Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 10783–10787, November 1993
Developmental Biology

The precursor region of a protein active in sperm–egg fusion contains a metalloprotease and a disintegrin domain: Structural, functional, and evolutionary implications

(PH-30/spermatogenesis/snake venom/astacin/cell adhesion)

TYRA G. WOLFSBERG[†‡], J. FERNANDO BAZAN[‡§], CARL P. BLOBEL[†‡¶], DIANA G. MYLES[‖], PAUL PRIMAKOFF[‖], AND JUDITH M. WHITE[†‡**]

Departments of [†]Pharmacology and [‡]Biochemistry and Biophysics, University of California, San Francisco, CA 94143; and [‖]Department of Physiology, University of Connecticut Health Center, Farmington, CT 06030

Communicated by Bruce M. Alberts, August 4, 1993

ABSTRACT **PH-30, a sperm surface protein involved in sperm–egg fusion, is composed of two subunits, α and β, which are synthesized as precursors and processed, during sperm development, to yield the mature forms. The mature PH-30 α/β complex resembles certain viral fusion proteins in membrane topology and predicted binding and fusion functions. Furthermore, the mature subunits are similar in sequence to each other and to a family of disintegrin domain-containing snake venom proteins. We report here the sequences of the PH-30 α and β precursor regions. Their domain organizations are similar to each other and to precursors of snake venom metalloproteases and disintegrins. The α precursor region contains, from amino to carboxyl terminus, pro, metalloprotease, and disintegrin domains. The β precursor region contains pro and metalloprotease domains. Residues diagnostic of a catalytically active metalloprotease are present in the α, but not the β, precursor region. We propose that the active sites of the PH-30 α and snake venom metalloproteases are structurally similar to that of astacin. PH-30, acting through its metalloprotease and/or disintegrin domains, could be involved in sperm development as well as sperm–egg binding and fusion. Phylogenetic analysis indicates that PH-30 stems from a multidomain ancestral protein.**

PH-30, a guinea pig sperm surface protein, is a candidate sperm–egg membrane binding and fusion protein (1–5). The PH-30 subunits found on fertilization-competent sperm, mature α and mature β, share membrane topologies and other characteristics with viral binding and fusion proteins. The β subunit contains a potential receptor binding domain, a disintegrin domain, related to soluble integrin ligands found in snake venom. The α subunit contains a potential fusion peptide. In addition, the two subunits share sequence similarity.

Snake venom disintegrins derive from precursors that also contain zinc-dependent metalloprotease domains (6, 7). Interestingly, PH-30 α and β are present on testicular spermatogenic cells as larger precursors, termed here pro-α and pro-β (2). Here we show that the precursor regions of PH-30 α and β (the regions amino-terminal to the mature proteins and found on developing, but not fertilization-competent, sperm) share further amino acid identity with each other as well as with this family of metalloprotease and disintegrin domain-containing snake venom proteins.[††]

MATERIALS AND METHODS

Cloning. A portion of the α precursor region sequence was obtained from a PCR product generated by the nested RACE (rapid amplification of cDNA ends) protocol (3). The sequences of the remainder of the α precursor region and the entire β precursor region were determined from clones of α and β isolated at high stringency (3) from a guinea pig whole-testis cDNA library (8).

Northern Analysis. RNA was isolated from adult male guinea pig tissues (9), electrophoresed in a formaldehyde/agarose gel, and transferred and cross-linked to a Hybond-N nylon membrane (Amersham). High-stringency prehybridization and hybridization with PH-30 α and β ^{32}P-labeled DNA probes was carried out at 65°C in 5× standard saline citrate (SSC)/5× Denhardt's solution/0.1% SDS containing salmon sperm DNA at 0.2 mg/ml. The membrane was washed, 10 min per wash, in 2× SSC/0.1% SDS once at room temperature and twice at 65°C and then in 0.2× SSC/0.1% SDS twice at 65°C. Hybridization with a mouse β-actin probe was carried out in the same solution at 55°C, with identical wash conditions.

RESULTS AND DISCUSSION

The amino acid sequences of the PH-30 α and β precursor regions were deduced from cDNA sequences and are shown in Fig. 1. Following their signal sequences, the α and β precursor regions contain sequences similar to those in the prodomains of disintegrin domain-containing snake venom proteins, and then sequences which align with the snake venom zinc-dependent metalloprotease domain (Figs. 1 and 2). α contains the consensus active-site residues for a metalloprotease (see below); β does not. Following the metalloprotease domain, both proteins contain a disintegrin domain (Figs. 1 and 2). The cleavage site which generates mature α falls within the disintegrin domain (3) (arrows, Figs. 1 and 2). The cleavage site which generates mature β lies at the amino terminus of the disintegrin domain (3) (arrow before position 383, Figs. 1 and 2). The sequence alignment of mature PH-30 α and β with the snake venom proteins continues through the cysteine-rich domain (Figs. 1 and 2). No snake venom proteins include either the epidermal growth factor repeat or the transmembrane and cytoplasmic segments of α and β (Figs. 1 and 2). Additional mammalian genes encode proteins with domain organizations identical to those of PH-30 α and β (Figs. 1 and 2). EAP I, cloned from rat and monkey, is an androgen-regulated protein located on the apical surface of epididymal epithelial cells (13). Cyritestin is a mouse testis cDNA (GenBank accession no. X64227).

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.

[§]Present address: Department of Molecular Biology, DNAX, Palo Alto, CA 94304.
[¶]Present address: Department of Cellular Biochemistry and Biophysics, Sloan–Kettering Institute, New York, NY 10021.
[**]To whom reprint requests should be addressed.
[††]The sequences reported in this paper have been deposited in the GenBank data base (accession nos. Z11719 and Z11720).

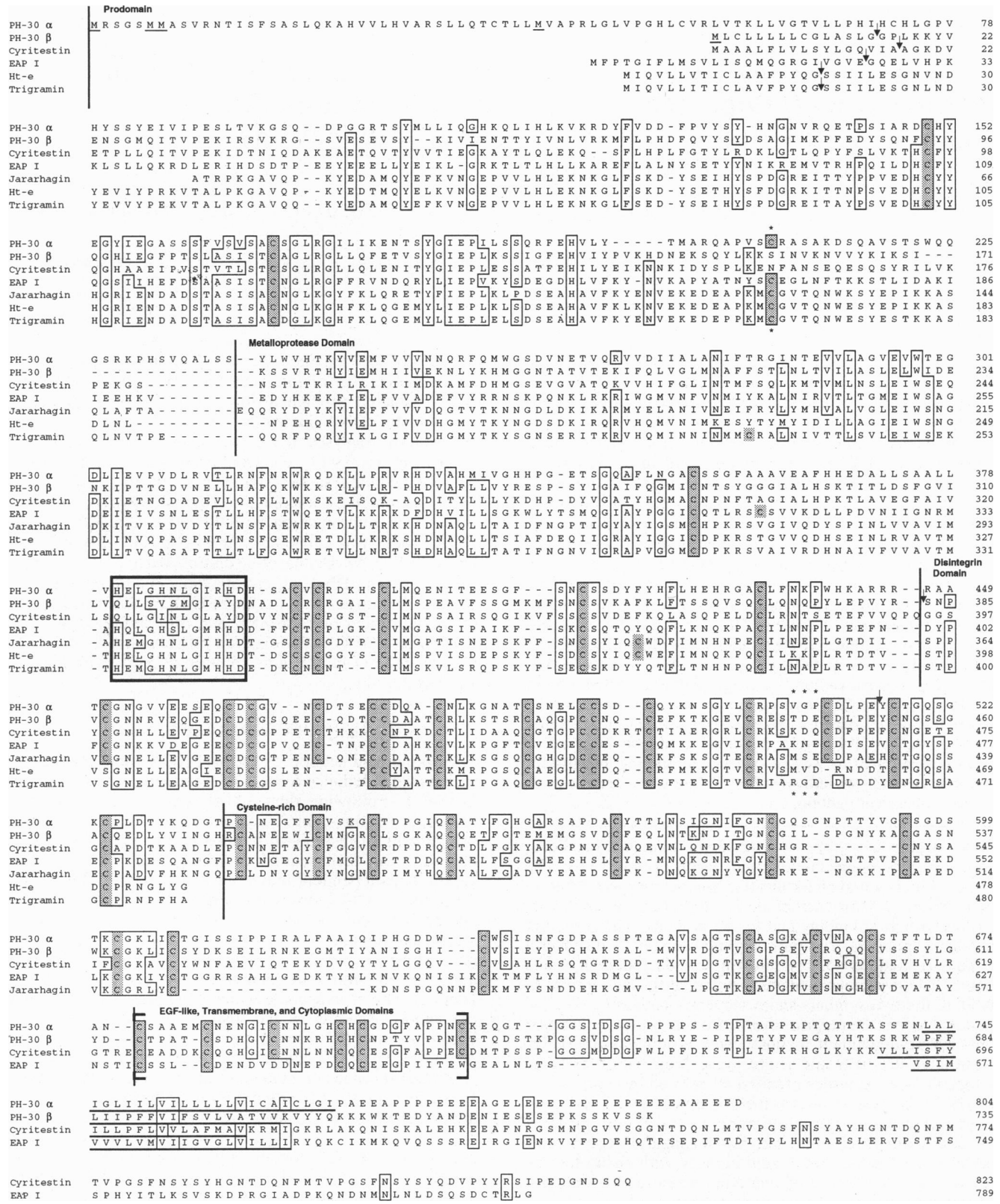

FIG. 1. Comparison of amino acid sequences. The entire coding sequence of each protein is depicted, with residues numbered consecutively. Cysteine residues are shaded. Amino acids are boxed at positions of >60% identity. The sequence of jararhagin is not complete at the 5′ end. Sequence similarities were found by using the program FASTA (3) with k-tuple = 1 to search the PIR (Release 36), Genpept (Release 76), and Swiss-Prot (Release 24) data bases. *Prodomain.* PH-30 α and β are 30% identical (excluding gaps) over this region. The first amino acid residue shown for α and β is that of the first in-frame methionine. α contains four potential start methionine residues (underlined), none of which is followed by an obvious signal sequence (10). The methionine at position 7 is encoded by an AUG codon in the most optimal context for translation initiation (11). The putative start methionine of β (underlined), encoded by an AUG codon in a good context to initiate translation (11), is followed by a potential signal sequence (10). Potential sites for signal-sequence cleavage are marked with arrows. Stars mark cysteine residues which may be involved in a matrix metalloprotease-like "cysteine switch" (12) activation of the metalloprotease domain. *Metalloprotease domain.* PH-30 α and β are 27% identical (excluding gaps) over this region. The consensus snake venom metalloprotease active-site sequence,

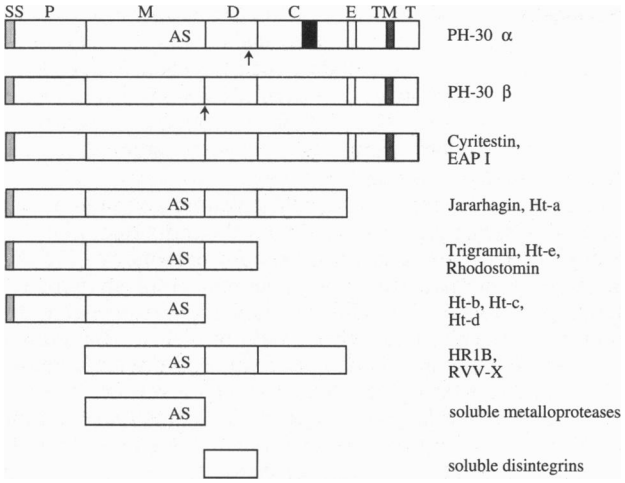

FIG. 2. Domain organization of disintegrin and metalloprotease domain-containing proteins. SS, signal sequence; P, prodomain; M, zinc-dependent metalloprotease domain; D, disintegrin domain; C, cysteine-rich domain; E, epidermal growth factor-like repeat; TM, transmembrane domain; T, cytoplasmic tail; AS, snake venom/PH-30 α metalloprotease active-site consensus sequence HEXGHXLGXXHD. Mammalian members of this protein family sequenced to date include all domains. Snake venoms contain three subfamilies of this protein family, all of which begin with a signal sequence, but which terminate with either the metalloprotease, the disintegrin, or the cysteine-rich domain. The sequences of mammalian PH-30 α and β, cyritestin (GenBank accession no. X64227), and EAP I (13), as well as the sequences of snake venom jararhagin (14), trigramin (15), Ht-e (12), rhodostomin (see ref. 14), and Ht-a, -b, -c, and -d (J. Fox, personal communication) are known from cDNA cloning. The sequences of the snake venom proteins HR1B (see ref. 7), RVV-X (16), and the soluble (see ref. 7) metalloproteases and disintegrins are known from direct amino acid sequencing only. The amino termini of mature PH-30 α and β, as determined by amino acid sequencing (3), are indicated by arrows. The potential fusion peptide of PH-30 α is indicated by a black box.

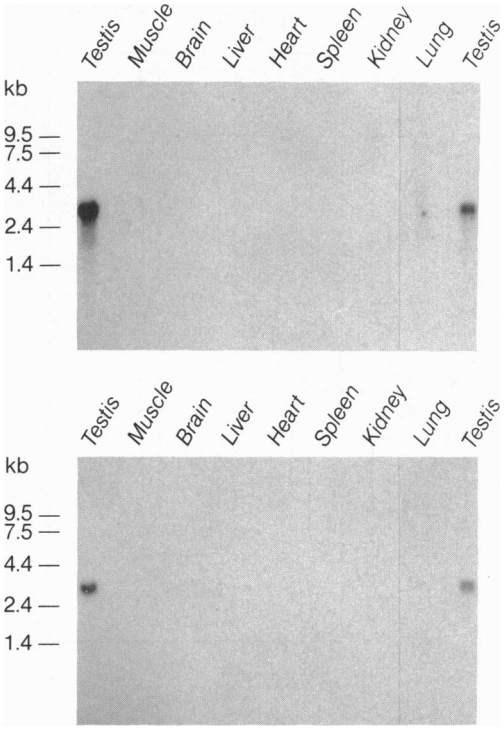

FIG. 3. Tissue distribution of PH-30 α (*Upper*) and β (*Lower*). A Northern blot containing total RNA from eight adult male guinea pig tissues was probed at high stringency with PH-30 α or PH-30 β DNA. Both hybridized to separate RNA species of ≈3000 nt in testis. These sizes are in good agreement with the lengths of PH-30 cDNAs (≈3300 nt for α and ≈2900 nt for β). All samples contained about 10 μg of RNA, except for the two rightmost lanes on each gel (lung and testis), which contained about 1 μg of RNA. The exposure time for samples containing less RNA was approximately 8-fold higher. β-Actin message was intact in all samples (data not shown). Although Northern analysis revealed only one mRNA species for both α and β, we have found sequence variations in the 3′ untranslated regions of PH-30 α (T.G.W., unpublished data) and PH-30 β (C.P.B. and P. D. Straight, unpublished data).

Neither is the respective species homologue of PH-30 α or β (ref. 13; T.G.W., unpublished data). A Northern blot of guinea pig tissue RNA, probed at high stringency, suggests that PH-30 α and β are testis-specific in the adult (Fig. 3).

Like the mature PH-30 β subunit, the PH-30 α precursor region contains a disintegrin domain. Soluble snake venom disintegrins which contain an RGD binding sequence interact with integrins on platelets and other cells (reviewed in refs. 6 and 7), presumably through a 13-amino acid RGD-containing loop (17, 18). Disintegrins which are linked to a carboxyl-terminal cysteine-rich domain (Fig. 2) lack the RGD motif, but contain instead a unique tripeptide as well as an adjacent cysteine (Fig. 1, stars). The latter snake venom disintegrins, as well as the disintegrin domains of EAP I and mature PH-30 β, maintain a negatively charged residue in the position of the RGD aspartic acid. The disintegrin domains of PH-30 α and cyritestin lack an acidic residue in this position (Fig. 1, stars).

The PH-30 α precursor region contains a metalloprotease-like domain. The sequence HEXXH is the active-site signature of a large class of zinc-dependent metalloproteases (19). As the structure of bacterial thermolysin demonstrates, the glutamic acid probably acts as a catalytic base, and the two histidines (underlined), which are adjacent on the same face of a helix, help ligate the zinc (Fig. 4) (20). The PH-30 α and sequence-similar snake venom metalloprotease domains feature an extended active-site signature motif (Fig. 1, box) which is similar to the conserved active-site sequence, HEXXHXX*G*XXHE, of a family of zinc-dependent metalloproteases homologous to the crayfish enzyme astacin (22, 26). The x-ray structure of astacin indicates that a glycine (in italics) terminates a thermolysin-like HEXXH-containing helix by making a tight turn, enabling the third underlined histidine to serve as a third zinc ligand. We predict that the histidine, glutamate, and glycine residues of the PH-30 α and snake venom metalloprotease active sites play analogous structural roles. The PH-30 β and cyritestin metalloprotease domains lack the critical histidine and glutamate residues but maintain the conserved glycine and (carboxyl-terminal) aspartic acid (Fig. 1, box). Thus, these latter proteins, although probably not active metalloproteases, may fold similarly.

Outside of the active-site consensus motif, the PH-30 α and snake venom metalloprotease domains share little sequence similarity with the astacin-like proteases. To assess the possibility that HEXXH-containing metalloproteases share a

HEXGHXLGXXHD, is boxed. PH-30 α contains the metalloprotease consensus sequence; PH-30 β and cyritestin do not. EAP I contains all residues of the consensus sequence except for the catalytic glutamic acid, suggesting that it may be able to ligate a zinc ion but be catalytically inert. *Disintegrin domain.* PH-30 α and β are 47% identical (excluding gaps) over this region. Stars mark the position of the RGD tripeptide of trigramin. Arrows indicate the amino termini of mature PH-30 α and β. *Cysteine-rich domain.* PH-30 α and β are 29% identical (excluding gaps) over this region. *EGF-like, transmembrane, and cytoplasmic domains.* The epidermal growth factor (EGF) repeat is bracketed. The potential transmembrane domains are underlined.

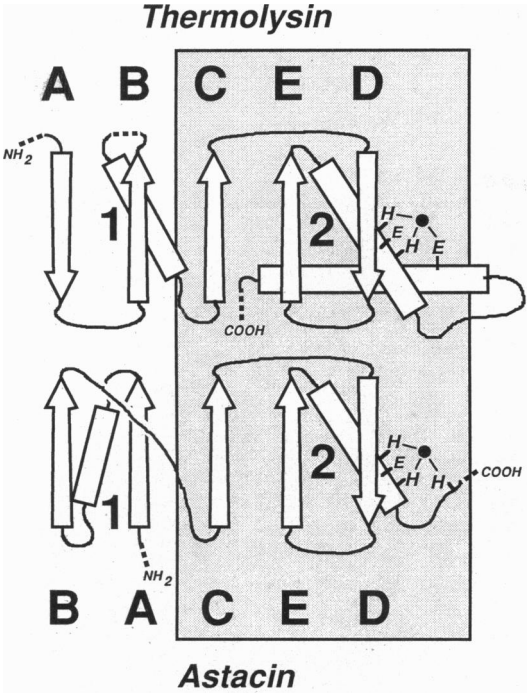

FIG. 4. The zinc-dependent metalloproteases of the thermolysin and astacin classes share an ≈50-amino acid core structural motif of three β-strands and a catalytic α-helix. Shown are the principal β-sheets of thermolysin (20, 21) and those we have located for astacin [by inspection of the ribbon diagram of astacin (figure 1*b* of ref. 22) and by structural prediction algorithms (23, 24)] with distinct ABCED and BACED strand topologies [corresponding to +1, +1*x*, +2, −1 and −1*x*, +2*x*, +2, −1 β-strand orders in Richardson notation (25), respectively]. A shaded box delineates the shared CED (+2, −1) three-β-strand motif, and the subsequent α-helix 2 that carries the signature HEXXH catalytic groups. The positions of structurally equivalent histidine and distal glutamic acid ligands to the zinc (black circle) are noted. Chain segments amino- and carboxyl-terminal to this common ≈50-amino acid supersecondary structure exhibit different chain topologies.

core framework in which the active site is embedded, we compared the x-ray structures of thermolysin-like bacterial enzymes (20, 21) and astacin (22). Although there is meager sequence identity, both structures are globally similar with an active site sandwiched between an amino-terminal β-sheet domain and a carboxyl-terminal helix/coil-rich structure. Only one discrete section of continuous chain is topologically conserved, an ≈50-amino acid core (Fig. 4) consisting of three β-strands (sheets C, E, and D) and the following catalytic α-helix (helix 2) that carries the HEXXH sequence. The existence of this core fold, as well as the astacin-like active-site geometry, in the PH-30 α and snake venom metalloproteases is supported by the recently determined x-ray structure of a protease from the venom of the snake *Crotalus adamanteus* (L. Kress and W. Bode, personal communication).

Analysis of the sequences of the snake venom and PH-30 α metalloprotease domains suggests at least two activation mechanisms. In astacin, the final glutamic acid of the active-site consensus sequence interacts with the amino-terminal ammonium group of the mature, active protease (22). The conserved aspartic acid of the mammalian and snake venom metalloprotease domains (Fig. 1, box) may undergo a similar interaction. However, an exactly analogous mechanism would require that astacin and the PH-30 α and snake venom metalloproteases have a similar protein fold outside of the conserved core. But, like thermolysin, the amino- and carboxyl-terminal extensions of the PH-30 α and snake venom enzymes may be structurally distinct from those of astacin. Alternatively, activation of the snake venom metalloproteases could occur by a matrix metalloprotease-like "cysteine switch" mechanism, through a conserved cysteine in the prodomain (12). The PH-30 α prodomain features a cysteine in a similar location (Fig. 1, star), although it lacks other postulated consensus residues.

To trace the history of the snake venom and mammalian PH-30-like sequences, we subjected the individual domains to phylogenetic analysis by maximum parsimony (27, 28). The trees generated from the prodomain (not shown), the metalloprotease domain (Fig. 5), and the cysteine-rich domain (not shown) contain three main branches. One branch contains PH-30 α and β and cyritestin, one branch contains EAP I, and one branch contains the snake venom proteins. That the trees are largely congruent indicates that the individual domains are evolving simultaneously and were probably all present in the progenitor gene.

Snake venoms contain soluble disintegrins and metalloproteases; by analogy, PH-30 α and β might be proteolytically processed at interdomain boundaries. The first proteolytic processing event for β, the transition from pro-β (≈88 kDa) to pro-β* (75 kDa) (2), is consistent with removal of the prodomain. Removal of the β metalloprotease-like domain and subsequent exposure of its disintegrin domain may render epididymal sperm fertilization-competent (2, 29). Al-

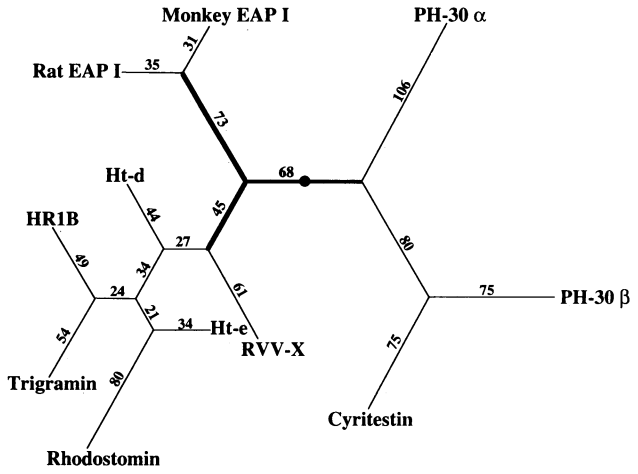

FIG. 5. Phylogenetic analysis of the snake venom/PH-30 zinc-dependent metalloprotease domain from a subset of available protease sequences. Numbers on the branches are proportional to the number of nucleotide changes between sequences; branch lengths are drawn to scale. Heavy lines indicate the three main branches of the tree: one to PH-30 and cyritestin, one to EAP I, and one to the snake venom proteins. EAP I is evolving separately from PH-30 and cyritestin, as it lies on a separate branch of the tree and contains a different linker module between its cysteine-rich and transmembrane domains (Fig. 1). The zinc-dependent metalloprotease domains from snake venom, Ht-e and Ht-d (12) (*Crotalus atrox*), HR1B (see ref. 7) (*Trimeresurus flavoviridis*), RVV-X (16) (*Vipera russelli*), rhodostomin (see ref. 14) (*Calloselasma rhodostoma*), and trigramin (15) (*Trimeresurus gramineus*), and mammalian EAP I (13), cyritestin (GenBank accession no. X64227), and PH-30 α and β were analyzed by the program PAUP (27). The tree was generated by a heuristic search algorithm, with 25 repetitions of random stepwise sequence addition. Tree validity was tested with the bootstrap approach, using 100 replications of a heuristic search, itself composed of 10 replications of random sequence addition. The tree shown is the most parsimonious, and conforms to criteria specified in ref. 28. Separate trees generated from a sequence alignment of critical residues in the conserved core domain (see Fig. 4) for bacterial thermolysin, astacin, as well as the proteins shown here, indicate that the root of this tree (the black dot) falls at the midpoint of the branch which joins PH-30 and cyritestin with EAP I and the snake venom proteins (J.F.B., unpublished data).

though we have not detected release of the α metalloprotease domain during the developmental processing of pro-α (2), there is a potential dibasic cleavage site ($R_4 \downarrow A$) in the short linker between the α metalloprotease and disintegrin domains.

The metalloprotease domain of the PH-30 α precursor region may be active in early spermatogenesis in the testis, either before or after proteolytic processing of pro-α. The protease could help spermatocytes cross the tight junction between adjacent Sertoli cells. It could be involved in the restructuring which occurs as sperm migrate toward the lumen of the seminiferous tubules. It might release spermatozoa into the lumen of the seminiferous tubules. It could be autocatalytic, as some snake venom proteins are autocatalytically processed at the boundary between the metalloprotease and disintegrin domains (12). Or it might process the β precursor.

The disintegrin domains of PH-30 pro-α or pro-β may also play roles in spermatogenesis. Integrins appear to be present at sites of attachment between spermatids and neighboring Sertoli cells (30). Interaction of PH-30 disintegrins with Sertoli cell integrins may be partially responsible for the adhesion of spermatogenic cells to Sertoli cells.

In summary, we have shown that the precursor regions of PH-30 α and β contain further similarity to each other as well as to a family of snake venom proteins which contain metalloprotease and disintegrin domains. PH-30 α and β are multidomain proteins which contain pro, metalloprotease, disintegrin, cysteine-rich, transmembrane, and cytoplasmic domains (Fig. 2). Their domain organization defines a family of mammalian integral membrane proteins, which now includes additional members [cyritestin, EAP I, MS2 (GenBank accession no. X13335), HUMORF09 (GenBank accession no. D14665), and T.G.W. and P. D. Straight, unpublished data]. The ancestors of these proteins may be the progenitor molecules from which the snake venom proteins evolved. These ancestors probably exist at least as far back as invertebrates, as a PH-30 homologue has been found in *Caenorhabditis elegans* (B. Podbilewicz and J. G. White, personal communication). Furthermore, PH-30 appears to be a multifunctional protein. The potential disintegrin and zinc-dependent metalloprotease domains in the precursor regions may play vital roles in early spermatogenesis. The disintegrin domain and the potential fusion peptide in the mature subunits may be necessary for the binding and fusion of sperm and egg plasma membranes. The putative functions of the PH-30 α precursor region suggest novel contraceptive strategies. Agents which inhibit metalloprotease or disintegrin domain activity may interfere with sperm development and be useful as male contraceptives.

We thank M. Skinner, C. Damsky, and G. Kemble for discussions; P. D. Straight for technical assistance; G. Gerton for the guinea pig testis cDNA library; and J. Fox, L. Kress, W. B. Bode, B. Podbilewicz, and J. G. White for communicating results prior to publication. This work was supported by grants to J.M.W. from the Muscular Dystrophy Association and the National Institutes of Health (GM48739). T.G.W. was a predoctoral fellow of the National Science Foundation.

1. Primakoff, P., Hyatt, H. & Tredick-Kline, J. (1987) *J. Cell Biol.* **104,** 141–149.
2. Blobel, C. P., Myles, D. G., Primakoff, P. & White, J. M. (1990) *J. Cell Biol.* **111,** 69–78.
3. Blobel, C. P., Wolfsberg, T. G., Turck, C. W., Myles, D. G., Primakoff, P. & White, J. M. (1992) *Nature (London)* **356,** 248–252.
4. White, J. M. (1992) *Science* **258,** 917–924.
5. Myles, D. G., Kimmel, L. H., Blobel, C., Turck, C., White, J. & Primakoff, P. (1992) *J. Cell Biol.* **3,** 212a (abstr.).
6. Blobel, C. P. & White, J. M. (1992) *Curr. Opin. Cell Biol.* **4,** 760–765.
7. Kini, R. M. & Evans, H. J. (1992) *Toxicon* **30,** 265–293.
8. Baba, T., Hoff, H. B. I., Nemoto, H., Lee, H., Orth, J., Arai, Y. & Gerton, G. L. (1993) *Mol. Reprod. Dev.* **34,** 233–243.
9. Shackleford, G. M. & Varmus, H. E. (1987) *Cell* **50,** 89–95.
10. von Heijne, G. (1986) *Nucleic Acids Res.* **14,** 4683–4690.
11. Kozak, M. (1991) *J. Biol. Chem.* **266,** 19867–19870.
12. Hite, L. A., Shannon, J. D., Bjarnason, J. B. & Fox, J. W. (1992) *Biochemistry* **31,** 6203–6211.
13. Perry, A. C. F., Jones, R., Barker, P. J. & Hall, L. (1992) *Biochem. J.* **286,** 671–675.
14. Paine, M. J. I., Desmond, H. P., Theakston, R. D. G. & Crampton, J. M. (1992) *J. Biol. Chem.* **267,** 22869–22876.
15. Neeper, M. P. & Jacobsen, M. A. (1990) *Nucleic Acids Res.* **18,** 4255.
16. Takeya, H., Nishida, S., Miyata, T., Kawada, S., Saisaka, Y., Morita, T. & Iwanaga, S. (1992) *J. Biol. Chem.* **267,** 14109–14117.
17. Saudek, V., Atkinson, R. A. & Pelton, J. T. (1991) *Biochemistry* **30,** 7369–7372.
18. Adler, M., Lazarus, R. A., Dennis, M. S. & Wagner, G. (1991) *Science* **253,** 445–448.
19. Jiang, W. & Bond, J. S. (1992) *FEBS Lett.* **312,** 110–114.
20. Holmes, M. A. & Matthews, B. W. (1982) *J. Mol. Biol.* **160,** 623–639.
21. Holland, D. R., Tronrud, D. E., Pley, H. W., Flaherty, K. M., Stark, W., Jansonius, J. N., McKay, D. B. & Matthews, B. W. (1992) *Biochemistry* **31,** 11310–11316.
22. Bode, W., Gomis-Rueth, F. X., Huber, R., Zwilling, R. & Stoecker, W. (1992) *Nature (London)* **358,** 164–167.
23. Rost, B. & Sander, C. (1992) *Nature (London)* **360,** 540.
24. Colloc'h, N. & Cohen, F. E. (1991) *J. Mol. Biol.* **221,** 603–613.
25. Richardson, J. S. (1981) *Adv. Protein Chem.* **34,** 167–339.
26. Gomis-Rueth, F. X., Stoecker, W., Huber, R., Zwilling, R. & Bode, W. (1993) *J. Mol. Biol.* **229,** 945–968.
27. Swofford, D. L. (1991) *PAUP: Phylogenetic Analysis Using Parsimony, Version 3.0s* (Illinois Natural History Survey, Champaign).
28. Stewart, C.-B. (1993) *Nature (London)* **361,** 603–607.
29. Hunnicutt, G. R., Beardsley, J. R. & Myles, D. G. (1990) *J. Cell Biol.* **111,** 361a (abstr.).
30. Pfeiffer, D. C., Dedhar, S., Byers, S. W. & Vogl, A. W. (1991) *J. Cell Biol.* **115,** 482a (abstr.).