Common evolutionary origin of α_2 -macroglobulin and complement components C3 and C4

(plasma proteins/internal thiol esters/protein structure/sequence homology/specific limited proteolysis)

LARS SOTTRUP-JENSEN*, TERRENCE M. STEPANIK*[†], TORSTEN KRISTENSEN*[‡], PETER B. LØNBLAD*[§],
CLAIRE M. JONES*[§], DIANE M. WIERZBICKI*[¶], STAFFAN MAGNUSSON*, HORST DOMDEY*^{||**},
RICK A. WETSEL^{††}, ÅKE LUNDWALL^{††}, BRIAN F. TACK^{††}, AND GEORG H. FEY^{††}

*Department of Molecular Biology, University of Aarhus, 8000 Aarhus C, Denmark; [†]Swiss Institute for Experimental Cancer Research, CH-1055 Epalinges, Switzerland; and ^{††}Department of Immunology, Scripps Clinic and Research Foundation, La Jolla, CA 92037

Communicated by Oscar D. Ratnoff, June 18, 1984

ABSTRACT A comparison of the sequence of the subunit of human α_2 -macroglobulin (α_2 M; 1451 amino acid residues) with that of murine complement component pro-C3 (1639 amino acid residues) reveals eight extended regions of sequence similarity. These regions contain between 19% and 31% identically placed residues and account for 75% and 67%, respectively, of the polypeptide chains of α_2 M and pro-C3. Published sequence data for complement component C4 show that segments of this protein match well with corresponding stretches in α_2 M and pro-C3. It is proposed that α_2 M, C3, and C4, which all contain a unique activatable β -cysteinyl- γ -glutamyl thiol ester, have a common evolutionary origin and are homologous proteins. Several larger regions of low sequence similarity indicate the presence of structural domains in each of these proteins that specifically modify an underlying common gross structure. The quartets of basic residues in pro-C3 and pro-C4, at which cleavage takes place to produce the mature subunits of these proteins, and most of the residues forming the anaphylatoxin peptides of C3 and C4 (C3a and C4a) are absent in α_2 M. In addition, C3 and C4 contain large portions, which extend beyond the COOH terminus of α_2 M.

Among the plasma proteins, α_2 -macroglobulin (α_2 M) and the complement components C3 and C4 share a unique sensitivity toward small nitrogen nucleophiles such as hydrazine, hydroxylamine, and methylamine. Upon incubation with these compounds, the proteinase binding properties of α_2 M and the hemolytic activity of C3 and C4 are gradually lost (1–3). It was recently shown that α_2 M, C3, and C4 have several other features in common: a characteristic pattern of heat/denaturation cleavage (4–9); an ability to incorporate methylamine, thereby forming a residue of γ -glutamylmethylamide (10–13); an appearance of SH groups after incubation with nitrogen nucleophiles or chaotropes (8, 14–19) or after activation with proteolytic enzymes (8, 14, 16, 18); and an ability of the proteolytically activated proteins to participate in covalent binding reactions (20–23), engaging the methylamine reactive Glx residue (24–27). These features result from the presence of a unique type of reactive site in each of these proteins, identified as an activatable internal β -cysteinyl- γ -glutamyl thiol ester (13, 14, 16), located in the α -chain part of C3 and C4 (13, 24, 28, 29) and in each of the four identical subunits of α_2 M (10, 17, 30). Since the thiol ester structure, -Gly-Cys-Gly-Glu-Glx-, is found in regions

of conserved amino acid sequence in all three proteins (10, 13, 17, 24, 28–30), an evolutionary relationship is apparent.

This is further substantiated by the similar size of the α_2 M subunits [M_r , 180,000 (31, 32)] and the single chains of the pro-C3 and pro-C4 molecules [M_r , 175,000 and 190,000, respectively (33–35)], and by the organization of the individual chains of C3 and C4 within pro-C3 and pro-C4 [β - α and β - α - γ , respectively (36–39)]. The high degree of sequence homology between the anaphylatoxins generated from C3 and C4 upon proteolytic activation (C3a and C4a) has been discussed (40, 41). The primary structure of human α_2 M has recently been determined (42). From a detailed comparison of this structure with the amino acid sequence of murine prepro-C3 (43, 44) and with published sequences of human and murine C4, it is evident that all three proteins share extended stretches of similar sequence, indicating that they have a common evolutionary origin.

MATERIALS AND METHODS

The primary structure of human α_2 M was determined by standard procedures of protein chemistry (42). The amino acid sequence of murine prepro-C3 was deduced from the nucleotide sequence of its cDNA, which was determined using the Maxam–Gilbert technique (35, 39, 45) and the dideoxynucleotide technique (43, 44). The amino acid sequences of α_2 M and prepro-C3 were compared using the DIAGON program set at a span of 25 residues and a percentage score of 280 (46, 47). A residue-by-residue comparison was performed manually and scores of homology were calculated using a gap penalty of 2.

RESULTS AND DISCUSSION

When the sequences of human α_2 M and murine prepro-C3 were compared by the diagonal matrix method, it became evident that these proteins share many short stretches of strong sequence similarity. Fig. 1 shows an alignment of the sequences of human α_2 M and murine pro-C3, emphasizing their mutual homology. Published partial sequences from human and murine C4 have been included in the comparison (24, 41, 48–51).

Most of the β -chain part of pro-C3 (residues 1–642) matches well with the NH₂-terminal 609 residues of α_2 M, and extended stretches from the α -chain part of pro-C3 (residues

Abbreviation: α_2 M, α_2 -macroglobulin.

[†]Present address: Division of Biological Sciences, National Research Council of Canada, Ottawa, ON, Canada.

[‡]Present address: Department of Immunology, Scripps Clinic and Research Foundation, La Jolla, CA 92037.

[§]Present address: Department of Biochemistry, Imperial College, London, England.

[¶]Present address: Abbott Laboratories, Abbott Diagnostic Division, North Chicago, IL 60064.

^{||**}Present address: Max Planck Institute of Biochemistry, Martinsried, F.R.G.

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.

FIG. 1. (Figure continues on the next page.)

FIG. 1. Alignment of the complete amino acid sequences (represented by standard one-letter abbreviations) of human α_2 M and murine pro-C3. Stretches from human or murine C4 have been included (24, 41, 48–51). Identically placed residues in any pair of sequences are boxed. Residues in pro-C3 or C4 that are chemically similar to residues in α_2 M are underlined. Chemically similar residues are defined as follows: D = E, T = S, N = Q, V = M = I = L, F = Y = H = W, and K = R. Gaps have been introduced to maximize the alignment of the sequences. Glucosamine-based oligosaccharide groups are attached to the following asparagine residues in α_2 M: 32, 47, 224, 373, 387, 846, 968, and 1401. Asn-915 and Asn-1593 in pro-C3 are candidates for the attachment of glucosamine-based oligosaccharide groups. The 24-residue leader sequence (47) of prepro-C3 has been omitted. The C3 numbering refers to pro-C3.

725–1639) match well with corresponding segments in the COOH-terminal 740 residues of α_2 M. Apart from 2 long gaps (34 residues in α_2 M and 14 residues in pro-C3), only short gaps needed to be introduced to maximize the alignment of the two sequences (Table 1). Eight extended regions, I–VIII (Table 1, Fig. 2), show between 19% and 31% sequence identity. If one includes chemically similar residues (see Fig. 1 legend), the sequence similarity ranges between 32% and 40%.

The boundaries between the regions are arbitrary, particularly for those that are separated by short stretches of dissimilar sequence. In the regions consisting of residues 410–570 and residues 1250–1400 of α_2 M, several slightly different alignments give almost the same score of sequence identity. The regions II, III, and IV in the NH₂-terminal part of α_2 M and C3 could be considered one large block of homologous sequence, having an overall sequence identity of $\approx$ 17%. Likewise, the regions VI, VII, and VIII could possibly constitute another large block of homologous sequence. The

overall score of sequence identity for that block is $\approx$ 22%.

The regions I–VIII account for 75% of the α_2 M sequence and for 67% of the pro-C3 sequence. The published short stretches of sequence for the α - and γ -chains of C4 (refs. 24, 41, 48–51; Fig. 1) match well with corresponding stretches in α_2 M and pro-C3, giving a similar degree of sequence identity.

From the eight regions of extended sequence similarity between α_2 M and C3 and the match with the available partial sequence data for C4, we propose that α_2 M, C3, and C4 must have a common evolutionary origin. Regions I–VIII probably constitute discrete domains (or parts thereof) that have essentially the same tertiary structure in α_2 M and C3 (and probably C4), suggesting a similar gross structure for the monomeric unit of α_2 M and the multiple chain structures of C3 and C4. These contentions are supported by the very similar circular dichroism spectra of α_2 M and the C3b and C4b parts (52–55), which suggest that these proteins contain extended regions of β -sheet conformation and very little α -helix (54,

Table 1. Score of sequence identity in longer stretches of α_2 M and C3

Position in		% identity		Size of gaps introduced in		Region of homology†
α_2 M	C3	Identically placed residues	Including chemically similar residues*	α_2 M	C3	
1–243	1–248	21	35	5, 2, 1, 2, 2	1, 1, 1	I
244–302	249–307	<5	<10	0	0	
303–407	308–408	19	32	1	1, 3, 1	II
408–419	409–417	<5	<10	0	3	
420–503	418–497	19	33	0	2, 2	III
504–513	498–507	<5	<10	0	0	
514–609	508–605	21	35	2	0	IV
610–696	606–726	<5	<9	34	1, 2	
697–889	727–927	19	36	1, 5, 1, 2, 2	3	V
890–941	928–978	<5	<10	0	1	
942–1133	979–1174	28	39	3, 1, 4	4	VI
1134–1174	1175–1197	<5	<5	0	1, 2, 2, 14	
1175–1241	1198–1258	31	40	0	6	VII
1242–1326	1259–1346	<10	<15	1, 1, 3	1	
1327–1440	1347–1469	29	35	1, 3, 3, 2, 1	1, 1	VIII

A gap penalty of 2 was used when calculating the score of identity.
 *Defined in legend to Fig. 1.
 †Shown in Fig. 2.

FIG. 2. Schematic representation of the sequences of α_2 M (Upper) and pro-C3 (Lower), emphasizing the eight larger regions of pronounced sequence similarity (see Table 1). Locations of oligosaccharide chains are not shown. Disulfide bridge pattern of α_2 M and C3a portion of pro-C3 are shown. Positions of half-cystine residues in α_2 M and pro-C3, which have not yet been identified in disulfide bridges, are indicated by short vertical bars. The reactive β -cysteinyl- γ -glutamyl thiol ester site in α_2 M, C3, and C4 is shown as an asterisk. Sites of activation cleavage in α_2 M and C3 are indicated by arrows above the lines. Sites in pro-C3 where cleavage, leading to the formation of the α - and β -chains of C3, occurs are shown as arrows below the lines.

55). Strikingly, none of the half-cystine residues (except the residue that is part of the reactive thiol ester site) found in C3 have counterparts in α_2 M, suggesting that the two proteins have different disulfide bridge patterns. In particular, 9 half-cystine residues are found in the COOH-terminal 158-residue stretch of the α -chain part of pro-C3, which is lacking in α_2 M.

Apart from this region, three other segments in α_2 M and pro-C3, located between regions I and II, IV and V, and V and VI, respectively, show very little sequence identity. The 59-residue section located between regions I and II is part of an area that in α_2 M contains several closely spaced disulfide bridges. This area probably constitutes a domain of distinct tertiary structure in α_2 M and C3. The segments between regions IV and V (87 residues in α_2 M and 121 residues in pro-C3) contain the activation cleavage sites of α_2 M and C3 and the arginyl quartet in pro-C3 where cleavage, resulting in the formation of the β - and α -chains of the mature protein, takes place (39). No quartet of basic residues is present in α_2 M, consistent with the fact that its monomers undergo no maturation cleavages into subunits comparable to the C3 α - and β -chains. α_2 M contains the sequence -Arg-Lys-Pro-Lys, 14 residues away from the Arg-Arg-Arg-Arg sequence of pro-C3 (Fig. 1). The presence of a prolyl residue in the α_2 M sequence possibly prevents a maturation cleavage at this position. The C3a anaphylatoxin peptide sequence is located between regions IV and V (39). This element and the analogous C4a peptide constitute homologous domains, as evidenced by their similar sequences (40, 41) and their high content of α -helical structures (55, 56, 57). However, no complete anaphylatoxin-like structure is present in α_2 M.

Although a number of identically placed residues are apparent in the regions of α_2 M, C3, and C4, preceding and including the activation cleavage sites (residues 680–708 in α_2 M), it is likely, especially in view of the rather restricted cleavage specificity of the C3 and C4 activating enzyme complexes, that the conformation of the individual polypeptide stretches in α_2 M, C3, and C4 containing these regions differ. In contrast with C3 and C4, the activation cleavage region of α_2 M is large and contains several clusters of residues where specific limited proteolysis can take place. Initial cleavage by proteinases presumably takes place in the region of residues 681–686 and is followed by one or more rapid secondary cleavage(s) that occur preferentially at residues 696–698. These residues are located in a region homologous to the activation cleavage sites of C3 and C4 (42, 58). Whether the resulting small peptides, which share partial sequence identity with the active COOH-terminal regions of C3a and C4a, are released from α_2 M and have any physiological function is not known. For all three proteins, the proteolytic events lead to the activation of the internal β -cysteinyl- γ -glutamyl thiol esters, which in α_2 M and pro-C3 are located 255–271 and 265 residues, respectively, from the activation cleavage region toward the COOH terminus. The structural features responsible for this activation are not known; but, by analogy with other proteins activated by specific limited proteolysis, they might reside in the homologous regions V

and VI. These two large regions are separated by an ≈ 50 -residue segment of low sequence similarity. As probed by proteolytic digestion (generation of the C3d and C4d fragments), this stretch presumably is a relatively exposed connecting strand between two larger domains. The C3d (and C4d) fragments appear to terminate in the stretch of weak homology, which connects regions VII and VIII.

Compared with α_2 M, both pro-C3 and pro-C4 contain large stretches extending beyond the COOH terminus of α_2 M. In pro-C3, the α -chain extension is 158 residues; in pro-C4, the γ -chain extension could be slightly longer. The α - γ -chain junction in pro-C4 seems to be located equivalent to residues 1325–1328 in α_2 M. Evidently, part of the γ -chain sequence of C4 is homologous to the α -chain of C3. These COOH-terminal extensions in pro-C3 and pro-C4 relative to α_2 M could have a distinct genetic origin, or more likely, they could represent structures that have been deleted from an ancestral α_2 M gene.

Thus, the single chains of pro-C3, pro-C4, and the monomer of the tetrameric α_2 M are most likely composed of a number of different relatively small domains having essentially the same tertiary structure in the three proteins. Presumably, these domains form an underlying common gross structure that is modified by other domains, some of which may be specific for each of these proteins. α_2 M, C3, and C4 form a class of large plasma proteins, containing internal activatable β -cysteinyl- γ -glutamyl thiol esters, which enable these proteins when proteolytically activated to bind covalently to other structures (the activating proteinase in the case of α_2 M and antigen-antibody aggregates or cell-surface structures in the case of C3 and C4). All three proteins undergo conformational changes upon proteolytic activation or upon cleavage of their thiol esters by amines. The altered conformation of each protein is readily recognized by specific cell-surface receptors leading to endocytosis of α_2 M-proteinase complexes and to the opsonization of complexes between activated C3 (C4) and cellular targets.

After submission of this manuscript, the complete sequences of human prepro-C3 and human pro- α_2 M (unpublished results) and of human pro-C4 (59), all derived from cDNA, became accessible to us. These data lend further support to the proposed common ancestry of these proteins.

We thank R. F. Doolittle for a computer-assisted comparison between α_2 M and parts of the C3 sequence and for helpful discussions. We are grateful to Keith Dunn for preparation of the manuscript. This work was supported by National Heart, Lung, and Blood Institute Grant HL16238, by the Danish Science Research Council, by the North Atlantic Treaty Organization Science Foundation, by U.S. Public Health Service Grants AI-19222 and AI19651, and by American Heart Association Grant AHA-80-841. This is publication no. 3420 IMM from the Research Institute of Scripps Clinic.

1. Ratnoff, O. D., Lepow, I. H. & Pillemer, L. (1954) *Bull. Johns Hopkins Hosp.* **94**, 196–279.
2. Müller-Eberhard, H. J. & Biro, C. E. (1963) *J. Exp. Med.* **118**, 447–466.
3. Steinbuch, M., Pejaudier, L., Quintin, M. & Martin, V. (1968)

- Biochim. Biophys. Acta* **154**, 228–231.
4. Harpel, P., Hayes, M. B. & Hugli, T. E. (1979) *J. Biol. Chem.* **254**, 8669–8678.
5. Barrett, A. J., Brown, M. A. & Sayers, C. A. (1979) *Biochem. J.* **181**, 401–408.
6. Howard, J. B., Vermeulen, M. & Swenson, R. P. (1980) *J. Biol. Chem.* **255**, 3820–3822.
7. Janatova, J., Tack, B. F. & Prahl, J. W. (1980) *Biochemistry* **19**, 4479–4485.
8. Janatova, J. & Tack, B. F. (1981) *Biochemistry* **20**, 2394–2402.
9. Sim, R. B. & Sim, E. (1981) *Biochem. J.* **193**, 129–141.
10. Swenson, R. P. & Howard, J. B. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 4313–4316.
11. Howard, J. B. (1980) *J. Biol. Chem.* **255**, 7082–7084.
12. Gorski, J. & Howard, J. B. (1980) *J. Biol. Chem.* **255**, 10025–10028.
13. Tack, B. F., Harrison, R. A., Janatova, J., Thomas, M. L. & Prahl, J. W. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 5764–5768.
14. Janatova, J., Lorenz, P. E., Schechter, A. N., Prahl, J. W. & Tack, B. F. (1980) *Biochemistry* **19**, 4471–4478.
15. Pangburn, M. K. & Müller-Eberhard, H. J. (1980) *J. Exp. Med.* **152**, 1102–1114.
16. Sottrup-Jensen, L., Petersen, T. E. & Magnusson, S. (1980) *FEBS Lett.* **121**, 275–279.
17. Sottrup-Jensen, L., Hansen, H. F., Mortensen, S. B., Petersen, T. E. & Magnusson, S. (1981) *FEBS Lett.* **123**, 145–148.
18. Salvesen, G. S., Sayers, C. A. & Barrett, A. J. (1981) *Biochem. J.* **195**, 453–461.
19. Howard, J. B. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 2235–2239.
20. Law, S. K. & Levine, R. P. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 2701–2705.
21. Campbell, R. D., Dodds, A. W. & Porter, R. R. (1980) *Biochem. J.* **189**, 67–80.
22. Gadd, K. J. & Reid, K. B. M. (1981) *Biochem. J.* **195**, 471–480.
23. Harpel, P. C. & Rosenberg, R. D. (1976) in *Progress in Hemostasis and Thrombosis*, ed. Spaet, T. H., ed. (Grune & Stratton, New York), Vol. 3, pp. 145–189.
24. Campbell, R. D., Gagnon, J. & Porter, R. R. (1981) *Biochem. J.* **199**, 359–370.
25. Hostetter, M. K., Thomas, M. L., Rosen, F. S. & Tack, B. F. (1982) *Nature (London)* **298**, 72–74.
26. Sottrup-Jensen, L. & Hansen, H. F. (1982) *Biochem. Biophys. Res. Commun.* **107**, 93–100.
27. Sottrup-Jensen, L., Hansen, H. F. & Christensen, U. (1984) *Ann. N.Y. Acad. Sci.*, 188–208.
28. Thomas, M. L., Janatova, J., Gray, W. R. & Tack, B. F. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 1054–1058.
29. Harrison, R. A., Thomas, M. L. & Tack, B. F. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 7388–7392.
30. Swenson, R. P. & Howard, J. B. (1980) *J. Biol. Chem.* **255**, 8087–8091.
31. Swenson, R. P. & Howard, J. B. (1979) *J. Biol. Chem.* **254**, 4452–4456.
32. Harpel, P. C. (1973) *J. Exp. Med.* **138**, 508–521.
33. Brade, V., Hall, R. E. & Colten, H. R. (1977) *J. Exp. Med.* **146**, 759–765.
34. Hall, R. E. & Colten, H. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 1707–1710.
35. Odink, K., Fey, G. H., Wiebauer, K. & Diggelmann, H. (1980) *J. Biol. Chem.* **256**, 1453–1458.
36. Parker, K. L., Schreffler, D. C. & Capra, J. D. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 4275–4278.
37. Goldberger, G., Abraham, G. N., Williams, J. & Colten, H. R. (1980) *J. Biol. Chem.* **255**, 7071–7074.
38. Goldberger, G., Thomas, M. L., Tack, B. F., Williams, J., Colten, H. R. & Abraham, G. N. (1981) *J. Biol. Chem.* **256**, 12612–12619.
39. Domdey, H., Wiebauer, K., Kazmaier, M., Muller, V., Odink, K. & Fey, G. H. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 7619–7623.
40. Hugli, T. E. (1975) *J. Biol. Chem.* **250**, 8293–8301.
41. Moon, K. E., Gorski, J. P. & Hugli, T. E. (1981) *J. Biol. Chem.* **256**, 8685–8692.
42. Sottrup-Jensen, L., Stepanik, T. M., Kristensen, T., Wierzbicki, D. M., Jones, C. M., Magnusson, S. & Petersen, T. E. (1984) *J. Biol. Chem.* **259**, 8318–8327.
43. Lundwall, Å., Wetsel, R. A., Domdey, H., Tack, B. F. & Fey, G. H. (1984) *J. Biol. Chem.*, in press.
44. Wetsel, R. A., Lundwall, Å., Davidson, F., Tack, B. F., Gibson, T. & Fey, G. H. (1984) *J. Biol. Chem.*, in press.
45. Wiebauer, K., Domdey, H., Diggelmann, H. & Fey, G. H. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 7077–7081.
46. Staden, R. (1982) *Nucleic Acids Res.* **10**, 2951–2961.
47. Fey, G. H., Lundwall, Å., Wetsel, R. A., Tack, B. F., de Bruijn, M. H. L. & Domdey, H. (1984) *Philos. Trans. R. Soc. London Ser. B* **306**, 333–344.
48. Press, E. M. & Gagnon, J. (1981) *Biochem. J.* **199**, 351–357.
49. Chakravarti, D. N., Campbell, R. D. & Gagnon, J. (1983) *FEBS Lett.* **154**, 387–390.
50. Gigli, I., Von Zabern, I. & Porter, R. R. (1977) *Biochem. J.* **165**, 439–446.
51. Ogata, R. T., Schreffler, D. C., Sepich, D. S. & Lilly, S. P. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 5061–5065.
52. Björk, I. & Fish, W. W. (1982) *Biochem. J.* **207**, 347–356.
53. Molenaar, J. L., Helder, A. W., Müller, M. A. C., Goris-Mulder, M., Lonker, L. S., Brouwer, M. & Pondman, K. W. (1975) *Immunochemistry* **12**, 359–364.
54. Isenman, D. E. & Cooper, N. R. (1981) *Mol. Immunol.* **18**, 331–339.
55. Gorski, J. P., Hugli, T. E. & Müller-Eberhard, H. J. (1981) *J. Biol. Chem.* **256**, 2707–2711.
56. Hugli, T. E., Morgan, W. T. & Müller-Eberhard, H. J. (1975) *J. Biol. Chem.* **250**, 1479–1483.
57. Huber, R., Scholze, H., Paques, E. P. & Deisenhofer, J. (1980) *Hoppe-Seyler's Z. Physiol. Chem.* **361**, 1389–1399.
58. Sottrup-Jensen, L., Lønblad, P. B., Stepanik, T. M., Petersen, T. E., Magnusson, S. & Jörnval, H. (1981) *FEBS Lett.* **127**, 167–173.
59. Belt, K. T., Carroll, M. C. & Porter, R. R. (1984) *Cell* **36**, 907–914.