When a module is not a domain: the case of the REJ module and the redefinition of the architecture of polycystin-1

Samantha SCHRÖDER*, Franca FRATERNALI†, Xueping QUAN†, David SCOTT‡, Feng QIAN§ and Mark PFUHL*†

*Department of Biochemistry, University of Leicester, Leicester LE1 9HN, U.K., †Randall Division of Cell and Molecular Biophysics, King’s College London, Guy’s Campus, London SE1 1UL, U.K., ‡Department of Biosciences, University of Nottingham, Sutton Bonington, Leicestershire LE12 5RD, U.K., and §School of Medicine, The Johns Hopkins University, Baltimore, MD 21205, U.S.A.

The extracellular region of a group of cell-surface receptors known as the polycystic kidney disease 1 family, containing, among others, polycystin-1, has been controversially described as containing four FNIII (fibronectin type III) domains or one REJ (receptor of egg jelly protein) module in the same portion of polypeptide. Stimulated by recent atomic force microscopy work, we re-examined the similarity of these four domains with a FNIII sequence profile showing the evolutionary relationship. Two of the predicted domains could be expressed in bacteria and refolded to give a protein suitable for biophysical study, and one of these expressed solubly. CD spectroscopy showed that both domains contain a significant amount of β-sheet, in good agreement with theoretical predictions. Confirmation of independent folding as a domain is obtained from highly cooperative thermal and urea unfolding curves. Excellent dispersion of peaks in the high-field region of one-dimensional NMR spectra confirms the presence of a hydrophobic core. Analytical ultracentrifugation and analytical gel filtration agree very well with the narrow linewidths in the NMR spectra that at least one of the domains is monomeric. On the basis of this combined theoretical and experimental analysis, we show that the extracellular portion of polycystin-1 does indeed contain β-sheet domains, probably FNIII, and that, consequently, the REJ module is not a single domain.

Key words: domain, fibronectin, module, polycystic kidney disease, polycystin-1 (PC1), receptor of egg jelly (REJ).

INTRODUCTION

A continuous segment of protein sequence that shows a high degree of sequence similarity in a range of different proteins is usually called a module [1,2]. This definition does not carry an explicit link to the structure into which the module might fold. Yet it is usually assumed that a module is a domain, i.e. that it is able to fold autonomously into a well-defined structure and that it cannot be cut down any further without losing its ability to fold properly [3,4]. For most modules this is the case, therefore the increased availability of newly sequenced proteins and the analysis of their module organization has given a significant boost to structural biology. The ability to cut large proteins down to their constituent modules greatly facilitated their structural and functional characterization. The automatic annotation of protein genes makes extensive use of established consensus sequences of modules even in cases where no experimental data have confirmed the relationship of modules and domains. Such information is extensively used as the base for numerous experimental studies of proteins where modules are mutated, added, swapped or deleted in the assumption that they are folded autonomously and make a defined contribution to the overall function of the protein.

In most cases, the assumptions made in the annotations of protein sequences turn out to be true. In others, however, even where structures are known, as in the case of the C2 domain fold [5], annotations may give a wrong estimate of the true size of the domain, resulting in a range of inconclusive experimental results. On the other hand, in the absence of any detailed information, large stretches of highly similar sequence are assigned as a module and thus classified as domain simply because they occur in a number of different proteins. One of such examples is the REJ module, which comprises a large portion of the extracellular region of a number of vertebrate cell-surface proteins. Its name derives from the protein in which this module was described for the first time, the receptor of egg jelly protein [6]. This module has a size of ~900 amino acids with no obvious homologues in sequence databases. It is found in suREJ (sperm receptor for egg jelly), PKDREJ (polycystic kidney disease and receptor for egg jelly-related protein), a number of uncharacterized proteins from genomic sequencing projects [7,8] and PC1 (polycystin-1). The latter is of specific medical interest because mutations in its gene, PKD1, are the main cause for ADPKD (autosomal dominant polycystic kidney disease) for which there is currently no cure [9]. ADPKD-related mutations are spread evenly throughout the entire PKD1 gene. At present, the only disease caused by mutations in PKD1 is ADPKD through the loss-of-function of PC1. REJ modules usually occur in the vicinity of the GPS (G-protein-coupled receptor proteolytic site) domain which contains an autoproteolytic motif [7,10]. Autoproteolysis is essential for full functionality of PC1 [11] and takes place after N-glycosylation of the protein [12–14]. Several mutations in the REJ module that cause ADPKD interfere with autoproteolysis [15,16] (Figure 1), suggesting an important function for the REJ module. Interestingly, in the first description of the gene for PC1 (PKD1) [17], there was no mention of the REJ module. Instead, it was proposed that the corresponding region should contain a significant amount of β-sheet, in good agreement with theoretical predictions. Confirmation of independent folding as a domain is obtained from highly cooperative thermal and urea unfolding curves. Excellent dispersion of peaks in the high-field region of one-dimensional NMR spectra confirms the presence of a hydrophobic core. Analytical ultracentrifugation and analytical gel filtration agree very well with the narrow linewidths in the NMR spectra that at least one of the domains is monomeric. On the basis of this combined theoretical and experimental analysis, we show that the extracellular portion of polycystin-1 does indeed contain β-sheet domains, probably FNIII, and that, consequently, the REJ module is not a single domain.

Key words: domain, fibronectin, module, polycystic kidney disease, polycystin-1 (PC1), receptor of egg jelly (REJ).

INTRODUCTION

A continuous segment of protein sequence that shows a high degree of sequence similarity in a range of different proteins is usually called a module [1,2]. This definition does not carry an explicit link to the structure into which the module might fold. Yet it is usually assumed that a module is a domain, i.e. that it is able to fold autonomously into a well-defined structure and that it cannot be cut down any further without losing its ability to fold properly [3,4]. For most modules this is the case, therefore the increased availability of newly sequenced proteins and the analysis of their module organization has given a significant boost to structural biology. The ability to cut large proteins down to their constituent modules greatly facilitated their structural and functional characterization. The automatic annotation of protein genes makes extensive use of established consensus sequences of modules even in cases where no experimental data have confirmed the relationship of modules and domains. Such information is extensively used as the base for numerous experimental studies of proteins where modules are mutated, added, swapped or deleted in the assumption that they are folded autonomously and make a defined contribution to the overall function of the protein.

In most cases, the assumptions made in the annotations of protein sequences turn out to be true. In others, however, even where structures are known, as in the case of the C2 domain fold [5], annotations may give a wrong estimate of the true size of the domain, resulting in a range of inconclusive experimental results. On the other hand, in the absence of any detailed information, large stretches of highly similar sequence are assigned as a module and thus classified as domain simply because they occur in a number of different proteins. One of such examples is the REJ module, which comprises a large portion of the extracellular region of a number of vertebrate cell-surface proteins. Its name derives from the protein in which this module was described for the first time, the receptor of egg jelly protein [6]. This module has a size of ~900 amino acids with no obvious homologues in sequence databases. It is found in suREJ (sperm receptor for egg jelly), PKDREJ (polycystic kidney disease and receptor for egg jelly-related protein), a number of uncharacterized proteins from genomic sequencing projects [7,8] and PC1 (polycystin-1). The latter is of specific medical interest because mutations in its gene, PKD1, are the main cause for ADPKD (autosomal dominant polycystic kidney disease) for which there is currently no cure [9]. ADPKD-related mutations are spread evenly throughout the entire PKD1 gene. At present, the only disease caused by mutations in PKD1 is ADPKD through the loss-of-function of PC1. REJ modules usually occur in the vicinity of the GPS (G-protein-coupled receptor proteolytic site) domain which contains an autoproteolytic motif [7,10]. Autoproteolysis is essential for full functionality of PC1 [11] and takes place after N-glycosylation of the protein [12–14]. Several mutations in the REJ module that cause ADPKD interfere with autoproteolysis [15,16] (Figure 1), suggesting an important function for the REJ module. Interestingly, in the first description of the gene for PC1 (PKD1) [17], there was no mention of the REJ module. Instead, it was proposed that the corresponding region should...
contain four FNIII (fibronectin type III) domains. This suggestion was subsequently dismissed after an unsuccessful bioinformatics screen of canonical FNIII domains [6] and the region was instead classified as a new type of module called REJ, named after the first gene in which it was identified. All of the subsequent literature on PC1 followed this definition and the FNIII domains were virtually forgotten about until recent AFM (atomic force microscopy) work on fragments of the extracellular portion of PC1 suggested the existence of smaller domains within the REJ module [18] with an unfolding pattern expected for FNIII domains. This led to a re-examination of the sequence of the REJ module by more advanced computational methods which confirmed the earliest suggestion of the presence of FNIII domains in PC1. To probe the combined evidence of sequence analysis and AFM data, we set out to perform an experimental analysis of the properties of the predicted FNIII domains. A reliable blueprint for the REJ module-containing proteins is essential for an understanding of their function, especially in the case of PC1, where this region of the protein harbours numerous point mutations involved in ADPKD (Figure 1).

MATERIALS AND METHODS

Sequence analysis

Four putative FNIII domains were tentatively identified in the REJ module of human protein PKD1 (Swiss-Prot [19] entry P98161, REJ module: residues 2146–2833; putative FNIII domains: residues 2155–2254, 2282–2361, 2392–2463 and 2485–2573). A total of 40 PDB structure fragments were selected from the SCOP FNIII domain family, with each subfamily with at least one representative structure. The structures with two or more consecutive FNIII domains were preferred in the selection. These 40 FNIII domain structures were superposed with the MAMMOTH-mult webserver [20] to build structural alignments of their sequences. Similarly, 40 PDB structure fragments were selected from the SCOP Ig I-set domain family, and superposed with MAMMOTH-mult. HMMs (hidden Markov models) [21] were constructed from these two MAMMOTH structural alignments by HMMER2.3. The four potential REJ module FNIII sequences were then aligned to the 40 SCOP FNIII structure sequences and Ig I-set structure sequences based on their HMMs by HMMER2.3 respectively.

Cloning and protein expression

All constructs for the FNIII domains were cloned using the In-Fusion method (Clontech) [22] into pLEICS-03 (Protein Expression Laboratory, University of Leicester). The constructs are expressed as a fusion protein with the sequence MHHIIHHHSGVLGENTLYFQSM, containing a His, tag and a TEV (tobacco etch virus) site N-terminally attached, which adds 23 residues and 2.7 kDa to each domain. After TEV digestion, the last two residues, Ser-Met, remain. For protein expression in inclusion bodies, constructs were transformed into Escherichia coli BL21* cells (Invitrogen). Bacterial cells were grown at 37°C and expression was induced with 0.5 mM IPTG (isopropyl β-D-thiogalactopyranoside) (Melford Laboratories) at a D600 of 0.8 for 4 h. Harvested cells were resuspended in wash buffer (20 mM phosphate buffer, pH 7.5, 500 mM NaCl, 1 mM 2-mercaptoethanol and 0.02% sodium azide) and opened using three cycles of French press at 6900 kPa. Cell debris was centrifuged at 5000 rev./min for 20 min in a Beckman JA30.5 rotor. At this speed, essentially only inclusion bodies are pelleted. The inclusion body pellet was separated and resuspended twice in wash buffer followed by centrifugation each time as before.
A third wash of the pellet was performed with wash buffer with additional 1 M urea. In this way, the amount of contaminating proteins is significantly reduced. The protein was then extracted from the inclusion bodies using wash buffer with 8 M urea for 2 h at room temperature (25 °C). The remaining insoluble debris was removed by centrifugation in a Beckman JA30.50 rotor at 15000 rev./min for 1 h. The supernatant was loaded on to a gravity flow column (empty PD10, GE Healthcare) filled with 2 ml of FF6 (Fast Flow HiTrap-binding resin) (GE Healthcare). The column was washed with 30 ml of wash buffer with 8 M urea, after which the bound protein was eluted with elution buffer (wash buffer with 500 mM imidazole and 8 M urea). The purity of protein samples was checked on SDS/PAGE 4–12% gradient gels (NuPAGE®, Invitrogen). Protein concentration was measured by absorption at 280 nm in a dual-beam UV-visible photometer with the respective buffer as a blank. For refolding, the protein concentration was adjusted to 5 mg/ml. Protein solution in elution buffer (250 μl) was then mixed with 4.5 ml of refolding buffer (50 mM Tris/HCl, pH 8.0) to which a l volume of NVoy (Expedeon) stock at a concentration of 25 mg/ml was added. The refolding reaction was left overnight at room temperature. The following day, an aliquot was taken before the reaction mixture was centrifuged at 4000 g for 30 min in a cooled Beckman benchtop centrifuge to remove precipitated protein. Another aliquot was taken of the supernatant afterwards. Both aliquots were analysed on SDS/PAGE 4–12% gradient gels (NuPAGE®, Invitrogen). NVoy polymer was removed following the manufacturer’s instructions for some samples. For expression of soluble protein, the constructs are transformed into ArcticExpress RIL cells (Stratagene) which contain the chaperonin system Cpn60/10 from Oleispira antarctica [23] for efficient protein folding at low temperature. After growth to a D[0.8] of ~0.8 at 37 °C, the temperature was lowered to 13 °C and expression was induced with 0.25 mM IPTG overnight. Cells were opened by French press followed by centrifugation at 18000 rev./min for 90 min in a Beckman JA30.50 rotor. The supernatant was then applied to a FF6 column and purified as described above, except without urea. To remove the His6 tag, 20 units of TEV protease were added per mg of protein to the soluble fraction which was then dialysed extensively against wash buffer to remove imidazole, usually 10–20 ml of solution three times against 1 litre of buffer. The solution was applied to the FF6 column as described above to remove the cleaved tag, TEV protease and remaining uncleaved protein. The flowthrough and wash fractions (10 ml) were checked by SDS/PAGE, pooled and dialysed as described above against measurement buffer [20 mM sodium phosphate, pH 7.5, 50 mM NaCl, 2 mM DTT (dithiothreitol) and 0.02 % sodium azide]. If required, the protein was polishes on a preparative gel-filtration column (HiLoad 16/60 Sephadex 75; GE Healthcare). After checking the concentration, the protein was concentrated in PES (polyethersulfone) VivaSpin20 concentrators with a 3 kDa molecular-mass cut-off.

AUC (analytical ultracentrifugation)

All AUC experiments were carried out on a Beckman XL-A analytical ultracentrifuge. Sedimentation equilibrium was attained at 18000 and 25000 rev./min in standard steel AUC cell using quartz windows and a six-channel centrepiece. Monomer molecular masses and partial specific volumes were calculated from the amino acid sequence using the program SEDNTERP [24]; these were determined to be 15721 Da and 0.7261 g/ml respectively. Data were processed using the programs SEDFIT and SEDPHAT [25,26] and fitted to single species.

CD spectroscopy

CD spectra were recorded on a Jasco J700 spectropolarimeter fitted with a Peltier temperature-control system. Spectra were recorded in rectangular quartz cuvettes (Starna) with 0.1 or 1 mm pathlength. A total of 20 scans were accumulated for one spectrum with a bandwidth of 2 nm, a slit width of 1 nm, one point per nm and 2 s averaging at each point. Samples of domains 1 and 2 were measured at protein concentrations from 20 to 100 μM in measurement buffer. Post-acquisition spectra were calibrated to molar ellipticity. Secondary-structure content was extracted using a home-written Mathematica macro by fitting the experimental spectrum to a synthetic spectrum made up of standard spectra for random coil, α-helix and β-sheet using a conjugate gradient minimizer. Thermal denaturation of the domains was monitored at a single wavelength of 214 nm using a temperature gradient of 1 °C/min from 5 to 90 °C. Data were recorded at one point per 1 °C. At each point, the CD signal was averaged for 1 s. The unfolding curve was fitted to a two-state unfolding equation in a home-written Mathematica macro which optimized the melting temperature and the slope at unfolding, while the initial and final slopes of the curve were optimized manually.

NMR spectroscopy

Spectra were recorded on a Bruker Avance 800 MHz spectrometer fitted with a cryoprobe at sample concentrations from 10 to 200 μM in 20 mM Tris/HCl or phosphate buffers, at pH values from 7.0 to 8.0, containing 50 mM NaCl, 2 mM DTT and 0.02 % sodium azide at temperatures of 25 and 30 °C. Water suppression in all spectra was achieved by WATERGATE with the offset on the water. The one-dimensional experiments were recorded with 256 scans and the two-dimensional HSQC (heteronuclear single-quantum coherence) with 128 scans. All spectra were recorded and processed with Topspin, version 2.1 (Bruker). One-dimensional spectra were apodized by exponential multiplication with a 4 Hz linewidth and zero-filled from 8192 to 16384 points before Fourier transformation followed by a standard baseline correction to remove offset effects. The HSQC experiment was processed by zero-filling F2 from 2048 to 4096 and F1 from 256 to 2048 points followed by apodization using a squared sine function shifted by π/2 in both dimensions before Fourier transformation that included an attenuation of the water signal by convolution. Points 2049–4096 in F2 were removed followed by an automatic polynomial baseline correction in F1 and F2. The HSQC spectrum was imported into CCPN analysis for peak picking, which was performed using the default parameters after manually optimizing the peak picking threshold.

RESULTS

Sequence analysis

The original description of the sequence of PC1 suggested the presence of four FNIII domains [17]. However, the sequence analysis was not complete, because additional domains such as the WSC domain, close to the N-terminus and the membrane-proximal GPS domain and PLAT/LH2 domain [27,28], were additionally identified later. The assignment of domains was significantly revised [6], leading to the introduction of the REJ module in place of the FNIII domains originally suggested (Figure 1). We followed up the original domain analysis with the aim of using newer methodologies not only to ascertain the presence of FNIII domains in the REJ module, but also to allow us to distinguish these from other potential β-strand-rich domains such as the very closely related Ig fold.
FNIII and Ig domains are structurally similar topologies composed of seven-strand β-sandwiches arranged in two sheets [29,30]. Structural alignments of 40 SCOP FNIII and Ig domain structures separately provide two sets of sequence conservation patterns to help in the classification of the four domain sequences from the REJ module as FNIII or Ig domains. These conservation patterns roughly correspond to the regions of the seven β-strands, which are labelled on FNIII and Ig modules in their alignments with the four PC1 sequences (Figure 2). These boundary regions were based on the assignments for the FNIII domains [30] (F8 in Figure 3 of [30] equal to 1I1f_1236–1326_A in our alignment in Figure 2), and of [31] for the Ig domains (1nc1_A and 1ncu_A in our alignment equal to TNM in Figure 3 of [31]).

The sequence conservation patterns in the strand regions are well maintained in the four PC1 sequences for strands A, E and F of FNIII modules, and do not completely match the other strands of FNIII modules. In contrast, the conservation patterns presented in the Ig structural alignments can only be incompletely observed in the regions of strands B, E and F and hardly observed in the region of other strands for the Ig module. As first observed [30], we noticed the conservation of a tryptophan residue in strand B; in addition, a tyrosine residue is strongly conserved in strand E of the FNIII modules and in strand F of the Ig modules is well aligned between the four PC1 sequences and the FNIII modules of strand E. The alignment of this region of the PC1 sequences and the Ig modules in strand F is more fuzzy (Figure 2A). On the basis of all of these observations, we can conclude that the four sequences from the REJ module are closer in evolution to FNIII modules rather than to Ig modules.

**Protein expression**

A range of expression constructs was designed for the four predicted FNIII domains (Figure 1) as shown in Figure 2 to cover the core domains plus parts of the linker sequences because of uncertainty about the precise location of the N- and C-termini. A selection of constructs and expression results is summarized in Table 1. Essentially, all constructs were expressed in inclusion bodies at 37°C in BL21* cells which could not be improved by reducing the IPTG concentration at induction from 0.75 to 0.1 mM and lowering the induction temperature to 20 and 15°C. Soluble protein for domains 1 and 2 was obtained by expression of the and lowering the induction temperature to 20 and 15°C reducing the IPTG concentration at induction from 0.75 to 0.1 mM 75 kDa corresponding to a soluble aggregate of at least six volume. This suggests an apparent molecular mass greater than 10 and 20 kDa, domain 1 appeared close to the exclusion column corresponding to a protein with a molecular mass between 20 and 120 kDa. Domain 2 appeared at an elution volume of the preparative column was successful for domain 2. Domain 1, however, could not be purified further using gel filtration (Figure 3C). Whereas domain 2 appeared at an elution volume of the preparative column corresponding to a protein with a molecular mass between 10 and 20 kDa, domain 1 appeared close to the exclusion volume. This suggests an apparent molecular mass greater than 75 kDa corresponding to a soluble aggregate of at least six molecules. As a result, we used refolded domains 1 and 2 as well as solubly expressed domain 2 for all of the biophysical experiments.

**CD spectroscopy**

CD spectra for domains 1 and 2 show the typical appearance of β-sheet proteins with a broad minimum between 210 and 220 nm (Figure 4) regardless of the method of production. Using a home-written Mathematica macro, the secondary-structure content of the domains according to these spectra was estimated to be approximately 62% β-sheet, 7% α-helix and 31% disordered, virtually identical for refolded and natively expressed domains. Both domains are thus assembled predominantly of β-sheet structure. CD spectroscopy was also used to measure the melting temperature by monitoring the CD signal at 214 nm over a range of temperatures from 5 to 90°C. The melting curves of both refolded domains showed little change from 5 to ~55°C from where the percentage of folded protein dropped within a short temperature interval from ~80–90% to less than 20%. The data measured were fitted to a two-state unfolding equilibrium using Mathematica, leading to melting temperatures of approximately 66°C for both, without any indication of significant deviation from the simple two-state model (see errors, lower panels of Figure 5). Interestingly, the natively expressed domain 2 showed hardly any sign of unfolding up to 90°C. In contrast, the intensity of the CD signal even increased from 20 to 40°C. As a result, a fit was not possible (Figure 5C). As an alternative, chemical denaturation with urea was performed using tryptophan fluorescence as a readout. A blueshift of approximately 12 nm from the lowest to the highest urea concentration was observed. This allowed the determination of the free energy of unfolding as 3.2 kcal/mol (1 kcal = 4.184 kJ) and the half maximum urea concentration as 4.4 M.

**Oligomeric state of domain 2**

Sedimentation equilibrium measurements at two velocities (Figure 6A) determined the molecular mass of refolded domain 2 in solution to be 15.2 kDa with a 68% confidence limit of 11.8–17.2 kDa. This is close to the calculated value of 15.7 kDa for a monomer with the His6 tag attached. Other more complex models such as monomer/dimer equilibrium did not improve the fit, and therefore domain 2 was judged to be monomeric under the conditions of the standard measurement buffer. AUC analysis of domain 1 under identical conditions did not lead to interpretable results, suggesting the presence of several species, presumably because of aggregation. The monomeric state of domain 2 was supported further by analytical gel filtration of natively expressed protein after removal of the His6 tag (Figure 6B). An elution at 12.6 ml corresponds to an apparent molecular mass of 14 kDa.

**NMR spectroscopy**

One-dimensional spectra were recorded at room temperature for refolded domains 1 (Figure 7A) and 2 (Figure 7B) and solubly expressed domain 2 (Figure 7C). The extreme high-field and low-field shifted regions are shown. The spectrum of domain 1 showed a few peaks around 0 p.p.m. in the high-field region and a good spread of peaks in the low-field region. The peaks were relatively broad for a protein with a molecular mass under 20 kDa and suggests that the protein is folded, but might aggregate. The spectra of both versions of domain 2 were of excellent quality. In the high-field region, the peaks were very sharp and very widely spread out up to ~1.0 p.p.m. Similarly, in the low-field region, a large number of well-dispersed sharp peaks are seen. The large
The four putative domains from PC1 are labelled FNIII1–FNIII4. All other sequences are taken from structures available from the PDB. All of these are labelled by their PDB accession number, beginning and end of the domain in the case of multidomain proteins and the molecule from which the sequence was taken. Expected $\beta$-strands for both folds are indicated by black boxes around the alignment which are labelled above. Sequence conservation is indicated by colouring of residues (green, hydrophobic; magenta, polar; orange, proline).

Figure 2  Sequence alignment of the predicted domains in the REJ module to a set of sequences representative of the FNIII fold (A) and the Ig fold (B)

The Authors Journal compilation © 2011 Biochemical Society
Table 1  FNIII domain constructs used in the present study and the results of their bacterial expression

FNIII domain constructs are as shown in Figure 3(A). Note that for constructs that did not express as soluble protein, the protein yield refers to soluble protein obtained after NvOy-assisted refolding. The start and end positions are the first and last residues in full-length human PC1. The bacterial expression host was either ArcticExpress (AE) or BL21* (Star).

<table>
<thead>
<tr>
<th>Domain</th>
<th>Start position</th>
<th>End position</th>
<th>Number of amino acids</th>
<th>Molecular mass (kDa)</th>
<th>$\varepsilon$ (M$^{-1}$·cm$^{-1}$)</th>
<th>Bacterial expression host</th>
<th>Is the expressed protein soluble?</th>
<th>Yield (mg of purified soluble protein in 1 litre of LB culture)</th>
<th>Solubility estimate after purification</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2152</td>
<td>2262</td>
<td>111</td>
<td>12.6</td>
<td>12.6</td>
<td>Star</td>
<td>No</td>
<td>50</td>
<td>Modest</td>
<td>–</td>
</tr>
<tr>
<td>1</td>
<td>2152</td>
<td>2262</td>
<td>111</td>
<td>12.6</td>
<td>12.6</td>
<td>AE</td>
<td>Yes</td>
<td>1</td>
<td>Modest</td>
<td>Soluble aggregates</td>
</tr>
<tr>
<td>1</td>
<td>2155</td>
<td>2254</td>
<td>100</td>
<td>11.4</td>
<td>11.4</td>
<td>Star</td>
<td>No</td>
<td>15</td>
<td>Poor</td>
<td>–</td>
</tr>
<tr>
<td>1</td>
<td>2155</td>
<td>2254</td>
<td>100</td>
<td>11.4</td>
<td>11.4</td>
<td>AE</td>
<td>No</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>2</td>
<td>2257</td>
<td>2369</td>
<td>113</td>
<td>12.3</td>
<td>20.97</td>
<td>Star</td>
<td>No</td>
<td>50</td>
<td>Good</td>
<td>–</td>
</tr>
<tr>
<td>2</td>
<td>2257</td>
<td>2369</td>
<td>113</td>
<td>12.3</td>
<td>20.97</td>
<td>AE</td>
<td>Yes</td>
<td>2</td>
<td>Good</td>
<td>–</td>
</tr>
<tr>
<td>2</td>
<td>2257</td>
<td>2374</td>
<td>118</td>
<td>12.9</td>
<td>20.97</td>
<td>Star</td>
<td>No</td>
<td>45</td>
<td>Modest</td>
<td>–</td>
</tr>
<tr>
<td>2</td>
<td>2257</td>
<td>2374</td>
<td>118</td>
<td>12.9</td>
<td>20.97</td>
<td>AE</td>
<td>Yes</td>
<td>1</td>
<td>Modest</td>
<td>Precipitated after short time</td>
</tr>
<tr>
<td>2</td>
<td>2282</td>
<td>2361</td>
<td>80</td>
<td>8.8</td>
<td>13.980</td>
<td>Star</td>
<td>No</td>
<td>–2</td>
<td>Very poor</td>
<td>–</td>
</tr>
<tr>
<td>2</td>
<td>2282</td>
<td>2361</td>
<td>80</td>
<td>8.8</td>
<td>13.980</td>
<td>AE</td>
<td>No</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>3</td>
<td>2371</td>
<td>2474</td>
<td>104</td>
<td>11.2</td>
<td>11.460</td>
<td>Star</td>
<td>No</td>
<td>20</td>
<td>Modest</td>
<td>Degradation</td>
</tr>
<tr>
<td>3</td>
<td>2374</td>
<td>2474</td>
<td>101</td>
<td>10.9</td>
<td>11.460</td>
<td>Star</td>
<td>No</td>
<td>30</td>
<td>Modest</td>
<td>Degradation</td>
</tr>
<tr>
<td>3</td>
<td>2374</td>
<td>2474</td>
<td>101</td>
<td>10.9</td>
<td>11.460</td>
<td>AE</td>
<td>No</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>3</td>
<td>2392</td>
<td>2463</td>
<td>76</td>
<td>8.2</td>
<td>6.990</td>
<td>Star</td>
<td>No</td>
<td>10</td>
<td>Poor</td>
<td>Degradation</td>
</tr>
<tr>
<td>4</td>
<td>2468</td>
<td>2577</td>
<td>110</td>
<td>11.8</td>
<td>9.970</td>
<td>Star</td>
<td>No</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>4</td>
<td>2477</td>
<td>2577</td>
<td>101</td>
<td>10.9</td>
<td>9.970</td>
<td>Star</td>
<td>No</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>4</td>
<td>2477</td>
<td>2577</td>
<td>101</td>
<td>10.9</td>
<td>9.970</td>
<td>AE</td>
<td>No</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>4</td>
<td>2485</td>
<td>2573</td>
<td>93</td>
<td>10</td>
<td>9.970</td>
<td>Star</td>
<td>No</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

Figure 3  Overview of expression constructs

(A) Overview of expression constructs. Shown for all domains are the various constructs that were created for expression trials in bacteria. The shaded box indicates the extent of the domain definition which we take to start two amino acids before the first residue of the first $\beta$-strand and to end two residues after the last amino acid of the last $\beta$-strand as shown in Figure 2. A few amino acids are shown at the start and the end of the box to aid orientation. Expression was tested for each domain with two constructs: one as indicated by the shaded box, the other indicated by the markers at the end points. The only variation exists for domain 2 where the new intermediate-length construct is indicated that is expressed solubly. (B) Refolding of domain 2. Shown is the purified protein before and after refolding in refolding buffer with and without NvOy polymer. Soluble (S) and insoluble (P) fractions are shown separated. M, molecular-mass markers (sizes given in kDa). (C) Preparative gel-filtration purification of domains 1 and 2 solubly expressed. Elution fractions of nickel-affinity purifications of both domains were loaded on to a Superdex 75 16/60 preparative column. Domain 2 emerges roughly in agreement with being a monomer, whereas domain 1 appears close to the exclusion volume, suggesting a heavily aggregated yet well-soluble state.

© The Authors Journal compilation © 2011 Biochemical Society
Figure 4 CD spectroscopy of predicted FNIII domains
(A) Domain REJ-1 (residues 2152–2262), refolded in NVoy. (B) Domain REJ-2 (residues 2257–2374), refolded in NVoy. (C) Domain REJ-2 (residues 2257–2369) expressed as soluble protein. All spectra were recorded at 5°C.

Figure 5 Stability of refolded and natively expressed FNIII domains
(A) Thermal denaturation from 5 to 90°C monitored by measuring the CD signal at 215 nm of domain 1, refolded in NVoy. (B) As in (A) for domain 2, refolded in NVoy. (C) As in (B), but with domain 2 expressed as soluble protein. Note that in this case, fitting was not possible because the curve did not have the expected shape for a denaturation. As a result, the y-axis is shown as molar ellipticity, not proportion folded. (D) Urea-denaturation curve of natively expressed domain 2. Experimental data are shown in the upper panels as black filled squares. A curve calculated using the fitting parameters is shown as a continuous line in the upper panels where fitting was possible. Fitting errors for each experimental point are shown in the lower panels.

Figure 6 Oligomeric state of REJ domain 2
(A) Sedimentation equilibrium AUC results at 18,000 rev./min (upper trace) and 25,000 rev./min (lower trace) of refolded protein in NVoy. For clarity, only one loading concentration has been shown. The determined molecular mass was 15.2 kDa, close to the expected monomeric mass of 15.7 kDa. Fits were determined globally using six datasets using the program SEDPHAT [26]. (B) Analytical gel filtration of domain 2 expressed in the soluble form, showing one peak at 12.6 ml which corresponds to a molecular mass of 15 kDa.

The excellent quality of the one-dimensional spectrum of domain 2 suggested that this domain might be best suited for the determination of the three-dimensional structure. To explore this further, an 15N-labelled sample was produced to record a two-dimensional 1H-15N-HSQC experiment (Figure 8). This two-dimensional spectrum was of equally excellent quality in line with the one-dimensional spectra. Its appearance shows the extensive dispersion of cross-peaks spreading them across most of the available space which is typical of proteins consisting mainly of β-sheet secondary structure. Automatic peak picking in CCPN analysis [32] gives a total of 116 peaks, excluding side chains, which is very close to the 121 peaks expected for domain 2 after removing the His6 tag.

DISCUSSION
The presence of FNIII-type domains in the extracellular part of PC1 was predicted when the sequence of the protein was presented for the first time [17]. Soon after, however, this interpretation was discounted because no strong signal evidence of a typical FNIII-related pattern was found in their analysis, and others opted in favour of classifying the entire region as the REJ module [6]. Since then, all analysis of functional features of PC1 has been based on this blueprint of PC1 (see Figure 1). A re-examination of the concept of the REJ module was prompted to us by the observation that AFM unfolding of extracellular fragments of PC1 comprising the REJ domain produced a number of unfolding peaks which disagrees with the idea that the REJ module is a single number of amide peaks was especially interesting given that the spectrum was recorded at a relatively high pH of 7.5.
co-operatively folded domain [18]. This result strongly suggested the presence of smaller domains such as the originally predicted FNIII domains.

A significant number of new FNIII structures have been added to the database since the earliest analysis, so that it was decided to first repeat the sequence analysis using a new sequence profile of the FNIII fold (Figure 2A). This was then compared with an alignment of the putative FNIII domains in PC1 against a profile for the Ig fold (Figure 2B). The \( \beta \)-strands in the FNIII profile are matched very well by the sequences of the putative FNIII domains in PC1. The only difference is seen for domain 3 in strand C. However, in two domains of the profile, the C-strand is also absent from the alignment, suggesting that this is probably not contributing as a strong signature for the FNIII fold. In contrast, the alignment with the Ig profile, the C-strand is absent from the alignment, suggesting that this is probably not contributing as a strong signature for the FNIII fold. The CD melting curve is very different, and unfolding appears to start at approximately 90°C. This is more than 20°C higher than observed for the refolded domain. Denaturation by urea clearly shows that this protein can be unfolded, that it happens in a co-operative manner and that it is indeed very stable, in good agreement with the failure to melt below 90°C (Figure 5). Also, a comparison of the NMR spectra suggests small but potentially significant differences in the way the protein folds when it is refolded or when it folds in cells. The overall peak pattern is very similar, but a close inspection reveals numerous variations such as the area around \(-1\) p.p.m. where the native protein has two peaks, whereas the refolded protein has only one. These differences cannot be explained by the different lengths of the constructs or by the absence or presence of the His\(_6\) tag. The low- and high-field ends of the one-dimensional NMR spectrum are dominated by resonances deeply buried in the hydrophobic core which is normally unaffected by changes at the N- or C-terminus. Combining these observations, it has to be concluded that, even though refolding appears to occur, it is not sufficient to produce a correctly folded protein. It is therefore necessary to use the low-yield low-temperature expression route to obtain a protein that has the correct structure.

The results for the natively expressed domain 2 are in good agreement with the predicted presence of FNIII domains: it is...
monomeric as is evident from analytical gel filtration as well as the good quality of the NMR spectra. The construct has a high $\beta$-sheet content, unfolds in a co-operative manner, is highly stable and produces an excellent two-dimensional $^{15}$N-HSQC spectrum (see Supplementary Figure S1 at http://www.BiochemJ.org/bj/435/bj4350651add.htm). The large number of high-field shifted peaks is slightly unusual for such a small domain, but can nevertheless be explained by the number of aromatic residues above the average (four phenylalanine, four tyrosine and three tryptophan; seven of these align with conserved hydrophobic positions of the FNIII fold in Figure 2).

It is quite intriguing to note that the other three predicted domains that ‘misbehaved’ appeared to do so independently of the way in which they were produced. In the case of domain 1, the refolded protein showed AUC data difficult to interpret and broad lines in the one-dimensional NMR spectrum suggesting at least partial aggregation, a tendency which was observed also for solubly expressed protein on the gel-filtration column (Figure 3). Domain 3 was expressed reasonably well in inclusion bodies, but was extremely unstable after refolding. The expression of domain 4 was so poor, even in inclusion bodies, that no effort was made to refold and investigate it further. The behaviour of these two domains does not appear to be caused by expression in bacteria because exactly the same pattern was observed in insect cells/baculovirus: domain 3 degraded quickly and domain 4 was hardly expressed at all (A.F. Oberhauser and F. Qian, unpublished work). These properties are thus inherent to these domains, so that further investigation would prove very challenging. The difficulty in producing these domains is not unusual for extracellular proteins and has been observed for a number of other domains and proteins [30,35].

In conclusion, we have provided extensive experimental evidence for the existence of at least two of the four predicted FNIII domains within the REJ module of PC1, suggesting that, in this case, it is not a single domain. We suggest avoiding the use of domain in this context and strictly refer to the REJ segment as a module. Further analysis of the remainder of the $\sim$600 amino acids in this part of PC1 might yield yet further domains, therefore this work should only be seen as the start of a re-evaluation of the domain architecture of PC1. Given the high sequence similarity that led to the creation of the term REJ domain, it is entirely conceivable that the domain organization of this region in PC1 might indeed be very similar in all REJ-module-containing proteins. The combination of the domains making up the REJ module is thus expected to be the same for all of the proteins in the REJ family. A precise understanding of the nature of the domains and their three-dimensional structure will facilitate the further investigation of these proteins to find out, e.g., why part of this module is required for autoproteolysis in the GPS domain and how point mutations in the REJ module (Figure 1) related to ADPKD interfere with this activity [15]. With the intent of improving our fundamental understanding of the relationship of sequence and structure in proteins, work such as this will also significantly contribute to the understanding of molecular mechanisms of inherited diseases.

ACKNOWLEDGEMENTS

We thank X. Yang and K. Hackmann for help with high-throughput cloning, Jennifer Moss for experiments with early constructs, F. Muskett for help with NMR experiments and K. Sidhu for computer support.

FUNDING

This work was supported by a project grant to M.P. and F.F. from Kidney Research UK [grant number RP22/2006], the National Institutes of Health [grant number DK 062199 to F.G.] and by the Johns Hopkins Polycystic Kidney Disease (PKD) Research and Clinical Core Center, National Institutes of Health P30 [grant number DK050868].

REFERENCES


© The Authors Journal compilation © 2011 Biochemical Society

AUTHOR CONTRIBUTION

Samantha Schröder carried out the experimental work, Xueping Qian carried out the sequence analysis, David Scott performed and analysed the analytical ultracentrifugation data, Franca Fraternali supervised the sequence analysis, planned the project and wrote the paper, Feng Qian planned the project, provided background information and wrote the paper, and Mark Pfuhl supervised the experimental work, planned the project and wrote the paper.

Received 5 November 2010/8 February 2011; accepted 14 February 2011
Published as BJ Immediate Publication 14 February 2011, doi:10.1042/BJ20101810

© The Authors Journal compilation © 2011 Biochemical Society
SUPPLEMENTARY ONLINE DATA

When a module is not a domain: the case of the REJ module and the redefinition of the architecture of polycystin-1

Samantha SCHRÖDER*, Franca FRATERNALI†, Xueping QUAN†, David SCOTT‡, Feng QIAN§ and Mark PFUHL*1

*Department of Biochemistry, University of Leicester, Leicester LE1 9HN, U.K., †Randall Division of Cell and Molecular Biophysics, King’s College London, Guy’s Campus, London SE1 1UL, U.K., ‡Department of Biosciences, University of Nottingham, Sutton Bonington, Leicestershire LE12 5RD, U.K., and §School of Medicine, The Johns Hopkins University, Baltimore, MD 21205, U.S.A.

Figure S1 Superposition of HSQC spectra of FNIII domain 2 in the presence (blue) and absence (red) of NVoy

Received 5 November 2010/8 February 2011; accepted 14 February 2011
Published as BJ Immediate Publication 14 February 2011, doi:10.1042/BJ20101810

1 To whom correspondence should be addressed at the present address: Cardiovascular Division and Randall Division, King’s College London, Guy’s Campus, London SE1 1UL, U.K. (email mark.pfuhl@kcl.ac.uk).