The structure of a PKD domain from polycystin-1: implications for polycystic kidney disease

Mark Bycroft^{1,2}, Alex Bateman⁴,
Jane Clarke³, Stefan J. Hamill¹,
Richard Sandford⁵, Ruth L. Thomas⁵
and Cyrus Chothia⁶

¹MRC Centre for Protein Engineering, ³University of Cambridge Department of Chemistry, Lensfield Road, Cambridge CB2 1EW, ⁴The Sanger Centre Wellcome Trust Genome Campus, Cambridge CB10 1SA, ⁵University of Cambridge Department of Medical Genetics, Box 238, Level 3, Addenbrookes Hospital, Hills Road, Cambridge CB2 2QQ and ⁶MRC Laboratory for Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

²Corresponding author
e-mail: mb10031@cus.cam.ac.uk

Most cases of autosomal dominant polycystic kidney disease (ADPKD) are the result of mutations in the PKD1 gene. The PKD1 gene codes for a large cell-surface glycoprotein, polycystin-1, of unknown function, which, based on its predicted domain structure, may be involved in protein–protein and protein–carbohydrate interactions. Approximately 30% of polycystin-1 consists of 16 copies of a novel protein module called the PKD domain. Here we show that this domain has a β -sandwich fold. Although this fold is common to a number of cell-surface modules, the PKD domain represents a distinct protein family. The tenth PKD domain of human and *Fugu* polycystin-1 show extensive conservation of surface residues suggesting that this region could be a ligand-binding site. This structure will allow the likely effects of missense mutations in a large part of the PKD1 gene to be determined.

Keywords: domain/Ig-like fold/PKD1/polycystin-1

Introduction

Autosomal dominant polycystic kidney disease (ADPKD) is a common condition affecting 1 in 800 live births from all ethnic groups. It results in progressive loss of renal function, with more than half of affected individuals requiring renal replacement therapy by their sixth decade. Overall, this one condition accounts for 8% of the renal dialysis population. ADPKD is characterized by the formation of multiple cysts in both kidneys derived from tubular epithelium which, in addition to a decline in renal function, also causes considerable renal enlargement and predisposes to haemorrhage, urinary tract infection and nephrolithiasis. Other more variable features of the condition include hepatic and pancreatic cysts, hypertension, cardiac valvular abnormalities and intracranial vascular malformations (Pirson *et al.*, 1998).

ADPKD is genetically heterogeneous with mutations in at least three different loci producing an identical phenotype. Mutations in the PKD1 and PKD2 genes account

for virtually all cases of ADPKD with most individuals (85%) linked to PKD1 (Peters and Sandkuijl, 1992). There is good evidence that, in addition to a germline mutation in PKD1 or PKD2, a second somatic mutation is also required for cyst formation (Qian *et al.*, 1996; Brasier and Henske, 1997).

The PKD1 gene is located in 16p13.3 and contains 46 exons that span 52 kb of genomic DNA encoding a 14 kb mRNA (Hughes *et al.*, 1995; The International Polycystic Kidney Disease Consortium, 1995). The PKD1 gene product, polycystin-1, is an evolutionarily conserved 4302 amino acid protein that is predicted to contain a large N-terminal extracellular domain of ~2500 residues, several multiple transmembrane domains and a short cytoplasmic C-terminal region (Hughes *et al.*, 1995; The International Polycystic Kidney Disease Consortium, 1995; Sandford *et al.*, 1997). Immunolocalization demonstrates that the protein is expressed widely in a variety of epithelial cell types and is located at lateral cell membranes (Geng *et al.*, 1996; Ibraghimov-Beskrovnaya *et al.*, 1997; vanAdelsberg *et al.*, 1997).

Most of the extracellular segment of polycystin-1 is made up of 16 copies of an 80–90 amino acid repeat (Hughes *et al.*, 1995; The International Polycystic Kidney Disease Consortium, 1995). One of the repeats is situated between a leucine-rich repeat domain and a C-type lectin domain, whilst the rest are arrayed in tandem between a LDL-A module and region which extends over 1000 amino acids and is homologous to sea urchin Receptor for Egg Jelly (REJ) (Moy *et al.*, 1996; see Figure 1). These repeats are also present in a number of other proteins. Single copies are found in Pmel17, MMP and Nmb, all of which are melanocyte cell-surface proteins. They are also found in three bacterial collagenases, a bacterial protease and a surface-layer protein from *Methanothermobacter* (The International Polycystic Kidney Disease Consortium, 1995). It has been predicted that the repeats would contain an antiparallel β -sheet and that they represent a novel protein module termed a PKD domain (The International Polycystic Kidney Disease Consortium, 1995). It has been suggested also that the repeats are members of the immunoglobulin superfamily and are most related to the I-set of this superfamily (Hughes *et al.*, 1995). The sequence similarity to the Ig superfamily is very low, however.

The precise structural organization of polycystin-1 is crucial to the understanding of its function. Therefore a detailed knowledge of the three-dimensional structure of the PKD domains will give valuable insights into the function of a significant proportion of the polycystin-1 molecule and provide the basis for determining potential ligand-binding sites and understanding of the effects of missense mutations in the PKD1 gene. In this paper, we report the determination of the solution structure of one

Fig. 1. A schematic representation of the domains in polycystin-1. LRR, leucine-rich repeat domain with flanking cysteine-rich regions. Lectin, C-type lectin domain. LDL, LDL-A module. PKD, PKD domains. REJ domain, similar to sea urchin REJ protein. TM, transmembrane domains.

Fig. 2. The ^1H ^{15}N HSQC spectrum of the first PKD domain of human polycystin-1. The backbone resonances of the domain are labelled. Other peaks are from side chains or from vector-derived amino acids.

of these repeats from human polycystin-1. We show that the repeat has a Greek key β -sandwich topology. Although this topology, known as the immunoglobulin (Ig)-like fold, is similar to that found in other domains present in cell-surface proteins, the details of the structure show that these domains represent a novel protein family.

Results and discussion

Structure determination

A construct corresponding to the first PKD domain of human polycystin-1 PKDd1 (amino acids V268–P356), was expressed in *Escherichia coli*, purified, and shown to give good nuclear magnetic resonance (NMR) spectra and be stable at concentrations suitable for NMR spectroscopy at pH 3.0, 25°C, for 3–4 weeks (Figure 2). Solution structures were calculated using distance and torsion angle restraints derived from multidimensional NMR spectra recorded using protein samples enriched with ^{13}C and/or ^{15}N . The structural statistics for the final 20 solution structures are summarized in Table I. The 78-residue region between residues A275 and E353 was found to be ordered and this concurs with the limits of the domain as judged from sequence alignment (see below). Within these

Table I. Structural statistics

Root mean squared deviations from experimental constraints	
Distance (Å)	0.037 ± 0.001
Angle (degrees)	0.47 ± 0.08
Root mean squared deviations from idealized covalent geometry	
Bonds (Å)	0.004 ± 0.0001
Angles (degrees)	0.93 ± 0.01
Improvers (degrees)	0.60 ± 0.01
Measures of structural quality for residues 8–87 ^a (Ramachandran plot)	
% residues in most favoured regions	69.5
% residues in additionally allowed regions	25.6
% residues in generously allowed regions	3.7
% residues in disallowed regions	0.0
Number of bad contacts	4 ± 1
Hydrogen bond energy	0.74 ± 0.034
Overall G-factor	-0.28 ± 0.06
Root mean squared distribution around the mean coordinate positions for residues 6–88	
All backbone atoms (Å)	0.47 ± 0.07
All heavy atoms (Å)	0.81 ± 0.10

^aDetermined using the program PROCHECK (Laskowski *et al.*, 1993). All variances are quoted ± 1 SD.

Fig. 3. (A) A stereo view of the superposition of the backbone atoms (residues 7–88) of the 20 NMR solution structures of the first PKD domain of human polycystin-1. (B) A stereo view showing the superposition of the main hydrophobic core residues with the backbone trace of the lowest energy structure. The core residues shown are: A8, L10, A23, F25, I27, T36, W38, F40, V47, A49, A54, H56, Y58, Y64, V66, A68, L70, A77, L79, T81, V83, V85. The residues have been renumbered with U268 set to residue 1.

limits the structures are well defined with the exception of the loop between the A and B strands (Figure 3).

The PKD domain has a β -sandwich fold

The PKD domain is built from two β -sheets, one of three strands and one of four strands, which are packed face to face (Figure 4). The number and arrangement of these

strands is the same as that found in some other proteins including certain members of the immunoglobulin superfamily and the fibronectin type III (FNIII) superfamily. This fold is called the Ig-like fold (Bork *et al.*, 1994). In the PKD domain, one β -sheet has its three strands labelled A, B and E and the other β -sheet has its four strands labelled G, F, C and C'. The labels are used to indicate

Fig. 4. (A) PKD domain 1 showing the elements of secondary structure. The β -strands are labelled. The figure was prepared using the program Molscript (Kraulis, 1991). (B) A schematic diagram of the fold of the PKD domain. Black lines are used for one sheet (strands G, F, C and C') and grey lines for the other sheet (strands A, B and E). The identity of each residue in the PKD domain is shown in the one letter code. The identities at buried sites are shown in boxes. Horizontal thin lines between residues represent the main-chain hydrogen bonds that form the β -sheets. The first and last residues in each strands are labelled with the strand and number, so A1 is the first residue in the A strand and A6 is the last.

their structural equivalence to the strands in other proteins with the Ig-like fold.

The two sheets pack together with a well defined hydrophobic core, centred around a conserved tryptophan located in the C strand. One unusual feature of the core is the presence of a conserved histidine residue (in position E5) close to the tryptophan. The E-F loop that spans the two sheets contains a tyrosine corner motif, with the hydroxyl oxygen of Tyr64 hydrogen bonding to the backbone amide of Leu60. This structural motif is found in many Greek key proteins (Hemmingsen *et al.*, 1994).

A noteworthy feature of this domain is the unusual amino acid composition. The sequence of PKDd1 is rich in light amino acids. Alanine and glycine alone comprise over a third of the residues in this domain. Residues 268 to 356 have an average mass per residue of 100 Da. This is lighter than 99% of sequences in SWISS-PROT 34 (Bairoch and Apweiler, 1998). Such light proteins are usually those with regions of low complexity and are uncharacteristic of globular structures (Wootton, 1994). In PKDd1 there are five positions in the core, which in most

PKD domains are occupied by larger hydrophobic residues, that contain alanine. The unusually high alanine and glycine content of the domain means that fewer nuclear Overhauser effects (NOEs) can be assigned compared with other proteins of similar size, as it reduces the number of potential proton-proton contacts and leads to considerable resonance overlap.

The amino acid composition of the domain is also unusual for an all β -protein, emphasizing that structural context is more important than the secondary-structure-forming propensity of the individual amino acids.

The structure of PKD domains 2–16

We have described here the structure of human polycystin-1 PKDd1. Polycystin-1 contains 16 homologous PKD domains. To what extent do the other 15 PKD domains have the same structure? In Figure 5 we present an alignment of the sequences of the 16 domains, together with information for domain 1 on the residues which are in strands and residue-accessible surface areas.

From the data in Figure 5, we see that the sequence

Strand:		A			B			C			C'			E			F			G		
Site:		123456			123456789			123456			12345			1234567			12345678901			123456789012		
ASA:		9795	057425	9063554	970305014	69938	409051	2469	29192	81	4407090	877	29390204045	79	450915370909							
PKD1	271	ASPG-	ATLVGP	HGPLASG	QLAAFHIAA	PLPVT	ATRWDF	GDGS	AEVDA	AG	PAASHRY	VLP	GRYHVTAVLAL	GA	GSALLGTDVQVE	353						
PKD2	844	ALVLQ	VDSGAn	a	aTARWPGGS	vsarfenvcp	ALVATF	vpqcpw	ETNDT	LF	SVVALFW	LSE	GEHVVDVVVEN	SA	SRANLSLRVTAE	929						
PKD3	930	EPICG	LRATPS	pearvlqg	VLVRYSPVV	eagsdm	VERWTI	ndk	qSLTF	QN	VVENVIY	QSA	AVFKLSLTASN	HV	SNVTVNYNVTVE	1014						
PKD4	1015	-RMRN	MOGLQV	stvpavlspn	ATLVLTGGV	lvdsavev	APLNWF	gd*ps	VAQVL	VE	HNVMHTY	AAP	GEYLLTVLASN	AF	ENLTQQQVPVSVR	1123						
PKD5	1124	ASLPS	VAVGVS	DGVLVAG	RPVTFYPHP	lpspggv	LXTWDF	GDGS	PVLTTQ	SQ	PAANHYY	ASR	GTYHVRLEVNN	TV	SGAAAQADVRVF	1209						
PKD6	1210	EELRG	LSVDMS	laveqg	APVVVSAAV	qtgd	NITWTF	dmgdg	TVLSG	PE	ATVEHVIY	LRA	QNCVTIVGAAS	PA	GHLARSLHLVLF	1292						
PKD7	1295	-EVL	VEPAAC	iptq	PDARLTAYV	tgnpahy	LEFWTF	gdgss	NTTVR	GC	PTVTHNF	TRS	GTFFPLALVLS	RV	NRAHYFTSICVE	1377						
PKD8	1378	PEVGN	VTIQPE	rqfvlqg	DEAWLVACA	wppfry	RYTWDF	gteea	APTRA	RG	PEVTIFY	RDP	GSYLVTIVTASN	NI	SAANDSALVEVQ	1463						
PKD9	1464	EPVLV	TSIKVN	gslglqlq	QPYLFSAVG	rgrpa	SYLWDL	gd	ggw1	EG	PEVTHAY	NST	GDFTVRVAGWN	EV	SRSEAWLNVTVK	1545						
PKD10	1546	RRVRG	LNVNAS	rtvvpln	GSVSFSTSL	eagsdv	RYSWVL	cdr	ctpip	GG	PTISYTF	RSV	GTENIIVTAEN	EV	GSAQDSIFVYVL	1629						
PKD11	1630	QLIEG	LQVVG	gryfptn	HTVQLQAVV	rdgtvn	SYSWTA	wrdrg	PALAG	SG	KGFSLTV	LEA	GTYHVQLRATN	ML	GSAAWADCTMDFV	1715						
PKD12	1716	EPVGV	LMVTAS	pnpaavn	TSVTLGAEL	aggsqv	VYTWSL	eeg	LSWET	SE	PFTTHSF	PTP	GLHLVTMTAGN	PL	GSANATVEVDVQ	1799						
PKD13	1800	VPVSG	LSIRAS	epggsfvaag	SSVFPWQQL	atgt	NVSWCW	avp	gGSSK	RG	PHVTMVF	PDA	GTFSIRLNASN	AV	SWVSATYNLTAE	1884						
PKD14	1885	EPIVG	LVLWAS	skvpap	QLVHFQILL	aagsav	TERLQV	gga	nPEVL	PG	PRFSHSF	PRV	GDHVSVRGKN	HV	SWAQVQVRIIVL	1968						
PKD15	1969	EAVSG	LQMPNC	cepgiatg	TERNFTARV	qrgsr	AYAWYF	slqkvqgd	SLVIL	SG	RDVITYP	VAA	GLLEIQVRAFN	AL	GSENRITLVLEVQ	2058						
PKD16	2059	DAVQY	VALQSG	pcftn	RSAQFEAAT	spsprv	AYHWDF	GDGS	PGQDT	DE	PRAEHSY	LRP	GDYRQVQNASN	LV	SFFVAQATVIVQ	2142						

Fig. 5. PKD domains of human polycystin-1. Structural alignment of domain 1 with the sequences of domains 2–16. In domains 1–15 the regions that very probably have the same conformation as the equivalent regions in domain 1 are shown in upper case letters. Regions that probably have different conformations are shown in lower case letters. Turns that are very short in domains 2–16, but not in domain 1, probably affect to a small but uncertain extent the conformation of regions that would be expected to form strands. These regions are indicated by lower case letters in the first or last residues of the relevant strands. ASA, residue accessible surface areas of domain 1 are given in units of 10. Thus, 1 indicates 0–9 Å², 2 indicates 20–29 Å² etc., until 9, which indicates 90+ Å². PKD4 has additional residues in the CC' loop *: GEQALHQFQPPYNESFPVPD. The alignment of PKD2 in the region of the C strand is uncertain.

identities between domain 1 and domains 2–15 are relatively low, mostly in the range 20–30%. In spite of these low identities, the residues' patterns, in the regions equivalent to those that form strands in domain 1, suggest that they have largely the same secondary structure. At sites buried in domain 1, the residues are hydrophobic or, occasionally, neutral. At buried sites on edge strands, or on the ends of inner strands, the substitution by arginine or glutamate can occur (see for example, positions G5 in domains 6, 12 and 15 and G7 in domain 12) because these residues are long enough for the hydrophobic part of the side chain to pack against the core, leaving the charged portion solvent exposed. At deeply buried sites the residues appear to be of similar size. Thus, the alignments in Figure 5 can be used to derive outline structures for the β -sheet regions of domains 2–15 and information on whether residues in these regions are on the surface or in the interior. Previously we have used similar alignments and structural comparisons to predict accurately the structure of a β -sandwich protein that is a member of the Ig superfamily (Fong *et al.*, 1996).

For the loop regions AB, BC and CC', the large variations in size and absence of conserved sites in domains 2–16 indicate that they have conformations quite different to that in domain 1 (the residues in these regions are shown in lower case letters in Figure 5). On the other hand, loops C'E, EF and FG are the same size as those in domain 1, and often conserve the residues immediately adjacent to the strands: at sites E1, E7, F1 and G1. Thus it is likely that most of the C'E, EF and FG loops in domains 2–15 have conformations that are the same, or close to, those found in domain 1.

PKDd1 is situated between the leucine-rich repeat region and C-type lectin domain of polycystin-1. The other 15 PKD domains are all arrayed in tandem elsewhere in the protein. Based on the alignments, most of the 15 domains are linked by six residues, the only exceptions

being the links between domains 3 and 4 and between domains 6 and 7, which are five and seven residues long respectively. The sequence of the interdomain linkers are fairly variable, with the exception of a hydrophobic residue equivalent to P273 in PKDd1 which is well conserved (Figure 5). The second residue in the FG turn is a large hydrophobic residue in most of the arrayed domains. This residue may participate in interdomain contacts with the preceding PKD domain.

PKD domains are not members of the Ig superfamily

Many proteins and protein domains have a β -sandwich topology similar to that seen in PKD domains (Bork *et al.*, 1994). The fold of these proteins is called the immunoglobulin-like fold after the first structure in which it was seen (Bork *et al.*, 1994). When classifying protein structures, it is important to distinguish between proteins that happen to share a similar fold and proteins that, based on sequence and structural evidence, are clearly evolutionarily related and fall into distinct protein families and superfamilies (Murzin *et al.*, 1995). Not all proteins with the Ig-like fold are related, nor are they all members of the Ig superfamily. Whilst they have the same overall fold these proteins differ in the arrangement of the peripheral regions, in particular in the positioning of the edge strands of the β -sheets. They also differ in the packing of the hydrophobic core. The topology of the PKD domain with sheets formed by the A, B, E and G, F, C and C' strands is the same as that found in FNIII domains and the C2-set of the Ig superfamily. Proteins have the same fold because either they are descended from a common ancestor or because physics and chemistry favour certain chain topologies and packing arrangements for secondary structures. Do PKD domains have an evolutionary relationship to other proteins with the Ig-like

Fig. 6. A diagram showing the relative position of the core tyrosine and tryptophan residues in the F and C strands of Ig and PKD domains. The squares represent residues pointing into the interior of the protein; the filled circles represent residues pointing into the solvent.

fold or do they just have one of the folds that are particularly favourable in common?

Hughes *et al.* (1995) analysed the sequences of 16 PKD domains and concluded that they are members of a branch of the Ig superfamily known as the I set (Harpaz and Chothia, 1994). This was because (i) residue patterns in the aligned sequences indicated β -strands like those in the immunoglobulins and (ii) as in the immunoglobulins, there was a strongly conserved tryptophan residue in the putative C strand and a conserved tyrosine at the beginning of the putative F strand. We have seen here that the fold of the PKD domain is indeed like that of immunoglobulins. However, the relative arrangement of the tryptophan and tyrosine residues is different. In both proteins the C and F strands are adjacent but, in PKD, the tryptophan in C is separated from a position adjacent to the tyrosine by four residues whereas in the immunoglobulins it is separated by two residues (Figure 6). This difference in position produces large differences in the way residues pack in the interior of the structures. Thus the structural data show that the argument for an evolutionary relationship between PKD domains and the immunoglobulin superfamily, based on the presence of tryptophan and tyrosine in the C and F strands, is not valid. The present structural evidence indicates that PKD domains and immunoglobulin domains share a common fold but not a common ancestor.

Comparisons of the PKD structure with those of non-immunoglobulin proteins with the Ig-like fold, such as FNIII domains, do not provide good evidence for an evolutionary relationship. The FNIII domains strongly (but not absolutely) conserve a tryptophan residue in the B strand. In the PKD domains the conserved tryptophan is not only in a different strand (C), but in a different sheet. This produces large differences in the core packing around the central tryptophan. Thus, at the present, the

Fig. 7. PKDd1 showing the conserved structurally important sequence WDFGDGS. The figure was prepared using the program Molscript (Kraulis, 1991).

PKD domains are best seen as proteins that have the Ig-like fold but are probably without an evolutionary relationship to other proteins of known structure that have this fold.

Like other β -sandwich proteins the origin of PKD domains is intriguing. PKD domains are found in all completely sequenced genomes of archaeobacteria where they are part of surface-layer proteins. This suggests that PKD domains existed in the ancestral archaeobacteria. Within eubacteria, the distribution of PKD domains is uneven. Where found, they are linked to collagenases and proteases. This distribution is reminiscent of FNIII domains in bacteria that have probably been horizontally transferred from animals (Little *et al.*, 1994). In eukaryotes, PKD domains have only been identified in two vertebrate proteins so far. Searches of the *Saccharomyces cerevisiae* genome do not detect any PKD domains. Thus, the evolution of PKD domains is unclear.

The WDFGDGS motif has a structural role

The sequence WDFGDGS is the most conserved sequence in PKD domains. The function of this motif is an important issue. Based on a proposed alignment of PKD domains with immunoglobulin I-set domains, it was suggested that this motif could be involved in ligand binding (Hughes *et al.*, 1995) since in an I-set domain of VCAM-1 the equivalent region participates in integrin binding (Jones *et al.*, 1995). In the PKD domain these residues form part of the C strand and the turn between the C and C' strands (Figure 7). This region of the structure is very different from the I-set domains as the C strand in I-set domains is followed by a D strand that is part of the other sheet. The analogy with the ligand-binding site in I-set proteins is thus not appropriate. The tryptophan and phenylalanine residues of the WDFGDGS motif are deeply buried and form an important part of the hydrophobic core of the domain. The residue between the tryptophan and phenylalanine is on the outside of the C strand and points into the solvent. Inspection of PKD domain sequences shows that it is not particularly well conserved. The other residues of this sequence comprise the turn at the end of the C strand and are involved in contacts that stabilize this turn. The conservation of the residues in this turn could possibly be of functional importance. It is, however, unlikely that all PKD domains with this motif have the same binding function, and it seems more likely that the conservation seen is due to constraints on structure rather than function.

Fig. 8. (A) A structural alignment of the sequences of human and *Fugu* PKD domain 10 (PKDd10) with that of human PKDd1. (B) Schematic representation of the sequence conservation in the sheets of PKDd10. The surface residues only are shown, in a view looking at the molecule from the outside. The filled circles represent residues pointing into the interior of the protein. The C'–C–F–G sheet is shown in the same orientation as used in Figure 3. The A–B–E sheet is shown with the entire molecule rotated 180° about the y-axis. The human sequence is on the left, the *Fugu* sequence on the right. Residues in the A–B–E sheet that are putative glycosylation sites are shaded.

Conservation of surface residues suggests PKD domain 10 has a functional role

The sequence of the PKD1 gene from the pufferfish *Fugu rubripes* has been determined (Sandford *et al.*, 1997). Given the large evolutionary distance between fish and humans it is probable that regions conserved between proteins in the two species are of functional importance. Both human and *Fugu* polycystin-1 contain 16 PKD domains but only PKDd10 (domain 10) is well conserved. Inspection of the sequences in light of the structure of the domain shows extensive conservation of surface residues (Figure 8). The sequence conservation is particularly striking for the C, F and G strands where all of the surface residues are conserved, making a conserved 'face' on one β -sheet. High levels of sequence conservation are also seen in the EF turn. (The other sheet shows no significant conservation of surface residues, except for residues in the A strand and AB loop, which are conserved glycosylation sites.) There is no apparent structural reason for this

surface conservation and it strongly suggests a functional role for PKDd10.

Conclusion

The cloning of the PKD1 gene (Hughes *et al.*, 1995; The International Polycystic Kidney Disease Consortium, 1995) provides an opportunity to gain a detailed understanding of the molecular pathogenesis of ADPKD. Currently, one of the most important areas of research is to define the normal cellular function of polycystin-1 so that the mechanism of cyst formation as a result of mutations in PKD1 can be understood. The C-terminal cytoplasmic portion of polycystin-1 contains a coiled-coil domain that interacts *in vitro* with a coiled-coil domain in the C-terminal cytoplasmic region of the PKD2 gene product (Qian *et al.*, 1997; Tsiokas *et al.*, 1997). In addition, it has a role in the activation of the transcription factor AP-1 (Arnould *et al.*, 1998) suggesting that PKD1 and PKD2 are components of a common signalling pathway that regulates renal tubular cell function. The N-terminal extracellular portion contains domains that in other proteins are known to participate in protein–protein and protein–carbohydrate interactions suggesting that polycystin-1 makes a complex series of interactions with different proteins and components of the extracellular matrix (Hughes *et al.*, 1995; The International Polycystic Kidney Disease Consortium, 1995).

To understand these interactions in such a large complex protein, it is necessary to determine the function of individual domains. In this paper we have shown that the PKD domain, of which there are 16 copies in polycystin-1, has an Ig-like fold. The PKD domain is not a member of any previously known protein family so direct comparison with other domains is not possible. However, since domains with the Ig-like fold have been shown to form the ligand-binding sites in cell-surface proteins, it is possible that PKD domains could play a similar role. The extensive conservation of surface residues on the GFCC' face of PKDd10 points to this domain as the strongest candidate for a ligand-binding site in the PKD domains of polycystin-1. These ideas can be tested by biochemical experiments.

The majority of cases of ADPKD are linked to mutations in the PKD1 gene and currently much effort is being put into identifying point mutations in the PKD1 gene. The structure presented here has been used as a template for structural modelling of one-third of the PKD1 protein. The identification of pathological mutations in the region of the gene containing the PKD domains has proved difficult. This is due almost exclusively to the reiteration of the majority of the PKD1 gene elsewhere in the genome (The European Polycystic Kidney Disease Consortium, 1994). Once the repeated region can be sequenced in patients, it will be possible, using the structure reported here, to predict if a missense mutation will disrupt the basic fold of the PKD domains or directly affect a binding function. These data will help to define the functions of polycystin-1 and help to unravel some of the complexities of ADPKD.

Materials and methods

Protein production

The DNA sequence corresponding to residues 268 to 356 of human polycystin-1 were amplified using PCR from a cDNA clone of the

human PKD1 gene (IbraghimovBeskrovnaya *et al.*, 1997) using primers: 5'-GCGCGGATCCGCTCTTCCCTGCCTCCCCAG-3'; and 5'-GGC-CGAATCTTAAGGTGCCGCTTCCACCTGC-3'. The PCR product was cloned into a pRSETc expression vector (Invitrogen) that had been modified to include a thrombin cleavage site after the hexahistidine sequence. The histidine-tagged PKD domain was purified from *E.coli* C41(DE3) cells (Miroux and Walker, 1996) transformed with the expression plasmid grown in M9 minimal media. ^{13}C glucose and ^{15}N ammonium chloride were used as the carbon and nitrogen sources when isotopically enriched protein was desired. When grown at 25°C, the fusion protein was recovered in the soluble fraction following sonication, but was in the insoluble fraction when the cells were grown at 37°C. The protein could be refolded by dialysis and the spectra of soluble and refolded protein were identical. Since at 37°C total protein production increased, the domain was purified using a refolding step. The fusion protein was purified in urea using Qiagen NTA-agarose resin according to the manufacturer's instructions. The protein was refolded by dialysis into pH 8.0 Tris buffer. The N-terminal histidine tag was removed by digestion with thrombin and the PKD domain was purified to homogeneity on a Fast Flow Sepharose Q column (Pharmacia). For NMR experiments, concentrated protein was diluted rapidly into pH 3.0 deuterated acetate buffer and concentrated to 1–2 mM in a Centrprep concentrator (Amicon).

NMR spectroscopy and structure calculations

A series of ^1H ^{15}N HSQC spectra were recorded on 1 mM samples of the PKD domain at a number of pH values ranging from pH 3.0 to pH 7.0 to determine the most suitable conditions for NMR. The spectra were very similar over this pH range. At higher pH values the sample life time was in the order of 1–2 days, whilst at pH 3.0 the sample was stable for several weeks. A full set of two- and three-dimensional spectra were, therefore, recorded at pH 3.0 in deuterated acetate buffer. Resonance assignments for the domain were obtained from standard double and triple resonance methods using isotopically enriched protein samples (reviewed in Bax and Grzesiek, 1993). Backbone assignments were obtained using 3D ^1H ^{15}N edited TOCSY, HBHACONH, CBCACONH and HNCACB spectra. Additional side-chain assignments were obtained using an HCCH-TOCSY experiment. The assignments are available from the BioMagResBank. Distance constraints were obtained from the analysis of ^1H ^1H 2D NOESY, 3D ^1H ^{15}N edited NOESY and 3D ^1H ^{13}C edited NOESY experiments all recorded with a 100 ms mixing time. Peaks intensities were classified as strong, medium or weak and were translated in upper distance bounds of 2.7, 3.3 and 5.0 Å. An additional 0.5 Å was added to upper distance bounds for atoms involving methyl protons. Structures were calculated using 209 intra-residue, 216 short-range and 275 long-range NOE constraints. Thirty-nine hydrogen-bond constraints were also used in the calculations. These were identified based on characteristic NOE patterns that are observed for residues in anti-parallel β -strands together with solvent exchange data (Wuthrich, 1986). Backbone ϕ angle constraints for 22 residues were used based on $J_{\alpha\text{H}}$ coupling constants measured in a ^1H ^{15}N HSQC spectrum (Stonehouse and Keeler 1995). χ_1 side-chain constraints for 11 residues were determined from $J_{\alpha\beta}$ and NOE data. Stereospecific assignments for the methyl groups of valine and leucine residues were obtained from the splitting patterns in a high resolution ^1H ^{13}C HSQC spectrum of a 10%, fractionally ^{13}C -labelled, sample (Neri *et al.*, 1989). Fifty structures were calculated using simulated annealing from random starting structures followed by restrained minimization using the program XPLOR 3.1 (Brunger, 1992). The 20 structures with the lowest total energy were selected. These structures have no distance constraint violations greater than 0.5 Å or dihedral angle constraint violations greater than 5°. The 20 final structures have been deposited in the Brookhaven Protein Data Bank.

Acknowledgements

This work was funded by the National Kidney Research Fund, the Medical Research Council and the Wellcome Trust. J.C. is a Wellcome Trust Career Development Fellow. R.S. is a Wellcome Trust Senior Fellow in Clinical Research.

References

- Arnould, T., Kim, E., Tsiokas, L., Jochimsen, F., Gruning, W., Chang, J.D. and Walz, G. (1998) The polycystic kidney disease 1 gene product mediates protein kinase C alpha-dependent and c-Jun N-terminal kinase-dependent activation of the transcription factor AP-1. *J. Biol. Chem.*, **273**, 6013–6018.
- Bairoch, A. and Apweiler, R. (1998) The SWISS-PROT protein sequence database and its supplement TrEMBL in 1998. *Nucleic Acids Res.*, **26**, 38–42.
- Bax, A. and Grzesiek, S. (1993) Methodological advances in protein NMR. *Acc. Chem. Res.*, **26**, 131–138.
- Bork, P., Holm, L. and Sander, C. (1994) The immunoglobulin fold—structural classification, sequence patterns and common core. *J. Mol. Biol.*, **242**, 309–320.
- Brasier, J.L. and Henske, E.P. (1997) Loss of the polycystic kidney disease (PKD1) region of chromosome 16p13 in renal cyst cells supports a loss-of-function model for cyst pathogenesis. *J. Clin. Invest.*, **99**, 194–199.
- Brunger, A.T. (1992) *X-PLOR v3.1 A System for X-ray Crystallography and NMR*. Yale University Press, New Haven, CT.
- The European Polycystic Kidney Disease Consortium (1994) The polycystic kidney disease 1 gene encodes a 14-kb transcript and lies within a duplicated region on chromosome-16. *Cell*, **77**, 881–894.
- Fong, S., Hamill, S.J., Proctor, M., Freund, S., Benian, G.M., Chothia, C., Bycroft, M. and Clarke, J. (1996) Structure and stability of an immunoglobulin superfamily domain from twitchin, a muscle protein of the nematode *Caenorhabditis elegans*. *J. Mol. Biol.*, **264**, 624–639.
- Geng, L. *et al.* (1996) Identification and localization of polycystin, the PKD1 gene product. *J. Clin. Invest.*, **98**, 2674–2682.
- Harpaz, Y. and Chothia, C. (1994) Many of the immunoglobulin superfamily domains in cell-adhesion molecules and surface-receptors belong to a new structural set which is close to that containing variable domains. *J. Mol. Biol.*, **238**, 528–539.
- Hemmingsen, J.M., Gernert, K.M., Richardson, J.S. and Richardson, D.C. (1994) The tyrosine corner—a feature of most Greek key β -barrel proteins. *Protein Sci.*, **3**, 1927–1937.
- Hughes, J., Ward, C.J., Peral, B., Aspinwall, R., Clark, K., Sanmillan, J.L., Gamble, V. and Harris, P.C. (1995) The polycystic kidney-disease-1 (PKD1) gene encodes a novel protein with multiple cell recognition domains. *Nature Genet.*, **10**, 151–160.
- IbraghimovBeskrovnaya, O. *et al.* (1997) Polycystin: *in vitro* synthesis, *in vivo* tissue expression and subcellular localization identifies a large membrane-associated protein. *Proc. Natl Acad. Sci. USA*, **94**, 6397–6402.
- The International Polycystic Kidney Disease Consortium (1995) Polycystic kidney disease: The complete structure of the PKD1 gene and its protein. *Cell*, **81**, 289–298.
- Jones, E.Y., Harlos, K., Bottomley, M.J., Robinson, R.C., Driscoll, P.C., Edwards, R.M., Clements, J.M., Dudgeon, T.J. and Stuart, D.I. (1995) Crystal-structure of an integrin-binding fragment of vascular cell-adhesion molecule-1 at 1.8 Å resolution. *Nature*, **373**, 539–544.
- Kraulis, P.J. (1991) Molscript—A program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.*, **24**, 946–950.
- Laskowski, R.A., MacArthur, M.W., Moss, D.S. and Thornton, J.M. (1993) Procheck—A program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.*, **26**, 283–291.
- Little, E., Bork, P. and Doolittle, R.F. (1994) Tracing the spread of fibronectin type III domains in bacterial glycohydrolases. *J. Mol. Evol.*, **39**, 631–643.
- Miroux, B. and Walker, J.E. (1996) Over-production of proteins in *Escherichia coli*—mutant hosts that allow synthesis of some membrane-proteins and globular-proteins at high-levels. *J. Mol. Biol.*, **260**, 289–298.
- Moy, G.W., Mendoza, L.M., Schulz, J.R., Swanson, W.J., Glabe, C.G. and Vacquier, V.D. (1996) The sea-urchin sperm receptor for egg jelly is a modular protein with extensive homology to the human polycystic kidney-disease protein, PKD1. *J. Cell Biol.*, **133**, 809–817.
- Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) SCOP—a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
- Neri, D., Szyperski, T., Otting, G., Senn, H. and Wuthrich, K. (1989) Stereospecific nuclear magnetic resonance assignments of the methyl groups of valine and leucine in the DNA binding domain of the 434-repressor by biosynthetically directed fractional C-13 labeling. *Biochemistry*, **28**, 7510–7516.
- Peters, D.J.M. and Sandkuyl, L.A. (1992) Genetic heterogeneity of polycystic kidney disease in Europe. In Breuning, M.H., Devoto, M. and Romeo, G. (eds), *Contributions to Nephrology, Polycystic Kidney Disease*. Karger, Basel, Switzerland.

- Pirson,Y., Chauveau,D. and Grunfeld,J.P. (1998) Autosomal dominant polycystic renal disease. In Alex Davison,J.S.C., Grunfeld,J.-P., Kerr,D., Ritz,E. and Winearls,C. (eds) *The Oxford Textbook of Clinical Nephrology*. 2nd edition. Oxford University Press, Oxford, UK, pp. 2393–2410.
- Qian,F., Watnick,T.J., Onuchic,L.F. and Germino,G.G. (1996) The molecular basis of focal cyst formation in human autosomal dominant polycystic kidney disease type-I. *Cell*, **87**, 979–987.
- Qian,F., Germino,F.J., Cai,Y.Q., Zhang,X.B., Somlo,S. and Germino,G.G. (1997) PKD1 interacts with PKD2 through a probable coiled-coil domain. *Nature Genet.*, **16**, 179–183.
- Sandford,R. *et al.* (1997) Comparative analysis of the polycystic kidney disease 1 (PKD1) gene reveals an integral membrane glycoprotein with multiple evolutionary conserved domains. *Hum. Mol. Genet.*, **6**, 1483–1489.
- Stonehouse,J. and Keeler,J. (1995) A convenient and accurate method for the measurement of the values of spin-spin coupling constants. *J. Magn. Res., A* **112**, 43–57.
- Tsiokas,L., Kim,E., Arnould,T., Sukhatme,V.P. and Walz,G. (1997) Homo- and heterodimeric interactions between the gene products of PKD1 and PKD2. *Proc. Natl Acad. Sci. USA*, **94**, 6965–6970.
- vanAdelsberg,J., Chamberlain,S. and Dagati,V. (1997) Polycystin expression is temporally and spatially regulated during renal development. *Am. J. Physiol.*, **272**, F602–F609.
- Wootton,J.C. (1994) Sequences with unusual amino-acid compositions. *Curr. Opin. Struct. Biol.*, **4**, 413–421.
- Wuthrich,K. (1986) *NMR of Proteins and Nucleic Acids*. John Wiley, New York, NY.

Received July 20, 1998; revised October 23, 1998;
accepted November 6, 1998