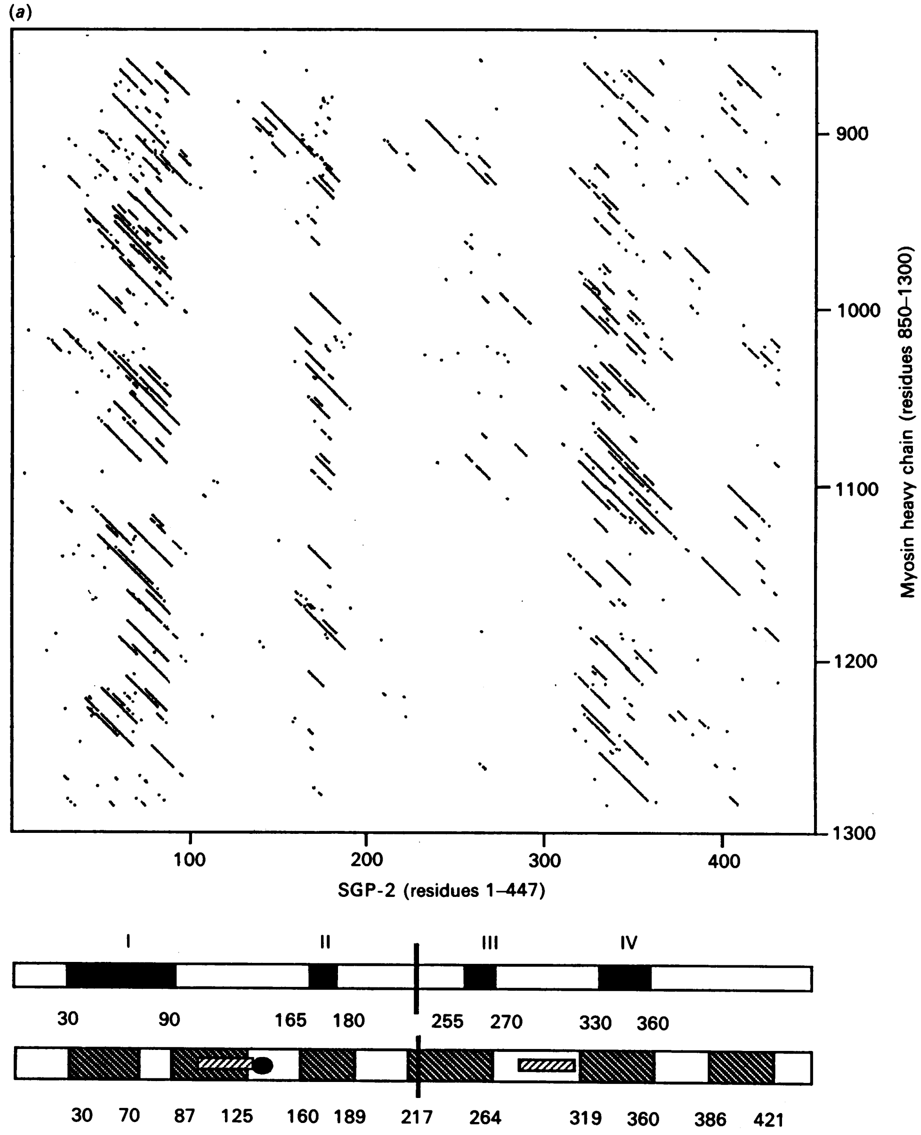
Parameter	SGP-2	SP-40,40
Similarity, identity (%)	83.8, 56.8	81.1, 63.9
RELATE score (S.D.)	11.69	9.3
Probability	$<6\times10^{-32}$	$<8\times10^{-21}$

```
SGP-2C       227   SLMPLSHYGPLSFHNMFQPFF
                    :||:   :|| ||::||||:
Cln-C        1     NVMPFPLLEPLNFHDVFQPFY
                    :|||   ||||||  :||||
SP-40,40N    206   SLMPFSPYEPLNFHAMFQPFL
```

Parameter	SGP-2	SP-40,40
Similarity, identity (%)	72.7, 45.5	68.2, 59.1
RELATE score (S.D.)	9.5	13.9
Probability	$<1.1\times10^{-21}$	$<3\times10^{-44}$

Fig. 2. Comparison of clusterin sequences with rat SGP-2 and human SP-40,40 precursor sequences

| represent identities and : denotes conserved replacement of amino acid residues between the clusterin sequences and the SGP-2/SP-40,40 sequences; overbars indicate extension of the Cln-N sequence via sequencing of V8-proteinase-derived peptides. ↓ indicates the final residue of the original *N*-terminal sequencing of Cln-N. *N*-Terminal sequences of rat SGP-2 and human SP-40,40 subunits are listed with the numbering convention of Collard & Griswold (1987) and Kirszbaum *et al.* (1989) respectively. The suffix *N* or *C* is appended to the precursor name to indicate the origin of the subunit with respect to the uncleaved precursor molecule. The RELATE comparison algorithm of Dayhoff *et al.* (1983) was used to determine the probability that the similarity observed between the clusterin sequences and the indicated sequences arose solely by chance. The RELATE analysis was performed with complete precursor sequences, the segment length was 15 residues and 100 random comparisons were performed.



(*b*)

Sequence alignments of *C. elegans* myosin heavy chain (upper line) and rat SGP-2 (lower line)

```
1684    FNAEKRATLLQSEKEE          1732    VSSLTSAKRKLEGEIQA
        :|:::   ||:|::::                  : ||  ||:| || ::
153     MNGDRIDSLLESDRQQ          70      LNSLEEAKKKKEGALDD

        Similarity 81.25%                 Similarity 64.7%
```

Sequence alignments of *D. discoideum* myosin heavy chain (upper line) and rat SGP-2 (lower line)

```
1337    VTEAKNKKESELDEIK          2013    ETELKEYRKK
        : ||| |||: ||: :                  |  | |||:|
73      LEEAKKKKEGALDDTR          434     EKALQEYRRK

        Similarity 75%                    Similarity 70%
```

Fig. 3. (*a*) Graphic analysis of the regions of similarity that exist between rat SGP-2 and myosin heavy chain from *C. elegans* (MWKW); (*b*) representative sequence alignments between SGP-2 and the myosin heavy chain from *C. elegans* and *D. discoideum*

(*a*) This graphic matrix illustrates amino acid sequence similarities between rat SGP-2 and the tail domain of *C. elegans* myosin heavy chain. The diagram was prepared by using the COMPARE and DOTPLOT programs from the University of Wisconsin Genetics Computer Group

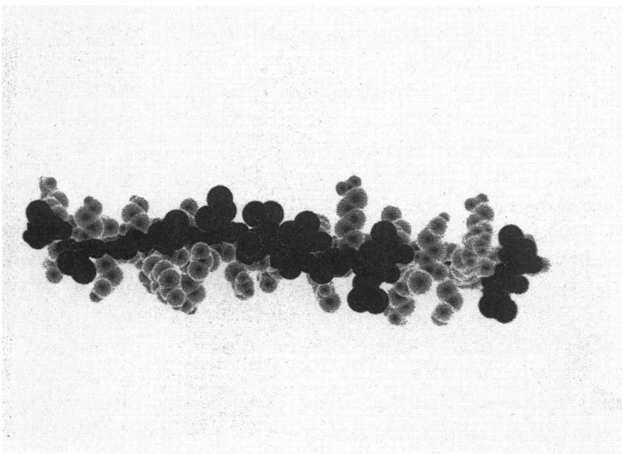


Fig. 4. CPK model of a portion of myosin-like domain I (residues 30–70) from SGP-2

Residues 30–70 of SGP-2 were modelled as a normal α-helix by using Macromodel to illustrate the amphipathic nature of this predicted helix. The hydrophobic residues have been rendered in black, whereas all other amino acids have been rendered in grey. The helix is oriented with the *N*-terminus to the left.

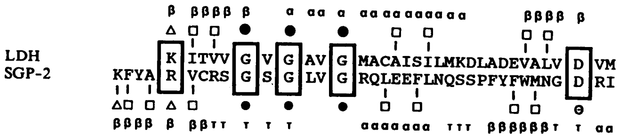

β ββββ β α αα α αααααααααα ββββ β
△ □ □ ● ● ● □ □ □ □
LDH
SGP-2
K ITVV G V G AV G MACAISILMKDLADEVALV D VM
KFYA R VCRS G S G LV G RQLEEFLNQSSPFYFWMNG D RI
△□ □ △ □ ● ● ● □ □ □ □ ϴ
ββββ β ββττ τ τ τ ααααααα τττ ββββββτ τ αα

Fig. 5. Alignment of a possible nucleotide-binding sequence conserved in SGP-2 and SP-40,40 with a fingerprint alignment deduced from five dinucleotide-binding enzymes

Abbreviations: LDH, pig lactate dehydrogenase; α, α-helix; β, β-sheet; τ, turn. Structural assignments are from X-ray data for LDH and from the prediction scheme of Chou & Fasman (1974) for SGP-2 and SP-40,40. Since the sequences of SGP-2 and SP-40,40 are absolutely conserved in this region, only data for SGP-2 are shown. The fingerprint and symbols are taken from Wierenga *et al.* (1986): ●, invariant glycine; □, neutral or hydrophobic groups forming the hydrophobic core of the β-α-β structural unit; ϴ, invariant negative charge involved in H-bonding to a ribose hydroxy group; △, invariant hydrophilic residue. All of the invariant residues in the fingerprint are conserved in the SGP-2/SP-40,40 sequence and are boxed for reference.

The results of these searches indicate that the sequences of rat SGP-2 and human SP-40,40 have more significant sequence similarity to both Cln-N and Cln-C than any of the other sequences contained within the database. Fig. 2 illustrates the alignment of the Cln-N and Cln-C sequences with rat SGP-2 and human SP-40,40 using the program GAP (Needleman & Wunsch, 1970). Cln-N has 56.8 % sequence identity with residues 25–60 of rat SGP-2; this increases to 83.8 % sequence similarity when conserved amino acid replacements are taken into account. Cln-C has 45.5 % sequence identity with residues 227–247 of rat SGP-2; this increases to 72.7 % sequence similarity when conserved amino acid replacements are taken into account. Similar degrees of identity and similarity are seen between the clusterin sequences and those of the SP-40,40 subunits. It was not necessary to introduce any gaps or deletions into any of these sequence comparisons. The high degree of sequence similarity between Cln-N and Cln-C and the corresponding subunits of rat SGP-2 clearly indicate the possibility of an evolutionary relationship between these two proteins.

To further examine the relationship between rat SGP-2, human SP-40,40 and ram clusterin, the sequences for Cln-N, Cln-C, human SP-40,40 and rat SGP-2 were analysed with the RELATE algorithm (Dayhoff *et al.*, 1983), which is designed to detect unusual similarity between sequences. The RELATE scores (in standard-deviation, S.D., units) between rat SGP-2 and the ram clusterin sequences are extremely high: 11.69 S.D. and 9.5 S.D. for Cln-N and Cln-C respectively. Similar RELATE scores are obtained for human SP-40,40 and the clusterin subunits: 9.3 S.D. and 13.9 S.D. respectively. These scores were obtained by using a segment length of 15 residues and 100 runs of the randomly permuted sequences. These RELATE scores, as an example, are comparable with the scores that are obtained when comparing segments of the highly conserved sequences contained in the cytochrome *c* superfamily (results not shown).

Comparison of rat SGP-2 precursor sequence with other known protein sequences: the search for significant sequence similarities

The SEARCH algorithm of Dayhoff *et al.* (1983) was used to screen the 5556 sequences contained in the PIR database (release 18.0; George *et al.*, 1986) against 15-amino-acid-long segments of the rat SGP-2 precursor sequence. This 'window' was advanced by five residues for each successive search.

A sequence that recurred a number of times in this SEARCH was that of myosin heavy chain from *Caenorhabditis elegans* (PIR sequence title code: MWKW). The RELATE comparison of Dayhoff *et al.* (1983) was used to compare MWKW with SGP-2 and SP-40,40. The RELATE scores were 3.85 S.D. and 4.53 S.D. respectively. These scores were obtained by using a segment length of 15 residues and 100 runs of the randomly permuted sequences and indicate that the probability that the sequence similarity observed arose solely by chance is less than 6×10^{-5} and 3×10^{-6}. These scores are of the same order of magnitude as the previously reported values for the similarity between SGP-2 and apolipoprotein A1 (Collard & Griswold, 1987). Additional myosin-heavy-chain sequences were analysed by the RELATE algorithm and of these the most significant score was obtained from the myosin heavy chain of *D. discoideum*. These RELATE scores were 6.0 S.D. and 6.6 S.D. for SGP-2 and SP-40,40 respectively and indicate that the probability that the sequence similarity observed arose solely by chance is less than 9.8×10^{-10} and 2×10^{-11} respectively.

(Devereux *et al.*, 1984). The DOTPLOT program compares a portion (residues 850–1300) of the tail domain of myosin heavy chain on the vertical axis with the full-length sequence of SGP-2 precursor on the horizontal axis in blocks of 30 amino acids. Each block of 30 amino acids in this segment of myosin heavy chain was compared with every 30-amino-acid block in the SGP-2 precursor. When the comparison score was greater than the stringency value of 15, a dot was placed at the position in the diagram that corresponds to the location of the compared amino acid blocks. Several features of the SGP-2 sequence are annotated as follows: solid black rectangles represent regions with similarity to the myosin tail domain; dark-shaded rectangles represent amphipathic helices predicted by the algorithm of Margalit *et al.* (1987); light-shaded rectangles represent the location of the two clusters of five cysteine residues in SGP-2 which span the regions 101–128 and 284–312; the solid black circle indicates the location of the putative nucleotide-binding site; the vertical bar represents the site at which proteolytic processing occurs to give the mature disulphide-linked heterodimer form of SGP-2. (*b*) Several representative sequence alignments between SGP-2 and the heavy myosin chains from *C. elegans* and *D. discoideum* are depicted. | represents identities and : represents conserved changes.

Graphic analysis of sequence similarities between rat SGP-2, human SP-40,40 and *C. elegans* myosin heavy chain

In an effort to examine further the significance of the sequence similarities between SGP-2, SP-40,40 and MWKW, these sequences were compared by using the COMPARE and DOTPLOT programs (Maizel & Lenk, 1981). Since results for SGP-2 and SP-40,40 and all of the myosin heavy chains were very similar, only results using SGP-2 and the myosin heavy chain from *C. elegans* are shown. The α-helical-tail domain of MWKW begins at residue 850 and continues to residue 1944 (McLachlan *et al.*, 1982; Karn *et al.*, 1983). Significant continuous similarities to SGP-2 and SP-40,40 abruptly begin at the start of the tail domain of MWKW and continue throughout its entire length (results not shown). This is in contrast with the head domain of MWKW, which shows only scattered similarities. Fig. 3 indicates that four distinct regions of SGP-2 have significant sequence similarities along the tail domain of MWKW; these will be referred to as 'myosin-like domains I, II, III and IV' in the present paper. Several sequence alignments representative of the degree of sequence similarity that exists between SGP-2 and the heavy chain of myosin are depicted in Fig. 3(*b*).

AMPHI, an implementation of the algorithm described by Margalit *et al.* (1987), was used to determine which regions of SGP-2 can form amphipathic helices. These regions are shown in Fig. 3 and correlate well with the regions of similarity between SGP-2 and MWKW. Fig. 4 depicts myosin-like region I from SGP-2 (residues 30–70) as a normal α-helix and is representative of the amphipathic character of these predicted helices.

Alignment of the amino acid fingerprint of a known nucleotide-binding sequence with a similar conserved sequence in rat SGP-2 and human SP-40,40

A number of dinucleotide-binding proteins contain a glycine-rich motif, GXGXXG in a *β-α-β* structure which can be identified by use of an amino acid fingerprint (Wierenga & Hol, 1983; Wierenga *et al.*, 1986). This fingerprint defines the type and position of amino acid residues necessary to form a nucleotide binding *β-α-β* structure. the longest contiguous stretch of identity between SGP-2 and SP-40,40 contains such a putative nucleotide-binding domain; the available protein sequence for clusterin does not span this region. As seen in Fig. 5, SGP-2/SP-40,40 follows the fingerprint exactly only for the boxed residues. These boxed residues are reported to be the only invariant residues in the fingerprint (Wierenga *et al.*, 1986). The predicted secondary structure (Chou & Fasman, 1974) does predict a *β-α-β* motif that is consistent with the Wierenga model for nucleotide-binding sites.

DISCUSSION

It has recently been shown that rat SGP-2 and human SP-40,40 are related proteins occurring in the reproductive system and the complement systems respectively (Kirszbaum *et al.*, 1989). It has long been suspected that rat SGP-2 and ram clusterin are closely related, owing to the number of their shared chemical properties. The results presented on the *N*-terminal sequencing of purified clusterin subunits and the results of the comparison of these sequences with the complete sequences of the SGP-2 and SP-40,40 subunits clearly demonstrate the relationship of clusterin to SGP-2 and SP-40,40. Clusterin, SGP-2 and SP-40,40 are related proteins occurring in diverse physiological systems in different species. The recent publication of the complete amino acid sequence of SP-40,40 has encouraged the undertaking of a more complete, segment by segment, search of known protein sequences in hope of uncovering sequences of functional importance that may help to elucidate the role of SP-40,40 in the complement system and the role of SGP-2 in the reproductive system.

Cheng *et al.* (1988*a*) have recently reported the isolation of clusterin from RTF, using methods different from those we initially described (Fritz *et al.*, 1983; Blaschuk *et al.*, 1983). Cheng *et al.* (1988*a*) have confirmed the previously published observations on the chemical characteristics of clusterin (Blaschuk & Fritz, 1984). In extensions of these observations, Cheng *et al.* (1988*a*) stated that the two subunits of clusterin consisted of two non-identical units having different *N*-terminal sequences. The sequence data provided, however, represent residues 4–25 of only one of the two clusterin subunits (see Fig. 6 of Cheng *et al.*, 1988*a*). The sequence of residues of the other subunit was deduced, but the two subunits were not physically separated and alkylated. The sequence analysis of the h.p.l.c.-purified carboxymethylated clusterin subunits reported here allows the unambiguous assignment of the *N*-terminal amino acid sequence of both clusterin subunits. This assignment includes the first four residues of both subunits; it was not possible to assign the initial four residues of either clusterin subunit in a direct or indirect fashion by the methods of Cheng *et al.* (1988*a*). An important consideration in the examination of the relationship between ram clusterin, rat SGP-2 and human SP-40,40 is that the sequences of both subunits of clusterin align well to the *N*-termini of the respective subunits of rat SGP-2 and human SP-40,40.

The hypothesis that ram clusterin, rat SGP-2 and human SP-40,40 share a common evolutionary relationship is supported by the good alignment and high degree of sequence similarity between the *N*-terminal sequences of the respective subunits of these three proteins. This is also supported by the results of the comparisons made with the RELATE algorithm. The RELATE algorithm is designed to detect unusual similarity between two sequences by comparing all possible segments of a given length from one sequence with all possible segments of the same length from the second sequence; this comparison is repeated with randomly permuted versions of the two sequences. The difference between the real and the random comparisons is expressed in units of S.D. The greater the difference between the real and random comparisons, the greater the likelihood that the two sequences are related in a statistically significant fashion.. The exceptionally high RELATE scores obtained when comparing the sequences of Cln-N and Cln-C with that of rat SGP-2 and human SP-40,40 are indicative of evolutionarily related protein sequences. The probability that two protein sequences developed independently and attained the degree of sequence similarity represented by a RELATE score of 9.5 S.D. (see the Results section) is less than 1.1×10^{-21} (Dayhoff *et al.*, 1983).

Recently, Cheng *et al.* (1988*b*) also reported the isolation and *N*-terminal sequencing of the two subunits of the protein responsible for clusterin activity from rat Sertoli-cell enriched culture media. In addition, a cDNA clone was isolated with antibodies raised from one of the purified subunits The cDNA clone isolated codes for the last 166 of 211 amino acids contained in the *C*-terminal subunit of rat SGP-2. The protein sequence obtained, as well as the deduced amino acid sequence from the cDNA clone of rat clusterin are identical with those of rat SGP-2 (Collard & Griswold, 1987; Cheng *et al.*, 1988*b*). Therefore clusterin activity in the rat is attributable to SGP-2; this corroborates well with our data, which suggests that ram clusterin and rat SGP-2 are two forms of the same conserved protein sequence.

The first functional domain to be tentatively identified in SGP-2 and SP-40,40 is contained in the longest continuous stretch of identity between these two sequences, which contains 45 identities (in a longer stretch of 63 identities out of 64). This is a putative nucleotide-binding β-α-β structure as identified by using the amino acid sequence fingerprint of Wierenga & Hol (1983). The Wierenga fingerprint appears in a number of ADP-binding proteins. This fingerprint essentially defines a set of rules for the occurrence of amino acids at 11 defined positions over a stretch of approx. 30 residues. Every sequence with known crystal structure that has been found to match this amino acid sequence fingerprint does indeed contain an ADP-binding β-α-β structure (Wierenga *et al.*, 1986). The glycine residues and the two charged residues are the only residues in the fingerprint that are reported to be absolutely conserved (Wierenga *et al.*, 1986). These residues are conserved in the sequence seen in SGP-2 and SP-40,40. The other residues of the fingerprint are present, but in shifted positions. The predicted secondary structure of the putative nucleotide-binding structure is consistent with the model of Wierenga. It will also be of interest to determine the complete sequence of ram clusterin in order to establish if this nucleotide-binding.sequence is also conserved in the ram.

The ATPase activity of myosin is localized to the head domain of the heavy chain, which comprises the first 850 amino acids. The vast majority of the remainder of the heavy chain is contained in the long α-helical-tail domain. In the myosin molecule, two heavy-chain tail domains interact in a coiled-coil fashion. A 28-amino-acid repeating motif with a distinct repeating pattern of alternating charged amino acids and the amphipathic nature of the helices help to mediate this interaction (McLachlan *et al.*, 1982). Graphic analysis (see Fig. 3) shows that the strongest regions of similarity between SGP-2/SP-40,40 and MWKW are between the predicted amphipathic α-helical segments of SGP-2/SP-40,40 and the rod-like tail domain of MWKW.

Evidence has been presented that SGP-2 is synthesized as a single polypeptide chain which is modified and then cleaved to the mature 47 and 34 kDa subunits (Collard & Griswold, 1987). The SGP-2 and SP-40,40 sequences have two conserved clusters of cysteine residues. They also show conservation of the proteolytic processing site at the junction between subunits. As a result of being a disulphide-linked heterodimer with a fairly discrete cluster of five cysteine residues approximately centred in the linear sequence of each subunit, it is possible to depict the linear sequence of either SGP-2 or SP-40,40 as a cross. The centre of the cross would represent the clustered disulphide link(s) between the two subunits. The distal portion of each arm would be an amphipathic helix with similarity to the myosin tail; just off-centre on the arm of the cross containing myosin-like domain I would be located the putative nucleotide-binding site. The linear sequence of the subunits may be oriented in an anti-parallel manner; this could facilitate disulphide-bonding and would be consistent with the precursor molecule folding back on itself in a 'hairpin' fashion. This cross analogy is obviously highly schematic and is presented only to emphasize the presence of four distinct amphipathic domains with a putative nucleotide-binding site located in a central position (with respect to the linear sequences involved).

The segregation of hydrophobic and hydrophilic residues to opposing faces of an α-helix results in a secondary structure that can behave in a much more hydrophobic fashion than the inspection of amino acid composition or the linear sequence would lead one to predict. For example, the linear sequence of myosin-like domain I in SGP-2 (residues 30–70) does not contain any significant local concentrations of hydrophobic residues, but when this segment is modelled as an α-helix (Fig. 4), it becomes apparent that this sequence is capable of forming a helix with a face that is predominantly hydrophobic. It has been previously reported that the hydrophobic-like behaviour of SGP-2 could not be predicted from amino acid composition, nor can it be predicted from examination of hydropathy profiles generated from the linear sequence. The formation of amphipathic α-helices had not been previously considered as a possible contributing factor in the physical properties of SGP-2.

The myosin-like amphipathic helices of SGP-2 may mediate inter- or intra-molecular interactions in a fashion analogous to the interaction of the tail domains of two myosin heavy chains. It is conceivable that these amphipathic domains may contribute to the structure or function of SGP-2 in a number of ways. These domains may mediate dimer formation, which has previously been demonstrated by gel-filtration chromatography in the presence of 5 mM-EDTA (Griswold *et al.*, 1982); they could aid in both the clusterin activity of SGP-2 and SGP-2s associated with spermatozoa by mediating an interaction with cellular membranes or they could be important in maintaining a rod-like tertiary structure, which would be consistent with velocity-sedimentation data that demonstrate that SGP-2 sediments with a velocity slightly less than 2-fold lower than expected for a globular 70 kDa protein (results not shown).

Although it is clear that clusterin, SGP-2 and SP-40,40 are related proteins, the roles of these proteins in physiologically distinct systems are not yet known. The possibility that clusterin and SGP-2 have a different role in the reproductive system from that of SP-40,40 in the complement system must be considered, since terminal complement components are not detectable in human seminal plasma by e.l.i.s.a. (Kirszbaum *et al.*, 1989). One must also consider that *S*-protein/vitronectin, another component of the SC5b-9 complex (Podack *et al.*, 1978) is known to have several functions unrelated to the complement system (Jenne & Stanley, 1985). The discovery of a putative nucleotide-binding domain and the presence of four myosin-like amphipathic domains may help to direct future investigation to solve this perplexing puzzle.

We thank the following individuals: Ms. Alice Karl and Mr. Dean Bays for Sertoli-cell culture work; Ms. Krystyna Burdzy for generating ascites fluid used for the isolation of ram clusterin and Dr. Steven Sylvester for critical reading of this manuscript. This research was funded by the following grants: Canadian Medical Research Council (no. MT 3292 to I.B.F.); National Institute of Child Health and Human Development (no. HD 10808 to M.D.G.).

REFERENCES

Blaschuk, O. W. & Fritz, I. B. (1984) Can. J. Biochem. Cell Biol. **62**, 456–461

Blaschuk, O. W., Burdzy, K. & Fritz, I. B. (1983) J. Biol. Chem. **258**, 7717–7720

Cheng, C. Y., Mathur, P. P. & Grima, J. (1988*a*) Biochemistry **27**, 4079–4085

Cheng, C. Y., Chen, C. C., Feng, Z., Marshall, A. & Bardin, C. W. (1988*b*) Biochem. Biophys. Res. Commun. **155**, 398–404

Chou, P. Y. & Fasman, G. D. (1974) Biochemistry **13**, 222–244

Clermont, Y., (1972) Physiol. Rev. **52**, 198–236

Collard, M. & Griswold, M. D. (1987) Biochemistry **26**, 3297–3303

Dayhoff, M. O., Barker, W. C. & Hunt, L. T. (1983) Methods Enzymol. **91**, 524–545

Devereux, J., Haeberli, P. & Smithies, O. (1984) Nucleic Acids Res. **12**, 387–395

Fawcett, D. W. (1975) in Handbook of Physiology (Hamilton, D. W. & Grup, R. D., eds.), pp. 21–55, American Physiological Society, Bethesda

Fritz, I. B. (1978) Biochem. Act. Horm. **5**, 249–278

Fritz, I. B., Burdzy, K., Setchell, B. & Blaschuk, O. (1983) Biol. Reprod. **58**, 1173–1188

Fritz, I. B., Blaschuk, O. W. & Burdzy, K. (1984) in Gonadal Proteins and Peptides and Their Biological Significance (Sairam, M. R. & Atkinson, L. E., eds.), pp. 311–325, World Scientific Publishing Co., Philadelphia

George, D. G., Barker, W. C. & Hunt, L. T. (1986) Nucleic Acids Res. **14**, 11–16

Griswold, M. D. (1988) Int. Rev. Cytol. **110**, 133–156

Griswold, M. D., Roberts, K. & Bishop, P. (1982) Biochemistry **25**, 7265–7270

Jenne, D. & Stanley, K. K. (1985) EMBO J. **4**, 3153–3157

Karn, J., Brenner, S. & Barnett, L. (1983) Proc. Natl. Acad. Sci. U.S.A. **80**, 4253–4257

Kirszbaum, L., Sharpe, J. A., Murphy, B., d'Apice, A. J. F., Classon, B., Hudson, P. & Walker, I. D. (1989) EMBO J. **8**, 711–718

Lipman, D. J. & Pearson, W. R. (1985) Science **227**, 1435–1441

Maizel, J. V. & Lenk, R. P. (1981) Proc. Natl. Acad. Sci. U.S.A. **78**, 7665–7669

Margalit, H., Spouge, J. L., Cornette, J. L., Cease, K. B., Delisi, C. & Berzofsky, J. A. (1987) J. Immunol. **138**, 2213–2229

McLachlan, A. D. & Karn, J. (1982) Nature (London) **299**, 226–231

Murphy, B. F., Kirszbaum, L., Walker, I. D. & d'Apice, A. J. F. (1988) J. Clin. Invest. **81**, 1858–1864

Needleman, S. B. & Wunsch, C. D. (1970). J. Mol. Biol. **48**, 443

Podack, E. R., Kolb, W. P. & Muller-Eberhard, H. J. (1978) J. Immunol. **120**, 1841–1848

Rosenior, J., Tung, P. S. & Fritz, I. B. (1987) Biol. Reprod. **36**, 1313–1320

Sylvester, S. R., Skinner, M. K. & Griswold, M. D. (1984) Biol. Reprod. **31**, 1087–1101

Tung, P. S. & Fritz, I. B. (1985) Biol. Reprod. **33**, 177–186

Warrick, H. M. & Spudich, J. A. (1987) Annu. Rev. Cell Biol. **3**, 379–421

Wierenga, R. K. & Hol, W. G. J. (1983) Nature (London) **302**, 842–844

Wierenga, R. K., Terpstra, P. & Hol, W. G. J. (1986) J. Mol. Biol. **187**, 101–107

Received 19 July 1989/8 February 1990; accepted 22 February 1990