The PAN module: the N-terminal domains of plasminogen and hepatocyte growth factor are homologous with the apple domains of the prekallikrein family and with a novel domain found in numerous nematode proteins

Hedvig Tordai, László Bányai, László Patthy*

Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, PO Box 7, H-1518 Budapest, Hungary

Received 29 July 1999; received in revised form 11 October 1999

Abstract Based on homology search and structure prediction methods we show that (1) the N-terminal N domains of members of the plasminogen/hepatocyte growth factor family, (2) the apple domains of the plasma prekallikrein/coagulation factor XI family, and (3) domains of various nematode proteins belong to the same module superfamily, hereafter referred to as the PAN module. The patterns of conserved residues correspond to secondary structural elements of the known three-dimensional structure of hepatocyte growth factor N domain, therefore we predict a similar fold for all members of this superfamily. Based on available functional informations on apple domains and N domains, it is clear that PAN modules have significant functional versatility, they fulfill diverse biological functions by mediating protein-protein or protein-carbohydrate interactions.

© 1999 Federation of European Biochemical Societies.

Key words: Apple module; Prekallikrein; Plasminogen; Hepatocyte growth factor; Coagulation factor XI

1. Introduction

Plasminogens, hepatocyte growth factors (HGFs) and macrophage stimulating proteins (MSP) are modular proteins belonging to the trypsin family of serine proteinases (Fig. 1). Their non-protease parts largely consist of kringle modules: there are five tandem kringle modules in the case of plasminogens, and four in the case of hepatocyte growth factors and macrophage stimulating proteins. At their N-terminal ends, these proteins contain a hairpin-loop structure with a characteristic pattern of four cysteines. Since this region of plasminogen is removed concomitant with its conversion to plasmin [1], it has been frequently referred to as preactivation peptide (PAP) or as N-terminal peptide (NTP) of plasminogen. These names are somewhat misleading since this domain has a compact fold. The corresponding N domain of hepatocyte growth factor contains seven β-strands, one α-helix and two short pieces of more irregular helical structure, the central feature of the domain being a five-stranded, antiparallel β-sheet [2,3]. The structural and functional independence of the N domains of the plasminogen/HGF family is underlined by the fact that there are some plasminogen-related genes (PRGs) which express proteins consisting of only a signal peptide and the N domain [4–6].

Studies on the genes of plasminogen [7], hepatocyte growth factor [8] and macrophage stimulating protein [9] have revealed that their genomic organization reflects their domain organization: at the boundaries of all kringle modules and at both boundaries of the N-terminal domain, phase 1 introns are found. Thus the N domain also satisfies the criteria for a mobile class 1-1 module (i.e. the module is flanked by phase 1 introns), nevertheless it has not been identified previously outside the family of plasminogen-related genes. The data described in the present work suggest that this class 1-1 module has actually been used for the construction of various modular proteins, including numerous nematode proteins as well as coagulation factor XI and plasma prekallikrein.

2. Materials and methods

Detection of distant homologies was carried out with the consensus sequence method [10,11]. The principle of this procedure is that in the case of distant homologues scores come primarily from positions that are characteristically conserved in the family. Accordingly, in this procedure the evolutionary information characterizing the given family is first condensed into consensus sequences in which low scoring positions are suppressed thereby sharpening key features of the family. When searching databases with such consensus sequences 'bad hits' with variable regions are eliminated thereby increasing the discriminating power of the consensus sequence.

In the present work multiple alignments for the three subgroups of the PAN module family (those of nematode proteins, plasminogen-related proteins, prekallikrein/factor XI) were constructed using Clustal W [12] and consensus sequences were defined for each group according to the consensus sequence procedure [10]. In the present work, the consensus sequences for the N-terminal domain of plasminogen-related proteins (conN in Fig. 2), for apple domains of prekallikrein and factor XI (conA in Fig. 2), for the tandem internal repeats of the different nematode proteins (conc16d9_1, conc29e6_1, conc30h6.5, conc34g6_6, conc52b11_1, conf38e11_4, conf41a4_1, conf52b11_3, conh42k12_3, conr07a4_4 in Fig. 2) were defined by including residues chemically conserved in more than 50% of the sequences. For the construction of these consensus sequences, amino acids were grouped as follows: Val, Ile, Leu, Met; Tyr, Phe, Trp; Ser, Thr; Asp, Asn, Glu, Gln; Arg, Lys; Cys; Gly; Pro; Ala; His. Unified multiple alignments of PAN modules were constructed using the consensus sequence-based progressive iterative alignment procedure and consensus sequences were also derived from these multiple alignments [10,11]. Iterative searches [11] of the SWISS-PROT, PIR and SP-TREMBL protein databases were performed with the FastA programme of the GCG Wisconsin Package, Version 9.1, using blosum40, blosum50, blosum62, pam120 and pam250 scoring matrices. To evaluate the significance of the results, *E*() scores were calculated to estimate the number of sequences that would be expected to produce, purely by chance, scores greater than or equal to the scores obtained in the search. Sequences with *E*() scores lower than 0.1 were considered to be homologous to the query sequence.

Secondary structure prediction based on multiple alignments of apple domains of prekallikreins and factor XI was carried out with described procedures [13–15] using the PHD mail server. The PHD

*Corresponding author. Fax: (36) (1) 4665-465.
E-mail: patthy@enzim.ha

0014-5793/99/$20.00 © 1999 Federation of European Biochemical Societies. All rights reserved.
PII: S0014-5793(99)01416-7

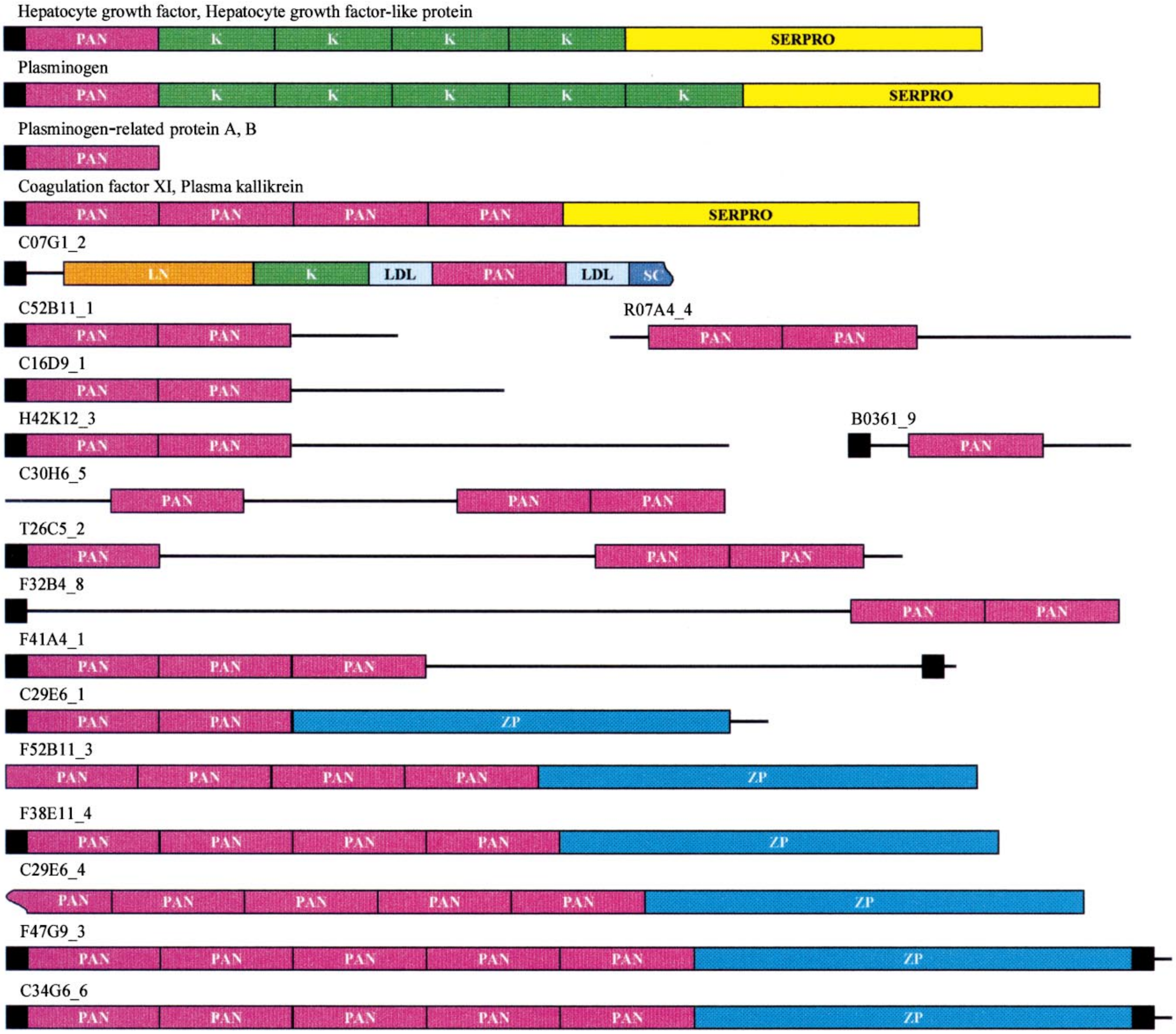

Fig. 1. Modular architecture of proteins containing the PAN module. The boxes representing domains of these modular proteins are drawn to scale. Accession numbers for sequences of these proteins are listed in the legend to Fig. 2. Note that the predicted sequences of some *C. elegans* proteins are probably incomplete (indicated by incomplete modules at their N-terminal and C-terminal ends, or the absence of signal peptides). The black boxes represent signal peptides or transmembrane helices, the colored boxes represent domains. The abbreviations of domains are: PAN, PAN module; K, kringle module; SERPRO, serine-protease domain; LN, C-type lectin module; LDL, LDL-receptor type A module; SC, scavenger receptor module; ZP, zone pellucida protein module.

(profile network prediction Heidelberg) server predicts secondary structural elements by evaluating the relative probabilities that a given segment can be assigned to helix, strand or loop. The estimated accuracy of this multiple alignment-based method for the correct prediction of secondary elements is about 72%.

3. Results and discussion

In the present work consensus sequences were determined for the N domains of plasminogen-related genes (e.g. conN in Fig. 2) and databases were searched to detect distant homologs, according to a described procedure [10,11]. These database searches have revealed that several nematode proteins harbor related domains (cf. Fig. 1). For example, using the consensus sequence conN of Fig. 2, domains of the predicted *Caenorhabditis elegans* proteins C07G1.2, C16D9.1, C34G6.6, F38E11.4, F41A4.1 and T26C5.2, were detected with *E*() scores lower than 0.09. Iterative searches with consensus sequences constructed for the internal repeats of these nematode proteins (e.g. conc16d9_1, conc34g6_6, conf38e11_4, conf41a4_1, in Fig. 2) have identified several additional nematode proteins that contain up to five related domains linked to each other and/or to other types of modules (Fig. 1). Consensus sequences of the internal repeats of the nematode proteins identified by iteration are also shown in Fig. 2 (e.g. conc29e6_1, conc30h6.5, conc52b11_1, conf52b11_3, conh42k12_3, conr07a4_4). The percent identity of the internal repeats of nematode proteins is rather low (c16d9_1: 14%; c29e6_1: 8%; c30h6.5: 10–19%; c34g6_6: 8–28%; c52b11_1: 13%; f38e11_4: 12–19%; f41a4_1: 11–15%; f52b11_3: 9–18%; h42k12_3: 13%; r07a4_4: 14%). Despite this low sequence similarity, the homology of the internal repeats is supported by a characteristic pattern of cysteines and hydrophobic residues (cf. Fig. 2).

In the case of the predicted protein C07G1.2 that is composed of several types of modules (a C-type lectin module, a kringle module, two LDL-receptor type A modules, and an incomplete scavenger receptor module), the PAN module is found between two LDL-receptor type A modules (cf. Fig. 1). (It must be pointed out that many of the predicted *C. elegans* proteins found in current databases appear to be incomplete, lacking their amino-terminal and/or carboxy-terminal parts.) In the case of the predicted proteins C29E6.1, C29E6.4, C34G6.6, F38E11.4, F47G9.3 and F52B11.3, the tandem PAN domains are linked to a ZP domain, a domain identified previously in sperm receptors of zona pellucida and related proteins [16]. The PAN modules of *C. elegans* proteins share the pattern of conserved residues characteristic of the N domains of plasminogen-related genes, but in the majority of cases they contain an extra pair of conserved cysteines at the amino- and carboxyl-terminal boundaries of the domain (C1 and C6 in Fig. 2). It should be pointed out that the

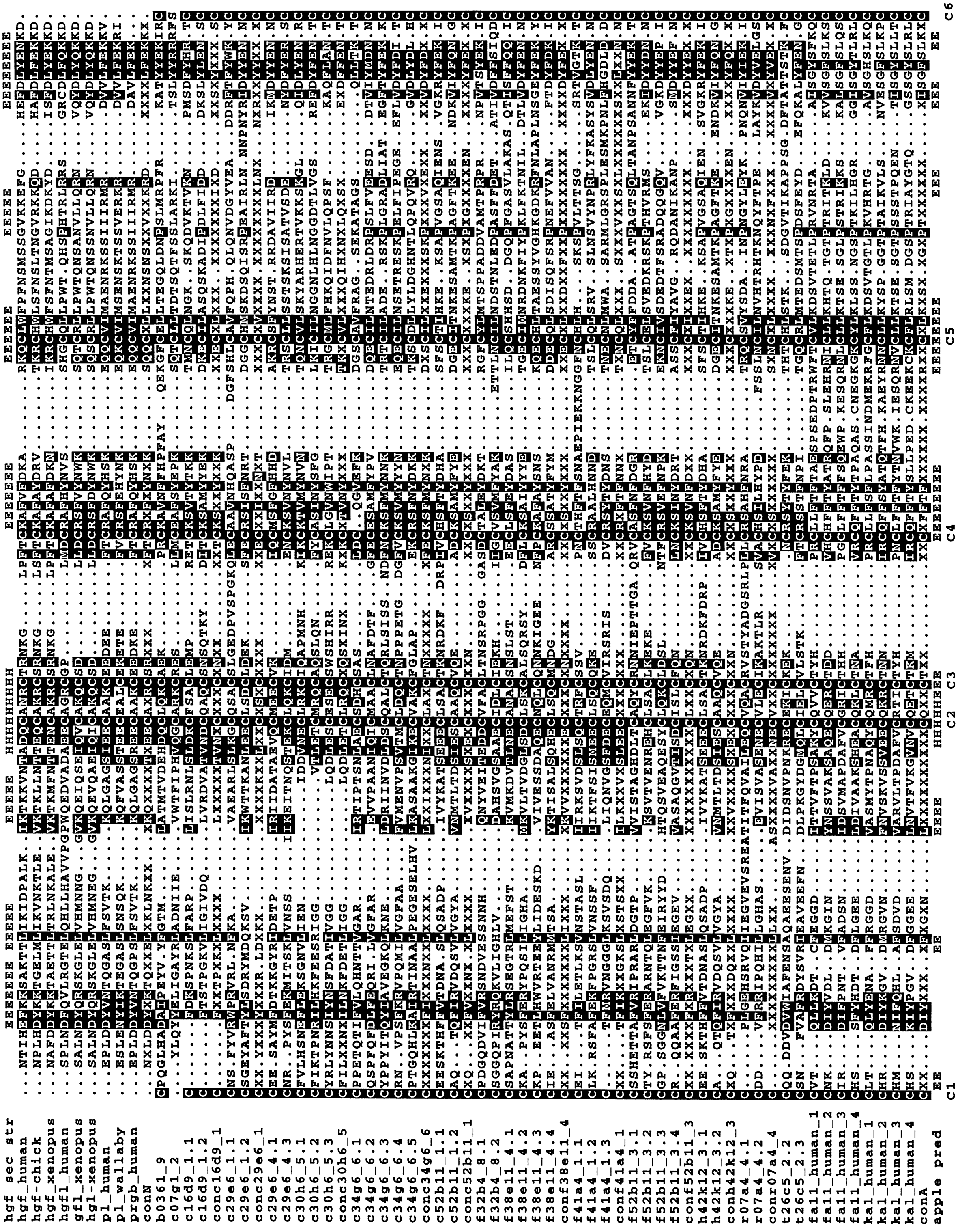

Fig. 2. Multiple alignment of the sequences of PAN modules of representative members of the plasminogen/hepatocyte growth factor family, various nematode proteins and the plasma prekallikrein/factor XI family. The sequences include the N-terminal domains of human (hgf_human, JH0579, residues 37–123), chick (hgf_chick, Q90978, residues 33–119) and *Xenopus* (hgf_xenopus, I51283, residues 24–110) HGFs, N-terminal domains of human (hgfl_human, P26927, residues 21–105) and *Xenopus* (gfl_xenopus, Q91691, residues 32–113; hgl_xenopus, P70006, residues 33–114) HGF-like proteins, the N-terminal domains of human (pl_human, P00747, residues 20–98) and wallaby (pl_wallaby, O18783, residues 20–98) plasminogens as well as the corresponding domain of human plasminogen-related gene B (prgb_human, Q02325, residues 20–96). The putative nematode proteins containing PAN modules are identified by their gene number, their tandem PAN modules are distinguished by a serial number: b0361_9, Q10952 (residues 49–134), c07g1_2, Q17797 (residues 351–429), c16d9_1, Q22902 (21–95, 114–194), c29e6_1 (LET-653), Q27394 (26–116, 123–209), c29e6_4, CAA96604 (88, 95–172, 179–251, 283–361, 382–461), c30h6.5, CAB02811 (81–161, 321–399, 420–498), c34g6_6, AAB52479 (25–97, 104–190, 112–300, 307–397, 404–491), c52b11_1, Q18777 (26–114, 123–203), f32b4_8, CAB04231 (residues 612–700, 709–794), f38e11_4, CAA92773 (27–113, 120–207, 212–304, 311–389), f41a4_1, AAC17538 (21–103, 108–192, 202–285), f52b11_3, CAB05199 (residues 21–112, 122–204, 211–295, 305–382), h42k12_3, O17347 (residues 26–114, 123–203), r07a4_4, CAA91764 (32–123, 133–220), t26c5_2, CAA90727 (24–101, 334–418, 429–514). The four apple domains of human factor XI (fa11_human, P03951, residues 20–103, 110–193, 200–283, 291–374), the four apple domains of human prekallikrein (kal_human, P03952, residues 21–104, 111–194, 201–284, 292–375) are distinguished by a serial number. ConN indicates a consensus sequence of N domains, conc16d9_1, conc29e6_1, conc30h6.5, conc34g6_6, conc52b11_1, conf38e11_4, conf41a4_1, conf52b11_3, conh42k12_3, conr07a4_4 represent consensus sequences for the tandem repeats of the respective proteins, conA represents the consensus sequence for the apple domains of factor XI and prekallikrein. To highlight conserved features of the PAN module, similar residues present in more than 50% of PAN modules are shaded. Amino acids were grouped according to their physicochemical properties into the following categories: hydrophobic residues (Val, Ile, Leu, Met, Tyr, Phe, Trp); hydroxylic amino acids (Ser, Thr); hydrophilic residues (Asp, Asn, Glu, Gln, Arg, Lys); Cys; Gly; Pro; Ala; His. The top row (hgf sec str) identifies the residues forming β-strands (E) and α-helix (H) in the known structure of the N domain of human HGF [2,3]. The secondary structural elements (H for α-helix, E for β-strand) predicted from the multiple alignment of the apple domains of prekallikrein with the PHD program are given at the bottom of the multiple alignment (apple pred). The last line identifies the positions of the six cysteines (C1, C2, C3, C4, C5, C6) conserved in all apple domains and the majority of the PAN modules of nematode proteins.

←

conserved residues shared by PAN modules of nematode proteins, PRGs and HGFs coincide with the secondary structural elements of the known structure of HGF's PAN module (cf. Fig. 2). Also note that gaps in the alignments of the PAN modules of nematode proteins coincide with gaps in the alignment of the sequences of N domains of plasminogen-related proteins. Since gap regions in multiple alignments identify surface loops connecting secondary structural elements, these observations indicate that the overall topology of the PAN modules of nematode proteins is similar to that of the N domain of HGF.

Iterative searches with the newly identified members of this module family have also revealed that it shows significant sequence similarity with the apple domains of the prekallikrein/factor XI family. For example, using the consensus sequence for the three tandem domains of protein F41A4.1 (conf41A41 in Fig. 2), the *E*() scores with the N domains of the majority of plasminogen-related proteins were in the range of 0.0032–0.011, whereas the *E*() scores for the apple domains of prekallikrein and factor XI were in the range of 0.028–0.19. The apple domains share the pattern of conserved residues characteristic of the PAN modules of nematode proteins, including all six conserved cysteines (Fig. 2). In the case of the multiple alignment of apple domains the segment connecting C4 and C5 contains gaps, suggesting that this region is a surface loop. Consistent with the surface location of this loop, it is known that the extra cysteine present in this region of the fourth apple domain of factor XI (cf. Fig. 2) forms a disulphide bond connecting two molecules of factor XI [17]. It is noteworthy that, with the exception of the first repeat of the nematode protein F41A4.1 (f41a4_1.1 in Fig. 2), this loop is significantly shorter in the other PAN modules. Aside from this unique feature of the apple domain, the homology suggests that the core structure of apple domains is similar to that of the N domain of HGF.

Coagulation factor XI and plasma prekallikrein are modular proteins that also belong to the trypsin family of serine proteinases (Fig. 1). Their non-protease parts consist of four tandem domains characterized by three conserved disulphide bonds, which give them an apple-like appearance [17,18]. Studies on the genes of factor XI [19], plasma prekallikrein [20] revealed that at both boundaries of all apple domains phase 1 introns are found. Thus the apple domain also satisfies the criteria for a mobile class 1-1 module. Our finding that the apple domains show significant sequence similarity with the N domain of plasminogen-related genes thus means that they form a single class 1-1 module family, and that this PAN domain does indeed correspond to a mobile module.

The significance of the sequence similarities detected in the present work is supported by the fact that the six cysteines conserved in the majority of nematode proteins (C1, C2, C3, C4, C5 and C6 in Fig. 2) align with the six cysteines conserved in all apple domains. Furthermore, four of these conserved cysteines (C2, C3, C4, C5) are also present in the N domains of plasminogen-related proteins as well as in the PAN module of protein C07G1.2. Importantly, the four cysteines shared by all PAN modules form disulphide bonds in the same pattern in apple domains of prekallikrein and factor XI [17,18] and in plasminogen [1]: C2–C5, C3–C4. On the basis of homology we may thus predict that the six conserved cysteines of nematode proteins form disulphide bonds in a pattern similar to that in apple domains.

To further test the significance of this sequence similarity (and disulphide bond pattern similarity), we have predicted the secondary structure of the apple domains of plasma prekallikrein and coagulation factor XI according to a described method [13–15]. As shown in the bottom line of Fig. 2 (apple pred), the positions and types of the predicted secondary structure elements of the apple domains are in good agreement with those in the known structure of the N domain of HGF (HGF sec str, top line of Fig. 2). On the basis of homology we may thus predict that the apple modules of prekallikrein/factor XI and the PAN modules of nematode proteins have a fold similar to that of HGF's N domain and their disulphide bond pattern is also similar.

In principle, the homology of the PAN modules of the nematode proteins with the apple domains of prekallikrein/factor XI and with the N domains of the plasminogen family could shed some light on the function and structure-function relationships of these molecules. The N domains of some plasminogen-related proteins are functionally well characterized. For example, the N domain of HGF is known to be important for the potent biological activity of this growth factor since it is involved in its binding to the c-Met receptor. Deletion of this region abolishes both the c-Met binding and heparin binding abilities of HGF [21–24]. In the case of plasminogen, the N-terminal domain is involved in intramolecular interactions controlling the conformation, activation and fibrin affinity of plasminogen [1,25]. The apple domains of plasma prekallikrein are known to mediate its binding to high molecular weight kininogen [26], the apple domains of factor XI carry binding sites for factor XIIa, platelets, kininogen, factor IX and heparin [27]. In view of such functional versatility of the apple domains and N domains, we can say very little about the function of the PAN modules of the nematode proteins, aside from suggesting that they may also be involved in binding interactions.

At present, very little is known about the biological function of the *C. elegans* proteins containing PAN modules. The only exception is protein C29E6.1 encoded by the let-653 gene. This protein (in addition to the PAN modules) contains a ZP module, a module identified previously in sperm receptors, TGF-β type III receptors, uromodulin, as well as GP2, the major glycoprotein of pancreatic secretory granules [16]. Mutations in the let-653 gene are early larval lethal, the lethal arrest is concurrent with the appearance of a vacuole anterior to the lower pharyngeal bulb. Mutants are characterized by cystic excretory canals, consistent with a dysfunction of the secretory/excretory apparatus [28]. It is noteworthy that GP2 is thought to play a crucial role in pancreatic secretion [29] and in this respect it shows some functional similarity to C29E6.1. Since the predicted proteins C29E6.4, C34G6.6, F38E11.4, F47G9.3 and F52B11.3 have domain organizations similar to that of C29E6.1 (cf. Fig. 1), it is possible that they fulfill similar functions in the secretory apparatus.

Acknowledgements: This work was supported by Grants OTKA T022949 and OTKA T014642.

References

[1] Wiman, B. (1973) Eur. J. Biochem. 39, 1–9.
[2] Zhou, H., Mazzulla, M.J., Kaufman, J.D., Stahl, S.J., Wingfield, P.T., Rubin, J.S., Bottaro, D.P. and Byrd, R.A. (1998) Structure 6, 109–116.
[3] Ultsch, M., Lokker, N.A., Godowski, P.J. and de Vos, A.M. (1998) Structure 6, 1383–1393.
[4] Ichinose, A. (1992) Biochemistry 31, 3113–3118.
[5] Tateno, T. and Ichinose, A. (1999) FEBS Lett. 445, 31–35.
[6] Lewis, V.O., Gehrmann, M., Weissbach, L., Hyman, J.E., Rielly, A., Jones, D.G., Llinas, M. and Schaller, J. (1999) Eur. J. Biochem. 259, 618–625.
[7] Petersen, T.E., Martzen, M.R., Ichinose, A. and Davie, E.W. (1990) J. Biol. Chem. 265, 6104–6111.
[8] Miyazawa, K., Kitamura, A. and Kitamura, N. (1991) Biochemistry 30, 9170–9176.
[9] Han, S., Stuart, L.A. and Friezner Degen, S.J. (1991) Biochemistry 30, 9768–9780.
[10] Patthy, L. (1987) J. Mol. Biol. 198, 567–577.
[11] Patthy, L. (1996) Methods Enzymol. 266, 184–198.
[12] Thompson, J.D., Higgins, D.G. and Gibson, T. (1994) Nucleic Acids Res. 12, 4673–4680.
[13] Rost, B. and Sander, C. (1994) Proteins 19, 55–77.
[14] Rost, B. (1995) in: The Third International Conference on Intelligent Systems for Molecular Biology (Rawlings, C., Clark, B., Altman, R., Hunter, L., Lengauer, T. and Wodak, S., Eds.), pp. 314–321, AAAI Press, Menlo Park, CA.
[15] Rost, B. (1996) Methods Enzymol. 266, 525–539.
[16] Bork, P. and Sander, C. (1992) FEBS Lett. 300, 237–240.
[17] McMullen, B.A., Fujikawa, K. and Davie, E.W. (1991) Biochemistry 30, 2056–2060.
[18] McMullen, B.A., Fujikawa, K. and Davie, E.W. (1991) Biochemistry 30, 2050–2056.
[19] Asakai, R., Davie, E.W. and Chung, D.W. (1987) Biochemistry 26, 7221–7228.
[20] Beaubien, G., Rosinski-Chupin, I., Mattei, M.G., Mbikay, M., Chretien, M. and Seidah, N.G. (1991) Biochemistry 30, 1628–1635.
[21] Matsumoto, K., Takehara, T., Iboue, H., Hagiya, M., Shimizu, S. and Nakamura, T. (1991) Biochem. Biophys. Res. Commun. 181, 691–699.
[22] Okogaki, M., Komada, M., Uehara, Y., Miyazawa, K. and Kitamura, N. (1992) Biochemistry 31, 9555–9561.
[23] Lokker, N.A., Presta, L.G. and Godowski, P.J. (1994) Protein Eng. 7, 895–903.
[24] Sakata, H. (1997) J. Biol. Chem 272, 9457–9463.
[25] Bányai, L. and Patthy, L. (1984) J. Biol. Chem. 259, 6466–6471.
[26] Herwald, H., Renne, T., Meijers, J.C.M., Chung, D.W., Page, J.D., Colman, R.W. and Muller-Esterl, W. (1996) J. Biol. Chem. 271, 13061–13067.
[27] Ho, D.H., Badellino, K., Baglia, F.A. and Walsh, P.N. (1998) J. Biol. Chem. 273, 16382–16390.
[28] Jones, S.J. and Baillie, D.L. (1995) Mol. Gen. Genet. 248, 719–726.
[29] Wong, S.M.E. and Lowe, A.W. (1996) Gene 171, 311–312.