Volume 282, number 1, 9–12 FEBS 09607 April 1991

ADONIS 001457939100317F

Complement components C1r/C1s, bone morphogenic protein 1 and *Xenopus laevis* developmentally regulated protein UVS.2 share common repeats

Peer Bork

Central Institute of Molecular Biology, Department of Biomathematics, Robert-Rössle-Str. 10, O-1115 Berlin-Buch, Germany

Received 24 January 1991

Property patterns were constructed, based on an alignment of related domains in human complement subcomponents C1r and C1s as well as in the sea urchin protein uEGF. This kind of consensus pattern was able to identify similar domains in a human bone morphogenic protein, in a *Xenopus laevis* embryonal protein involved in dorsoanterior development and in a calcium-dependent serine protease secreted from malignant hamster embryo fibroblast cells. Because of the high level of overall sequence homology this protease may be the hamsters' equivalent of the human complement subcomponent C1s. The resulting multiple alignment of all studied domains suggests functionally and structurally important regions.

Homology search; Property pattern; Exon shuffling; Mosaic protein

1. INTRODUCTION

Nearly every week a new primary structure of an extracellular protein becomes available. Many of them are called mosaic proteins, because they result from exon shuffling [1] and contain therefore a set of different structural units, domains (for reviews see [2–4]). These domains have defined functions, which remain conserved when being shuffled into different proteins. Often the function is unknown and a comparative domain analysis may shed some light on the molecular basis of many seemingly unrelated extracellular proteins. Therefore, we have collected consensus patterns for known domains of extracellular mosaic proteins [5] and have looked for so far unidentified connections.

The best characterized, well-defined extracellular systems are the coagulation and the complement system, for which the domain assemblies of nearly all involved proteins are known [6,7]. Only for a few sequence segments of the many proteins in both cascades no counterpart could be found in other proteins. One example is a domain of the first complement component, in particular of the homologous subcomponents C1r and C1s. C1r and C1s have a typical mosaic structure (Fig. 1) containing the above mentioned domain twice separated by an epidermal growth factor (EGF)-related domain [8]. Recently, the C1r/C1s specific domains of unknown function were proposed to be homologous to a region in a sea urchin protein (uEGF) of the EGF superfamily, based on similar fragments that match the N-terminal half of the C1r/C1s domains [9]. Thus, the occurrence of this domain in further seemingly unrelated proteins may be anticipated. We used our property pattern method [10] to search for the described domain in other extracellular mosaic proteins.

Correspondence address: P. Bork, Central Institute of Molecular Biology, Department of Biomathematics, Robert-Rössle-Str. 10, O-1115 Berlin-Buch, Germany

Abbreviations: C1r (C1s), complement subcomponents C1r (C1s); BMP1, bone morphogenic protein 1; UVS.2, *Xenopus laevis* protein UVS.2; uEGF, sea urchin gene uEGF product; CASP, calcium-dependent serine protease isolated from malignant hamster embryo fibroblasts

2. MATERIALS AND METHODS

The multiple alignments were carried out by our program PULIGN [11]. The most conserved (reliable) regions [12] were extracted and a consensus was described by property patterns, which consider the conservation of defined steric and physicochemical amino acid properties in each position of the multiple alignment (for reviews of pattern recognition methods see [13,14]). Having assigned a vector of such properties to each amino acid, a position dependent weighting procedure is able to record those properties that are common in all amino acids of a position, and those which never occur in any amino acid of that position [10,11]. Possible deviations from this kind of consensus pattern as well as insertions or deletions can be taken into account. The number of mismatches corresponds to amino acid properties deviating from the pattern. Gap penalties depend on the degree of property conservation and the occurrence of deletions/insertions in the multiple alignment (learning set).

For homology search in MIPSX (a merged database, containing 32 800 protein sequences in release 17) we used a Needleman–Wunsch algorithm [15] modified for use of the property patterns. When related domains are found a second alignment including the newly detected sequence segments has to be carried out and the whole procedure is started again. To verify the results every domain detected by our method was compared by FASTP [16] with all proteins of the MIPSX database.

Published by Elsevier Science Publishers B. V.

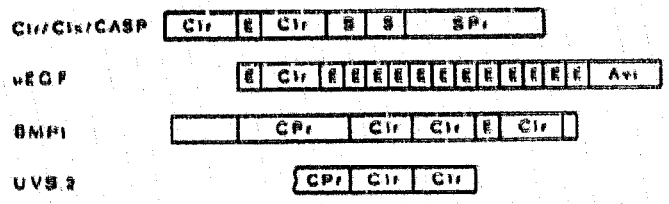

Fig. 1. Domain assembly of proteins containing the studied sequence repeat (C1r). E = EGF related domain; S = short consensus repeat (e.g. [7]); Spr = serine protease; Avi = Avidin like domain; and CPr = segments similar to an unusual crayfish protease [23]. White boxes represent domains for which no similar segment could be found in proteins with other domain compositions. Only the short form of uEGF is shown. Another splicing variant containing 6 additional EGF-like domains was also sequenced [9].

3. RESULTS AND DISCUSSION

In a first step the property patterns were derived from a multiple alignment comprising the 2 domains of C1r (residues 16–127, 195–301 [17]) and C1s (residues 11–122, 175–284 [18,19]) as well as the fragment of uEGF (residues 54–114 of the cDNA clone uEGF-7 [9]). A subsequent screening of the MIPSX sequence database revealed similarities with 2 regions of a calcium-dependent serine protease (CASP) from malignant hamster embryo fibroblasts capable of degrading extracellular matrix proteins [19] as well as 2 repeats of a human bone morphogenic protein (BMP1 [20]). By including these segments into the learning set (i.e. into the multiple alignment) an improved property pattern could be derived and a second run identified a further, 3rd repeat in BMP1 as well as 2 segments of a *Xenopus laevis* protein called UVS.2 [21]. A 3rd run did not detect any further related domains.

The average pairwise amino acid identities range between 25% and 40% (Fig. 2). Because of the high similarity between the CASP and C1s domains the whole proteins were compared and they show an overall homology of about 79% amino acid identity. Thus, they share the same domain composition (Fig. 1). We suppose that CASP is a C1s-equivalent in hamster, because some known functional features of CASP like calcium dependence and capability of degrading extracellular matrix proteins [19,22] also agree with those in human C1s.

The alignment of all domains is shown in Fig. 3. Only in 7 out of 110 positions amino acids remain absolutely conserved in the 12 different domains (underlined in the consensus line of Fig. 3). There are a lot of positions, where certain amino acids are clearly preferred or to which defined amino acid properties can be assigned (Fig. 3). The most conserved region is located in the N-terminal part of the domain and contains the consensus peptide SPpYPpxY (for nomenclature see Fig. 3; x =

	C1r (1)	C1r (2)	C1s (1)	C1s (2)	CASP(1)	CASP(2)	uEGF	BMP1(1)	BMP1(2)	BMP1(3)	UVS2(1)	UVS2(2)
C1r (1)	569	29.5%	38.7%	30.0%	42.3%	30.0%	(14.9%)	25.8%	27.6%	32.3%	27.6%	30.6
C1r (2)	108	565	25.8%	37.2%	25.8%	32.6%	(22.2%)	25.5%	31.1%	26.4%	28.3%	22.7%
C1s (1)	230	129	572	26.4%	82.0%	25.2%	(24.3%)	29.5%	32.3%	30.4%	27.6%	24.1%
C1s (2)	124	218	103	563	28.2%	80.0%	(38.5%)	21.4%	35.3%	30.7%	26.0%	29.0%
CASP(1)	228	115	493	98	554	27.3%	(18.7%)	32.3%	32.3%	26.7%	28.6%	25.9%
CASP(2)	120	199	96	486	92	552	(36.7%)	21.4%	34.4%	29.8%	27.0%	29.9%
uEGF	(51)	(75)	(44)	(92)	(45)	(82)	(274)	(21.0%)	(30.5%)	(30.5%)	(32.4%)	(24.8%)
BMP1(1)	127	103	123	86	123	96	(85)	607	41.5%	34.9%	37.7%	30.3%
BMP1(2)	144	143	168	130	150	119	(90)	276	574	42.4%	37.7%	35.1%
BMP1(3)	146	120	159	111	151	117	(61)	243	295	560	32.1%	34.1%
UVS2(1)	118	86	109	90	104	77	(91)	182	191	82	522	27.5%
UVS2(2)	147	125	135	131	142	115	(88)	194	188	181	158	540

Fig. 2. Pairwise similarities of the detected domains. *Upper right*: amino acid identities; *lower left*: optimized FASTP-scores. The maximum scores are shown in the diagonal. Optimized FASTP-scores consider similar amino acids as well as insertion/deletions and therefore can be used for comparison. Relatively low scores between some of the domains mirror undetected relations in the C-terminal parts of the respective domains. The scores between uEGF and all other domains are shown in parentheses, because values for only the N-terminal part could be established.

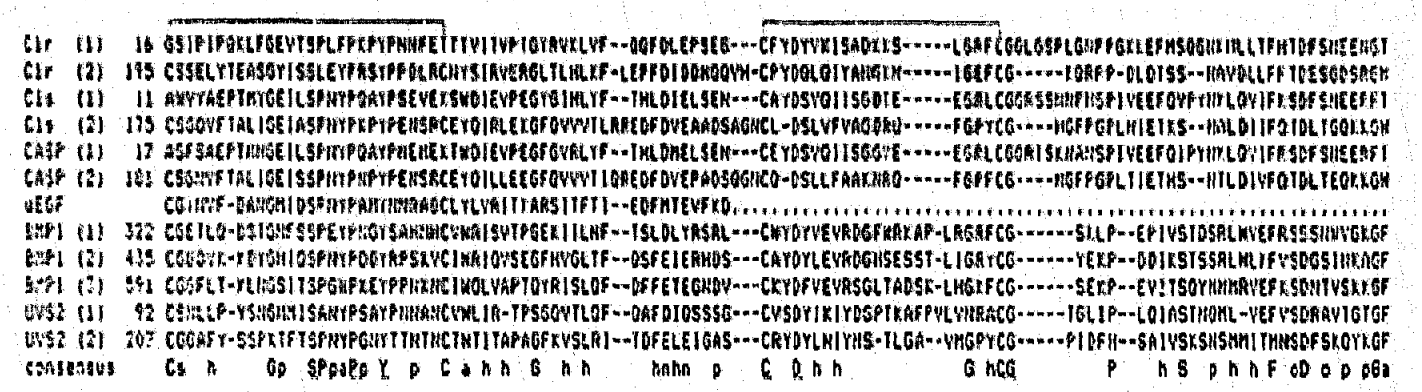

Fig. 3. Multiple alignment of the detected domains. In the 'consensus' line, capital letters mark extremely conserved amino acids (they must be present in at least 10 sequences). The underlined ones are absolutely conserved in all sequences. Lower case letters are used to show the conservation of amino acid properties (s = small, h = hydrophobic, p = polar, a = aromatic, n = negatively charged, o = presence of a hydroxyl group). The probable linkage of the 2 disulfide bonds is also indicated.

any amino acid), which could not be found in any other sequence segment of the MIPSX protein sequence database. Such a highly conserved consensus peptide in otherwise more distantly related segments points to a defined function, supported by the fact that the first of the 2 disulfide bonds stabilizes the folding pattern in the respective region.

In addition to the C1r/C1s related repeats, proteins BMP1 and UVS.2 have a common region homologous to a unique protease isolated from crayfish (Fig. 1). The mechanism of this protease with unusual cleavage specificity is not known and no sequence similarity to any other protease family has been found so far [23]. All domains of BMP1 and UVS.2, but especially those related to C1r/C1s, are more similar to each other than to those of the other proteins (Figs 2, 3). Thus, the studied domains can be divided into 2 subgroups. In particular, the positions around the 2nd disulfide bond differ in both groups. Even though the corresponding segment in uEGF is not available for comparison and the highest number of identical amino acids was recorded for the 2nd domain in C1s, the uEGF fragment shows more similarity to BMP1 and UVS.2 domains in positions, where both subgroups can be distinguished (Fig. 3).

With the exception of C1r and C1s no defined physiological function could be related to the studied proteins. BMP1 is one of 3 proteins purified from a fraction with a capability of inducing bone formation [20,24]. The other 2 bone morphogenic proteins also represent mosaic proteins and belong to the tumor growth factor β superfamily. It is likely that bone morphogenesis requires the combined action of many factors with multiple interactions. For defined binding and regulation mosaic structures are favorable. UVS.2 has the same domain assembly as the central part of BMP1 (Fig. 1), but was described to be exclusively present in the anterior neural fold of neurula stage embryos from *X. laevis* [21]. Interestingly, uEGF also represents a developmentally regulated protein, most prevalently located in embryonic ectoderm (as UVS.2), but also accumulated in the primary mesenchyme [9,25]. Since UVS.2 lacks EGF-like domains which are thought to play an important role in cell growth, differentiation and development [26,27], we conclude that the studied C1r/C1s related domain has specific functions in developmental processes (at least in the subgroup comprising BMP1, UVS.2 and uEGF). In this connection, it should be mentioned that CASP was also isolated from cells in an embryonal stage.

Although the function of the domain common in C1s, C1r, CASP, BMP1, UVS.2 and uEGF remains speculative, the detected homologies represent a further step in solving the puzzle of domains in extracellular mosaic proteins. The more homologies that can be identified among the various domains, the better will be the understanding of their functional network.

Acknowledgements: The author is indebted to Prof. J.C. Reich and Dr K. Rohde for critical reading of the manuscript and thanks E. Wolf for technical assistance.

REFERENCES

[1] Gilbert, W. (1978) Nature 271, 501.
[2] Doolittle, R.F. (1985) Trends Biochem. Sci. 10, 233–237.
[3] Patthy, L. (1987) FEBS Lett. 214, 1–7.
[4] Patthy, L. (1988) J. Mol. Biol. 202, 689–696.
[5] Bork, P. (1989) FEBS Lett. 257, 191–195.
[6] Furie, B. and Furie, B.C. (1988) Cell 53, 505–518.
[7] Reid, K.B.M. and Day, A.J. (1989) Immunol. today 10, 177–180.
[8] Leytus, S.P., Kurachi, K., Sakariassen, K.S. and Davie, E.W. (1986) Biochemistry 25, 4855–4863.
[9] Delgadillo-Reynoso, M.G., Rollo, D.R., Hursh, D.A. and Raff, R.A. (1989) J. Mol. Evol. 29, 314–327.
[10] Bork, P. and Grunwald, C. (1990) Eur. J. Biochem. 191, 347–358.
[11] Bork, P. and Rohde, K. (1990) Biochem. Biophys. Res. Commun. 171, 1319–1325.
[12] Vingron, M. and Argos, P. (1990) Prot. Eng. 3, 565–569.
[13] Taylor, W.R. (1988) Prot. Eng. 2, 77–86.

[14] Barton, G.J. and Sternberg, M.J.E. (1990) J. Mol. Biol. 212, 389–402.
[15] Needleman, S.B. and Wunsch, C.D. (1970) J. Mol. Biol. 48, 443–453.
[16] Lipman, D.J. and Pearson, W.R. (1985) Science 227, 1435–1441.
[17] Mackinnon, C.M., Carter, P.E., Smyth, S., Dunbar, B. and Fothergill, J.E. (1987) Eur. J. Biochem. 169, 547–553.
[18] Tosi, M., Duponchel, C., Meo, T., Julier, C. (1987) Biochemistry 26, 8516–8524.
[19] Konishita, H., Sakiyama, H., Tokunaga, K., Imajoh-Ohmi, S., Hamada, Y., Isono, K. and Sakiyama, S. (1989) FEBS Lett. 250, 411–415.
[20] Wozney, J.M., Rosen, V., Celeste, A.J., Mitsock, L.M., Whitters, M.J., Kriz, R.W., Hewick, R.M. and Wang, E. (1988) Science 242, 1528–1534.
[21] Sato, S.M. and Sargent, T.D. (1990) Dev. Biol. 137, 135–141.
[22] Sakiyama, H., Nishino, Y., Tanaka, T., Tomosawa, T., Kinoshita, H., Nagata, K., Chiba, K., Sakiyama, S. (1989) Biochim. Biophys. Acta 990, 156–161.
[23] Titani, K., Torff, H.-J., Hormel, S., Kumar, S., Walsh, K.A., Rödl, J., Neurath, H. and Willing, R. (1987) Biochemistry 26, 222–226.
[24] Wang, E.A., Rosen, V., Cordes, P., Hewick, R.M., Kriz, M.J., Luxenberg, D.P., Sibley, B.S. and Wozney, J.M. (1988) Proc. Natl. Acad. Sci. USA 85, 9484–9488.
[25] Hursh, D.A., Andrews, M.E. and Raff, R.A. (1987) Science 237, 1487–1490.
[26] Apella, E., Weber, I.T. and Blasi, F. (1988) FEBS Lett. 231, 1–4.
[27] Engel, J. (1989) FEBS Lett. 251, 1–6.