The solution structure of the MANEC-type domain from hepatocyte growth factor activator inhibitor-1 reveals an unexpected PAN/apple domain-type fold

Zebin Hong*, Michal Nowakowski†, Chris Spronk‡, Steen V. Petersen§, Peter A. Andreasen*, Wiktor Koźmiński†, Frans A.A. Mulder¶ and Jan K. Jensen*†

*Department of Molecular Biology and Genetics, Danish-Chinese Centre for Proteases and Cancer, Aarhus University, 8000 Aarhus C, Denmark
†Faculty of Chemistry, Biological and Chemical Research Centre, University of Warsaw, 00-927 Warsaw, Poland
‡Spronk NMR, LT-01104, Lithuania
§Department of Biomedicine, Aarhus University, 8000 Aarhus C, Denmark
¶Interdisciplinary Nanoscience Center (iNANO) and Department of Chemistry, Aarhus University, 8000 Aarhus C, Denmark

INTRODUCTION

Motif at N-terminus with eight-cysteines (MANEC) was discovered in 2004 as a new class of protein domains with unknown function and structure. Initially, the domain was believed to contain seven cysteines, hence the early name MANSC [1]. However, subsequent sequence analysis added the eighth cysteine to the conserved pattern (BLAST and SMART, http://smart.embl-heidelberg.de/) [2]. The domain encompasses 80–100 amino acid residues and is found in more than 420 proteins exclusively from animals. Apart from the uncharacterized protein UPI0002B46A4F from Hydra vulgaris predicted to contain 11 MANEC domains, MANEC domains occur as a single domain motif in the primo part of type-I transmembrane multi-domain proteins.

Only one MANEC-containing protein, hepatocyte growth factor activator inhibitor-1 (HAI-1), has been functionally characterized. The functions of two other MANEC-containing proteins, LRP-11 and KIAA0319, were predicted from sequence homology (see Supplementary Figure S1). The three proteins are all among the five MANEC-containing proteins in humans (Supplementary Figure S2A). HAI-1 was originally identified as an inhibitor of hepatocyte growth factor activator (HGFA) [3,4]. The inhibitory repertoire of HAI-1 has subsequently expanded to include the type-II transmembrane serine proteinases matriptase [4,5] and hepsin [6,7] as well as the membrane-associated serine protease prostasin [6]. The observation that HAI-1 deficient mice die in utero due to placental defects further supported the importance of its protease inhibitory activity [8]. A mild overexpression of matriptase relative to HAI-1 results in ∼70% of the mice developing spontaneous squamous cell carcinoma. The oncogenic effects of matriptase were removed by additional expression of HAI-1 [9]. Moreover, HAI-1 appears to be an important regulator of a delicate balance between homeostasis and the development of cancer. As a result, low HAI-1 levels have been proposed as a prognostic marker of a poor prognosis in several cancers [10,11]. The mature HAI-1 protein has a molecular mass of ∼53 kDa and encompasses a large uncharacterized N-terminal region containing the MANEC domain followed by a predicted ‘internal’ domain, a Kunitz-type inhibitor domain, a LDLR (low-density lipoprotein receptor) class A domain, a second Kunitz-type inhibitor domain, a transmembrane region, and a short C-terminal cytoplasmic tail (Figure 1 and Supplementary Figure S1). HAI-1 exists in two splice variants, the full-length isoform 1 [3] and isoform 2 [12] with a 16 amino acid deletion (306–321).

 fail to predict structural homology. In the present paper, we present the NMR solution structure of the MANEC domain from HAI-1, the first three-dimensional (3D) structure from the MANEC domain family. Unexpectedly, MANEC is a new subclass of the PAN/apple domain family, with its own unifying features, such as two additional disulfide bonds, two extended loop regions and additional α-helical elements. As shown for other PAN/apple domain-containing proteins, we propose a similar active role of the MANEC domain in intramolecular and intermolecular interactions. The structure provides a tool for the further elucidation of HAI-1 function as well as a reference for the study of other MANEC-containing proteins.

Key words: NMR, serine protease inhibitor, structure homology.
Figure 1  HAI-1 full length protein sequence with domain locations

Signal sequence, domain architecture, and transmembrane region are shown as labelled boxes below the corresponding amino acid residues (isoform 1, Uniprot #O43278-1). The location of signal sequence, Kunitz-type inhibitor and LDLR class A domains and the transmembrane region is as previously reported in Shimomura et al. (1997) [3]. The location of MANEC sequence corresponds to what was determined in the present study. The location of the internal domain is our prediction based on Kojima et al. (2008) [14]. The positions of the three potential N-glycosylation sites are indicated with an asterisk (*).

between the first Kunitz-type inhibitor domain and the LDLR class A domain. Three potential N-glycosylation sites are found at positions 66, 235 and 523 (507 in isoform 2).

The protease inhibitory activity resides strictly in the first Kunitz-type inhibitor domain. The N-terminal region encompassing the MANEC domain and the internal domain has been proposed to play a role in the regulation of the inhibitory activity of HAI-1. Moreover, the MANEC domain has been suggested to participate in both intramolecular interactions with other domains of HAI-1 and intermolecular interactions with the target protease [13,14].

Characterization of the MANEC domain remains a challenge for a comprehensive biochemical understanding of MANEC-containing proteins. A keystone for elucidating protein function is the availability of structural information. Unfortunately, sequence-based homology searches failed to predict the MANEC fold. We have generated the MANEC domain from human HAI-1 and show that the expressed protein domain is soluble and stable. Using state-of-the-art four-dimensional NMR technology we solved the three-dimensional solution structure. The structure revealed a well-defined fold with a four disulfide bond pattern. Unexpectedly, structure-based homology searches revealed a close homology to the PAN/apple domain family but also distinct differences from proteins of this family. Based on our data, we identify MANEC domains as a new subclass of the PAN/apple domain family. The homology to the PAN/apple domains suggests a similar role of the MANEC domain as a mediator of molecular interactions with potential regulatory properties for the activity of the parent protein.

EXPERIMENTAL

Protein production

The DNA sequence encoding the MANEC domain (G47-L152) was amplified from full-length human HAI-1 cDNA (isoform 1, Uniprot #O43278-1). The N66Q variant was generated by site-directed mutagenesis. The PCR product was subcloned into the XhoI–SalI sites of the pPICZaA expression vector (Invitrogen) and linearized by digestion with the restriction enzyme SacI prior to transformation into Pichia pastoris X-33 strain (Invitrogen). Protein producing clones were stored at −80°C in 50% glycerol.

Protein production followed the manufacturer’s recommendations (Invitrogen).

Protein purification

The medium was cleared from yeast cells by centrifugation and concentrated >20-fold using a stirring cell concentrator (3.5 kDa cut-off). The concentrate was loaded on to a Ni-NTA (nitrotriacetate) column equilibrated with 20 mM NaH2PO4, pH 6.5, 100 mM NaCl, washed with buffer supplemented with 20 mM imidazole and finally bound protein was eluted by increasing the imidazole concentration to 300 mM. The sample was dialysed against 20 mM NaH2PO4, pH 6.5, 100 mM NaCl. The identity and integrity of the purified protein was verified by Western blotting and mass spectrometry (MS). Protein concentration was measured by absorbance or attenuation at 280 nm using a calculated molar absorption
Uniform $^{13}$C and $^{15}$N protein labelling

MANEC producing yeast was inoculated in 5 ml of YPD (yeast extract-peptone-dextrose) medium and grown for 1 day. The yeast cells (250 μl) were pelleted by centrifugation at 5000 g, resuspended in 5 ml of growth medium containing 100 mM KH$_2$PO$_4$, pH 6.0, 0.34 % yeast nitrogen base (YNB) [lacking (NH$_4$)$_2$SO$_4$], 0.5 % (w/v) ($^{13}$NH$_4$)SO$_4$ and 2 % (w/v) $^{13}$C-glucose (99 %, Cambridge Isotope Laboratories) buffer medium and grown for an additional day. The following day, an additional 100 ml of labelled growth medium was added and cells allowed to grow 1 day more. The yeast cells were harvested by centrifugation at 5000 g, resuspended in 500 ml of induction-medium containing a similar composition to growth medium with methanol instead of glucose as a carbon source and supplemented daily with 1 % (v/v) $^{13}$C-methanol for 4 days. Suspension cultures were grown at 28 °C with orbital shaking at 300 rev./min in ruffled conical flasks. Labelling efficiency was determined to >99 % by MS (Supplementary Figure S3A). Protein labelled uniformly with $^{15}$N only was obtained by a similar procedure (mass spectrum not shown).

Mass spectrometry

Approximately 2 μg of purified intact MANEC protein was acidified and desalted using Poros50 R1 micro columns essentially as described by Gobom et al. [15]. Bound protein was eluted using 90 % acetonitrile, 0.1 % TFA (trifluoroacetic acid), lyophilized, and resuspended in 2 μl of 1 % TFA and mixed with an equal volume of DHAP (2,5-dihydroxycetophenone) matrix solution prepared in 20 mM diammonium hydrogen citrate, 75 % (v/v) ethanol. After thorough mixing, samples were spotted on to a matrix-assisted laser desorption-ionization (MALDI) target and mass spectra acquired using a Bruker Autoflex III instrument operated in linear mode and calibrated in the mass range of 5000–17500 Da using Protein calibration standard I (Bruker Daltonics). The degree of $^{13}$C and/or $^{15}$N labelling was estimated on the basis of mass increase relative to the unlabelled material. For fragmentation studies, purified MANEC (3 μg) was digested for 1 h at 37 °C using porcine trypsin (Promega) in 50 mM ammonium bicarbonate containing 5 mM iodoacetamide to block any free cysteines. Approximately 20 pmol protein was removed and desalted using a StageTip (C8, Thermo Scientific). Bound peptides were eluted directly on to the MALDI target plate using α-cyano-4-hydroxy-cinnamic acid in 70 % acetonitrile, 0.1 % TFA. Peptides were subsequently analysed by using an Autoflex Smartbeam III instrument (Bruker) operated in positive and linear mode. Prior to analysis, the instrument was calibrated by external calibration using a peptide mixture containing seven calibrants (Bruker). The obtained data were evaluated by using the GPMAW software (gpmaw.com).

CD spectroscopy

CD spectroscopy was performed on a JASCO-810 circular dichroism (CD) system with Peltier temperature control. Wavelength scans were averages of five scans between 190 nm and 350 nm collected at 25 °C in a quartz cuvette with a 1-mm path length with 5 μM protein in 10 mM NaH$_2$PO$_4$, pH 7.4. Secondary structure analysis was performed using an online service (http://dichroweb.cryst.bbk.ac.uk/html/home.shtml). For thermal denaturation experiments, 10 μM samples were heated from 25 °C to 95 °C at 1 °C/min with or without 10 mM DTT (dithiothreitol). All CD data were buffer subtracted. $T_m$ values were determined by sigmoidal curve fit between the low (native) and high (unfolded) temperature plateaus, as averages of at least three experiments.

NMR sequence-specific resonance assignment

NMR measurements used for backbone resonance assignment and collection of distance restraints were acquired at 25 °C on an Agilent DDR2 800 MHz spectrometer equipped with a cryogenic probe-head and an Agilent DDR2 600 MHz spectrometer equipped with a ‘PENTA’ ($^1$H,$^1$C,$^{15}$N,$^3$P,$^1$H) probe-head.

Both 3D and 4D experiments were performed using sparse random sampling of indirectly detected dimensions to increase resolution (relevant parameters of all NMR spectra measured with NUS (non-uniform sampling) are given in Supplementary Table S4). Chemical shifts in $^1$H NMR spectra were reported with respect to external deuterated 4,4-dimethyl-4-silapentane-1-sulfonic acid. Chemical shifts of $^{13}$C and $^{15}$N signals were referenced indirectly using the 0.251449530 and 0.101329118 frequency ratios for $^{13}$C/H and $^{15}$N/H, respectively [18]. The 2D and conventionally-sampled 3D experiments were processed with NMRpipe [16], 3D and 4D NUS experiments were processed by SSA software package available at http://nmr.cent3.uw.edu.pl/software [19–21]. Processed spectra were analysed with SPARKY [17]. Aliphatic side chain assignment was achieved using $^1$H–$^{13}$C H$\alpha$QC, 4D $^1$H$_\alpha$$^1$C$_\alpha$($^15$N,$^1$H) NOESY [22,23] and 4D HCCH-TOCSY [21,24,25]. The aromatic side chain resonances were assigned from the analysis of $^1$H–$^{13}$C HSQC, 4D $^1$H$_\alpha$$^1$C$_\alpha$($^15$N,$^1$H) NOESY [22,23] and 4D HCCH-TOCSY [21,24,25]. The aromatic side chain resonances were assigned from the analysis of $^1$H–$^{13}$C HSQC tuned to aromatic carbons, 2D (HB)CB(CGCD)H, 2D (HB)CB(CGCDCE)H, 3D HBCB(CGCD)H, 3D HBCB(CGCDCE)H [26] acquired at 800 MHz, 3D $^{13}$C-edited NOESY HSQC tuned to aromatic carbons (measured at 600 MHz) [27] and 800 MHz 4D $^{13}$C,$^{13}$C$\alpha$-edited HMQC–NOESY–HSQC [28]. Distance constraints were obtained at 800 MHz from the $^{15}$N-edited NOESY–HSQC [29], 4D $^{13}$C$\alpha$,$^{13}$C$\alpha$-edited HMQC–NOESY–HSQC, 4D $^{13}$C$_\alpha$,$^{13}$C$_\alpha$-edited HMQC–NOESY–HMQC [30,31], and 4D $^{15}$N,$^{15}$N-edited HMQC–NOESY–HSQC [21].

NMR structure calculation procedure

The NMR structures were calculated using the CYANA 3.96 automated NOE assignment and structure-calculation protocol [32]. Input for the structure calculations were peak lists from the following five NOESY spectra: 4D $^{13}$C-edited aliphatic–aliphatic NOESY (2479 peaks); 4D $^{13}$C-edited aliphatic – $^{15}$N-edited NOESY (811 peaks); 4D $^{13}$C-edited aromatic–aliphatic NOESY (190 peaks); 3D $^{15}$N-edited NOESY (1662 peaks); 3D $^{13}$C$\alpha$-edited aromatic NOESY (59 peaks).

Tolerances for chemical shift assignments were set to 0.02 ppm for all $^1$H dimensions, and 0.4 for $^{15}$N and $^{13}$C dimensions.
in all spectra. In each structure calculation cycle, 80 structures were calculated of which 20 were selected for CYANA analysis. Ambiguous assignments were kept in the final cycle. In order to cross-validate the final structures, no dihedral-angle restraints from TALOS+ (http://spin.niddk.nih.gov/bax/nmrserver/talos/) [33] were included in the structure calculations.

**NMR structure refinement and validation**

Structure refinement was done using a simulated-annealing protocol with explicit solvent in the software YASARA Structure [34]. The protocol employed the YASARA2 force field with potentials for Van der Waals and electrostatic interactions to improve non-bonded interactions, and knowledge-based potentials to define rotamer states. For structure refinement, the 20 best (lowest energy) conformers were used, together with the upper-distance restraint limits of the last CYANA cycle. The upper-distance restraint limits were exported to XPLOR format for use in YASARA. Structural water molecules (g=0) and violations of distance restraints above 0.5 Å (1 Å = 0.1 nm) were found. Local and global geometry of the final ensemble of 20 refined structures were checked using the WHATIF web server (http://swift.cmbi.ru.nl/whatif/). In addition, two independent measures of validation using experimental data were applied to the structures: (1) determination of the disulfide bonding partners by MS and (2) comparison of TALOS+ chemical shift-based secondary-structure predictions with secondary-structure elements in the final refined structure ensemble.

**SAXS data acquisition and analysis**

The SAXS data were collected on the laboratory-based instrument at iNANO, Aarhus University, Denmark as described in [35]. Samples were placed in re-usable quartz capillaries at a controlled temperature of 25°C. The acquisition time was 3600 s for a MANEC protein sample of 1.0 mg/ml in 20 mM NaH2PO4, pH 7.4, 100 mM NaCl. Corresponding buffer data were collected. Background buffer subtraction and conversion of the data to absolute scale by use of water as a primary standard was performed using the SUPERSAXS program package (Jan S. Pedersen and Cristiano L. Pinto Oliveira, Aarhus University, unpublished). The instrumental sample-to-detector distance was set to 640 cm, giving a q range of 0.01–0.345 Å−1, where q is the length of the scattering vector, defined as $q = 4\pi \sin(\theta)/\lambda$, where $\lambda$ is the X-ray wavelength at 1.54 Å and $2\theta$ is the scattering angle between the incident and scattered beam. The final intensity is $I(q)$, in units of cm−1. Indirect Fourier transformation analysis was performed using the program WIFT ([36] and Jan S. Pedersen and Cristiano L. Pinto Oliveira, Aarhus University, unpublished) to obtain the pair distance distribution function, $p(r)$, along with the characteristic parameters: the maximum diameter, $D_{\text{max}}$, the radius of gyration, $R_g$ and the forward scattering, $I(q=0)$. The theoretical scattering profile of the NMR structure was computed and compared to the experimental data using the program CRYSON [37]. To obtain a structural model of the scattering molecule in solution without the use of any assumptions other than $D_{\text{max}}$, ab initio modelling was performed using the program DAMMIF [38]. Finally, our NMR structure ensemble was directly compared to the ab initio SAXS model using part of the DAMAVER program package, specifically the programs SUPCOMB [39] and DAMSEL [40].

**ACCESSION NUMBERS**

The atomic coordinates, resonance assignment lists and distance restraints lists (PDB ID code: 2msx) have been deposited in the Protein Data Bank, Research Collaboratory for Structural Bioinformatics, Rutgers University, New Brunswick, NJ, U.S.A. (http://www.rcsb.org/) and to BMRB (accession code: 25139), University of Wisconsin, WI, U.S.A. (http://www.bmrb.wisc.edu).

**RESULTS**

The MANEC domain from HAI-1 is expressed as a soluble entity

To allow for the expression of the MANEC domain, we aligned all available HAI-1 sequences to obtain a consensus sequence of the MANEC domain. This analysis identified G47-L152 as the MANEC domain of human HAI-1 (Supplementary Figure S2B). The N-terminal region of human HAI-1 encompassing M1-A35 is predicted to represent the signal peptide, producing mature HAI-1 containing a short segment of 11 amino acid residues preceding the MANEC domain. This segment P36-A46 was predicted to be unstructured due to the high content of Pro and Gly/Ala (I-TASSER [41]). Likewise, the segment succeeding L152 is also predicted to be unstructured (I-TASSER), and a good candidate for an interdomain-linker region between MANEC and the predicted internal domain (Figure 1). Hence, we decided to clone and express the 106 amino acid residue segment representing G47-L152. The MANEC domain was expressed in *P. pastoris* to facilitate correct folding including the establishment of four predicted disulfide bonds. An N66Q substitution was introduced to remove the predicted glycosylation site, thus producing a protein migrating similarly to deglycosylated wild-type MANEC on SDS/PAGE (Supplementary Figure 2A). Analysis by MS produced the expected mass of 12815 Da, showing the absence of any post-translational modifications (Supplementary Figure S3A). When subjected to size-exclusion chromatography, MANEC eluted as a single peak with an elution time corresponding well to the theoretical mass of the monomer (12.8 kDa) (Figure 2B) with no tendency for interdomain cysteine bond formation (Figure 2A, lane 5). The purified protein was soluble with no sign of precipitation at high concentrations (≤40 mg/ml). In summary, the purified HAI-1 MANEC domain appears as a well-behaved isolated protein suitable for structural studies.

The MANEC domain is thermodynamically stable and homogenously folded

The secondary structure content of MANEC was investigated by CD spectroscopy. From the shape of the wavelength scan spectrum, it was predicted that the protein contains both α-helical and β-sheet structures (Figure 2C). The thermal denaturation profile of MANEC evaluated by CD spectroscopy revealed a minimal loss of secondary structure up to ~80°C (Figure 2D). When the experiment was performed under reducing conditions, a single melting point transition was observed with a $T_m$ of 64.0 ± 0.1 °C (Figure 2D). This is a surprisingly high thermal stability for a small protein in the absence of disulfide bonds. Finally, a two-dimensional [1H–15N] NMR HSQC spectrum was recorded on a uniformly 15N-labelled sample. In the well-dispersed spectrum, each amino acid was represented by a unique peak as expected for a homogeneous sample of a folded protein (Figure 3). In concert, the CD and NMR data show that purified MANEC domain is a stable protein with a defined fold.

NMR reveals a well-defined MANEC structure stabilized by four disulfide bonds

Uniformly 13C- and 15N-labelled MANEC was prepared (Figure S3A) and state-of-the-art 3D and 4D liquid-state NMR
Figure 2  Biochemical and biophysical characterization of the purified MANEC

(A) SDS/PAGE analysis (15%) of MANEC purified from Pichia pastoris. Lanes 1–5 contain: 3 μg wild-type MANEC (lane 1), wild-type MANEC incubated with PNGaseF to remove N-linked glycans (lane 2), N66Q MANEC (lane 3), PNGaseF (lane 4) and N66Q MANEC under nonreducing conditions (lane 5). The samples in all lanes but lane 5 were incubated with 1% 2-mercaptoethanol to disrupt disulphide bridges. Prestained protein marker (Fermentas) is shown between lanes 4 and 5 with corresponding masses shown on the right. (B) Elution profile of MANEC protein established by size-exclusion chromatography by using a Superdex75 column (GE Healthcare). MANEC eluted as a single peak at 13.8 ml. The void volume and elution volumes of protein standards are indicated above the profile. (C) A representative CD wavelength scan spectrum for a 5 μM MANEC sample recorded at 25 °C. The ellipticity (mdeg) is given as a function of the wavelength of the circular polarized light (nm). (D) Thermal denaturation curve for MANEC protein by CD spectroscopy. The change in CD signal as a function of increasing temperature was monitored at 220 nm. A representative thermal scan performed in standard phosphate buffer is shown with grey circles, and supplemented with 10 mM DTT (dithiothreitol) to reduce disulfide bonds is shown with black squares. The corresponding fit to a sigmoidal curve to obtain the Tm value for the DTT (dithiothreitol) experiment is shown as a solid grey line.

Experiments were used to obtain 88% complete chemical shift assignments and NOESY peak lists. The assignments and the NOESY data from 3D and the highly unambiguous and information-rich 4D experiments were used as input for automated NOE assignment and structure calculation in CYANA. An example of NOESY spectra quality and NOE assignment is given in Supplementary Figure S4. The carbon-13 chemical shift-values of all assigned cysteines was in agreement with an oxidized state [42]. Initial structure calculations immediately revealed three of the four possible disulfide bridges: C87–C116, C91–C97, and C121–C129, with a consistent maximum distance of 4 Å between the β-carbon atoms within each cysteine pair. This emphasized the quality of the NMR data and was a strong indication of the validity of the calculated structure. Additional restraints for the four disulfide bonds were included in the subsequent CYANA calculations. The assignment of ~40 distance-restraints per residue (average) (Figure S5A) resulted in a well-defined structure with very low ensemble-averages of backbone and heavy atom with RMSD values of 0.47 ± 0.13 Å and 0.97 ± 0.12 Å, respectively (Supplementary Table S1). High conformational variability is only observed for the three N-terminal residues and the C-terminal hexa-His-tag (Figure 4) for which only few or no NOE restraints were observed (Supplementary Figure S5A). To further validate the convergence of the structure calculations, five additional independent CYANA calculations were performed using different random seeds. All resulting 120 structures agreed very well with average backbone and heavy atom RMSD values of 0.68 ± 0.16 Å and 1.20 ± 0.16 Å, respectively. Overall, the sequence-predicted secondary structure elements by I-TASSER [41] as well as those predicted by the TALOS analysis of the backbone chemical shifts (Supplementary Figure S5B) agrees well with the calculated structure. However, in two cases, TALOS + predicted α-helical (96–97) and β-strand (119–123) secondary structure for an unstructured loop region in the structure. In both cases, the presence of disulfide bridges affects the local structure, and affects chemical shifts in the vicinity of the cysteine residues (Cys96 and Cys121). These segments may therefore not adopt canonical secondary structure, or may no longer be identified as such on the basis of chemical shifts alone.
The NMR-derived disulfide bond pattern was verified by mass spectrometry

To verify the connectivity of the disulfide bonds in the NMR structure, a MANEC sample was digested with trypsin and analysed by MALDI-MS (Supplementary Figure S3B). The ion of \( m/z \) 3194.0 corresponds to the disulfide-linked double peptide encompassing G84–R89 and G109–R130, corroborating the disulfide bridge between C87 and C116 identified by NMR analysis (Table 1). The absence of any S-carbamidomethylated cysteine residues confirmed the intrapeptide bond between C121 and C129 within the G109–R130 peptide. A second ion of \( m/z \) 3350.2 represented the same disulfide-linked double peptide encompassing the peptides R83–R89 and G109–R130 (Table 1). The remaining two disulfide bonds were represented by \( m/z \) 5713.1 corresponding to the disulfide linked G47–R82 and A90–R108 peptides with no S-carbamidomethylated cysteine residues (Table 1). The presence of this ion validated the NMR data, establishing the C50–C92 and C91–C97 disulfide bridges. In addition to ions representing disulfide linked peptides, a triplet of ions of \( m/z \) 3594.6, 3627.6 and 3660.1 was observed. This cluster represented the MALDI-induced fragmentation of the C50–C91 intermolecular disulfide bond present in the \( m/z \) 5713.1 ion producing the G47–R82 peptide without sulfur (\( m/z \) 3594.6), with sulfur (\( m/z \) 3627.6) or with an additional sulfur atom originating from C91 (\( m/z \) 3660.1) [43]. In summary, the data obtained by MALDI-MS analysis correlates with the findings by NMR analysis establishing disulfide bonds between C50–C92, C87–C116, C91–C97 and C121–C129.

MANEC domains represent a new subclass of the PAN/apple domain family

Next, we wanted to investigate whether our new structure represented a novel fold as suggested by the sequence-based homology searches. A three-dimensional structure-based search using the DALI server (http://ekhidna.biocenter.helsinki.fi/dali_server/start) [44] identified several proteins with homologous structures including proteins from the plasminogen subfamily of the serine proteases, plasminogen, hepatocyte growth factor (HGF) and coagulation factor XI (Supplementary Table S2). In all cases, the MANEC structure matched the fold of a PAN or apple domain from the respective proteins. The PAN domain is a divergent subclass of the apple domain family defined by a single central five-stranded \( \beta \)-sheet, a short \( \alpha \)-helix connecting strands 5 and 3 (equivalent to MANEC \( \alpha \)-helix 2), and a short \( \alpha \)-helix in the C-terminus (equivalent to MANEC \( \alpha \)-helix 3) [45]. The domain is stabilized by the presence of two conserved disulphide bonds interlocking the fold (Figures 5B and 5D). The apple domain differs from the PAN domain by encompassing an additional conserved disulfide bond bridging the N- and the C-terminus to form a structure with an ‘apple-like’ appearance (Figures 5C and 5D). Another characteristic feature of the common PAN/apple fold is the way \( \beta \)-strands 1 and 2 bend and adopt a conformation in which strand 2 ‘wraps around’ strand 1 (Figures 5B and 5C). By comparing our structure to the HGF PAN domain with the best fit (Supplementary Table S2), it was possible to identify several additional features of the HAI-1 MANEC domain not found in the common PAN/apple domain fold (Figures 5B–5D). The most obvious difference is...
Table 1  MANEC disulfide bonds analysis by MS

The assignment of [M + H]+ ions observed in the range m/z 3000–6000 corresponding to disulfide-linked fragments following tryptic digest. The observed and theoretical calculated molecular mass of the indicated peptide sequences are shown in Da.

<table>
<thead>
<tr>
<th>Observed mass (Da)a</th>
<th>Calculated mass (Da)b</th>
<th>Disulfide-linked peptides</th>
</tr>
</thead>
<tbody>
<tr>
<td>3193.0</td>
<td>3193.4</td>
<td>Gly84–Arg89 (Cys87)</td>
</tr>
<tr>
<td>733.3</td>
<td>733.3</td>
<td></td>
</tr>
<tr>
<td>2460.1</td>
<td>2460.1</td>
<td>Gly109–Arg130 (Cys116, Cys121, Cys129)</td>
</tr>
<tr>
<td>3349.2</td>
<td>3349.5</td>
<td>Gly83–Arg89 (Cys87)</td>
</tr>
<tr>
<td>889.4</td>
<td>889.4</td>
<td></td>
</tr>
<tr>
<td>2460.1</td>
<td>2460.1</td>
<td>Gly109–Arg130 (Cys116, Cys121, Cys129)</td>
</tr>
<tr>
<td>3593.6</td>
<td>3593.5</td>
<td>Gly47–Arg82 (Cys50)</td>
</tr>
<tr>
<td>3626.6</td>
<td>3626.8</td>
<td>Ions generated by in source disruption of the disulfide-linkage of 5712.1 Da</td>
</tr>
<tr>
<td>3659.1</td>
<td>3659.1</td>
<td></td>
</tr>
<tr>
<td>5712.1</td>
<td>5714.7</td>
<td>Gly47–Arg82 (Cys50)</td>
</tr>
<tr>
<td>3626.8</td>
<td>3626.8</td>
<td>Ala90–Arg108 (Cys91, Cys92, Cys97)</td>
</tr>
<tr>
<td>2087.9</td>
<td>2087.9</td>
<td></td>
</tr>
</tbody>
</table>

aThe given mass is calculated on the basis of the observed m/z value ([M + H]+).
bThe calculated mass includes all cysteines in oxidized state. The mass of the individual peptides are given in italics.

Figure 4  The MANEC NMR solution structure ensemble

The Figure shows an alignment of the 20 structures with the lowest energy from the final CYANA calculation (PDB ID: 2msx). The backbone traces are shown as a blue ribbon. The flexible N- and C-termini are shown to the left and right, respectively. The structure Figures and alignments were prepared using PyMOL (the PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC).

Figure 5A and 5D, and Supplementary Table S2), it is now clear why sequence-based homology searches fail in predicting homology to the PAN/apple domain. Only 11 amino acid residues, including four cysteines, are conserved, with an overall sequence homology as low as 13% between the HAI-1 MANEC domain and HGF PAN, the closest structural homologue based on RMSD. Impressively, an in silico structure predicted by the Quark server (http://zhanglab.ccmb.med.umich.edu/QUARK/) [46], compared very well with our NMR structure (Supplementary Figure S7) with an RMSD of 2.5 Å. DALI searches using the predicted structure was again able to pick up homology to the PAN/apple domain family (Supplementary Table S3), showing that in silico structure folding may represent a useful tool to predict structure homologies where sequence-based searches fail.

SAXS analysis shows that the recombinant HAI-1 MANEC domain is a monomer

It is well documented that the PAN domain of HGF supports in the formation of a HGF homodimer as a crucial step for receptor activation [47,48]. In order to verify that our recombinant MANEC domain is monomeric in solution, as suggested by our data obtained by size-exclusion chromatography, we decided to analyse our MANEC preparation by small angle X-ray scattering (SAXS). The shape of the scattering curve for MANEC showed a well-behaved homogeneous sample with only low tendency to aggregation (upturn at low q values) (Figure 6A). Comparing the theoretically calculated scattering profile of our monomeric NMR structure to the SAXS data using CRYOSOL, an almost perfect fit was obtained (Figure 6A). By indirect Fourier transformation of the SAXS data, a Dmax of 43 Å, a radius of gyration of 16.3 ± 0.9 Å and the pair distance distribution function, p(r) (Supplementary...
Figure 5  MANEC represent a new subclass of the PAN/Apple domain family

© The Authors Journal compilation © 2015 Biochemical Society
Figure S6) was determined. The shape of the \( p(r) \) function, in combination with the small radius of gyration, clearly shows that MANEC in solution is a small globular domain. From the SAXS data, it was also possible to construct an \textit{ab initio} low resolution molecular model representing the overall shape of MANEC which fitted very well to our NMR ensemble (Figure 6B). In summary, the SAXS data confirmed that the recombinant HAI-1 MANEC domain is a monomer in solution.

**DISCUSSION**

The hitherto uncharacterized N-terminal region of HAI-1, encompassing the MANEC domain, constitutes a large portion of the extracellular part of the protein (Figure 1). To facilitate a deeper insight into the biological properties of HAI-1 and the MANEC domains as a family, we decided to perform a structural characterization of the MANEC domain of HAI-1.

The purified HAI-1 MANEC domain proved to be highly soluble with a thermodynamically stable fold, allowing structure determination by liquid-state NMR. Our studies produced a structure of the MANEC domain with a resolution of 1.8 Å (calculated during the validation and upload process to http://www.bmrb.wisc.edu). As observed in our structure presentation (Figure 5A) and shown in the secondary-structure labels above the sequence alignment (Figure 5D) some of our \( \beta \)-sheet strands have only partially ideal hydrogen bonding patterns and thus are not picked up by PDB viewing software using standard settings.

After determining the structure, it was possible to use a structure-based homology search with the DALI server, and to demonstrate structural homology between MANEC and members of the PAN/apple domain family. PAN/apple domains have been ascribed important functions in numerous studies, all linked to the ability of mediating protein–protein or protein–oligosaccharide interactions. Deletion of the PAN domain from HGF abolishes both receptor binding as well as heparin binding [49,50]. Our MANEC domain shows no binding to heparin as it passes unaffected through heparin sepharose columns at physiological ionic strength and elutes as a particle of monomeric size from a size-exclusion column in the presence of excess heparin (data not shown). HGF has two natural splice variants, NK1 and NK2, which contain the N-terminal PAN domain, followed by one or two kringle domains, respectively, and finally the inactive protease domain (Supplementary Figure S1). NK1, which is a receptor agonist, has been shown to form a head-to-tail dimer by X-ray crystallography, partly via the PAN domain. Mutations in the NK1 dimer interface convert NK1 into an antagonist [51]. The crystal structure of NK2 was shown to form a ‘closed’ monomeric conformation through interdomain interactions between the PAN domain and the second kringle domain [52]. In the case of plasminogen, the PAN domain is involved in intramolecular domain–domain interactions determining the overall conformation and thus the activation of the proform of the serine protease. In the structure of the full-length ‘closed form’ of plasminogen, the PAN domain appear to interact with kringle IV and kringle V out of the five kringle domains between the PAN and the protease domain (Supplementary Figure S1) [53]. Through its role in the formation of the ‘closed form’ that exhibits weak affinity towards fibrin, the PAN domain is also important for the biological localization of the enzyme [54]. The apple domain of plasma prekallikrein (Supplementary Figure S1) is known to mediate the binding of high molecular mass kininogen [55]. Factor XI of the coagulation cascade comprises of four tandem apple domains and a C-terminal serine protease domain (Supplementary Figure S1). The apple domains of factor XI have been shown to contain binding sites for factor XIIa, platelets, kininogen, factor IX and heparin [56]. In an NMR solution structure of the apple domain IV of factor XI, it was shown that the domain mediates the formation of the disulfide-linked factor XI dimer [57]. These observations underline essential roles of the PAN/apple domain, primarily as mediator of protein:protein interactions. Also, the \textit{primo} location in the N-terminus of a multi-domain protein appears in many cases to be important in order to potentiate a structural effect of interdomain interactions on the overall function of the protein.

In 2008, Kojima et al. proposed a model in which the MANEC domain of HAI-1 plays an active role in the regulation of the inhibitory activity. It was proposed that the tertiary structure of HAI-1 collapses via direct interaction between the MANEC domain and the second Kunitz-type inhibitor domain, resulting in steric interference with the protease–inhibitor interaction [14]. Our findings support this model by introducing homology to other proteins, such as plasminogen and certain forms of HGF, in which intermolecular interactions involving the MANEC/PAN/apple domain have been proven for regulation of protein function. Our MANEC structure also provides clues to where such interaction surfaces may be localized. The presence of two additional conserved disulfide bonds not only stabilizes the already well-defined PAN-like fold, but also appear to introduce the aforementioned unique structural features. The additional fourth disulfide bond between C121 and C129 forces a seven-residue loop between the two cysteines into an extended conformation projecting out from the overall globular shape of the MANEC fold. This loop region (L122–Y128) along with the loop region Q105–A112, are both longer than corresponding linkers of the typical PAN-domain (Figure 5D) and project in opposite directions from both poles of the central \( \beta \)-sheet (Figure 4A). As both extended loops contain multiple charged side chains, they both represent likely areas for mediating protein–protein interactions. Furthermore, the location of the N- and C-terminals and the additional \( \alpha \)-helix with a verified N-linked glycosylation appears to shield a large part of the remaining MANEC surface. We also postulate that the conserved \textit{primo} position of the MANEC in the N-terminus of the parent proteins may reflects a general mechanism by which it participates in the regulation of the

\( (A) \) A representative structure solution (2) of the MANEC domain is shown in cartoon (PDB ID: 2zmx). Secondary structure elements are numbered and highlighted by different colours: \( \alpha \)-helices as sky blue, \( \beta \)-strands as green and loops as wheat. Cysteine side chains including disulfide bonds and are shown as yellow sticks. The \( \beta \)-carbon of the potential glycosylation site at residue 66 is shown as a purple sphere. The location of N- and C-terminal is highlighted by labels. \( (B) \) Alignment of the MANEC domain (colours as in (A)) with the PAN domain from HGF (PDB ID: 1NK1) in red cartoon reveals the high degree of overall structural similarity with an RMSD of 2.3 Å. The N- and C-terminal labels and the cysteine side chains shown as yellow sticks belong to the PAN domain. \( (C) \) Alignment of the MANEC domain (colours as in (A)) with the apple domain from coagulation factor XI (PDB ID: 2F83) in red cartoon with an RMSD of 2.9 Å. The N- and C-terminal labels and the cysteine side chains shown as yellow sticks belong to the apple domain. The structure orientation in (B) and (C) is as in (A) left panel. \( (D) \) Alignment of sequence and secondary structure elements of MANEC domain of HAI-1, PAN domain of HGF and apple domain of coagulation factor XI using the structure-based DALI-server. HAI-1 numbering and secondary structural elements including numbers are indicated above the sequence: \( \alpha \)-helices are drawn as helices, \( \beta \)-strands as arrows, other elements as solid lines. Partially defined elements in our MANEC structure are shown as dashed lines. Disulfide connection-lines are shown below the sequences. The glycosylation site (N66) in HAI-1 MANEC is highlighted by an asterisk (*). The structure Figures and Alignments were prepared using PyMOL (the PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC).
tertiary fold and thus the function of these multi-domain proteins. Why a MANEC domain is preferred by membrane proteins over the common PAN/appeel fold, as found in circulating proteins, remains to be uncovered.

In conclusion, we provide here the first structural characterization of the MANEC domain. The structure will allow for a detailed mapping of potential protein–protein interaction sites in the parent protein HAI-1 by site-directed mutagenesis. The structure will also become a general reference structure for all predicted MANEC domains from the many so far uncharacterized proteins. We were able to use our structure to define MANEC domains as a new subclass of the PAN/appeel domain family. The homology to the PAN/appeel domains seems to suggest a similar function of the MANEC domain as a mediator of protein–protein interactions with potential regulatory properties for the parent protein.

**AUTHOR CONTRIBUTION**

Zebin Hong carried out the majority of the experiments, initial NMR assignment and participated in the preparation of the manuscript. Michal Nowakowski performed NMR and side chain assignment, Chris Spronk supported NMR analysis and structure calculation, Steen Petersen performed mass spectrometry analyses, Witko Koziński supported NMR setup and supplied instrumentation, Frans Mulder performed initial NMR experiments and NMR data analysis and Jan Jensen designed the experiments, analysed the data and wrote the manuscript.

**ACKNOWLEDGEMENTS**

We thank Professor Jan Skov Pedersen for access to SAXS equipment and Christine R. Schar for useful comments on the manuscript.

**FUNDING**

This work is supported by the Danish National Research Foundation [grant number 26-331-6]; the Lundbeck Foundation [grant number R34-A3528] and Danish Cancer Society [grant number R56-A2997-12-S2] and Foundation. M.N. and W.K. was supported by the Danish National Research Foundation [grant number R56-A2997-12-S2] and Foundation. M.N. and W.K. was supported by the Lundbeck Foundation [grant number R34-A3528] and Danish Cancer Society [grant number R56-A2997-12-S2] and Foundation. M.N. and W.K. was supported by the Foundation for Polish Science, TEAM programme.

**REFERENCES**


Received 6 October 2014/12 December 2014; accepted 16 December 2014


37


Received 6 October 2014/12 December 2014; accepted 16 December 2014

Published as BJ Immediate Publication 16 December 2014, doi:10.1042/BJ20141236

© The Authors Journal compilation © 2015 Biochemical Society