NMR analysis of the N-terminal SRCR domain of human CD5: engineering of a glycoprotein for superior characteristics in NMR experiments

Mark S.B.McAlister¹, Ben Davis², Mark Pfuhl¹ and Paul C.Driscoll¹,²,³

¹Department of Biochemistry and Molecular Biology, University College London, Gower Street, London WC1E 6BT and ²Ludwig Institute for Cancer Research, 91 Riding House Street, London, W1P 8BT, UK
³To whom correspondence should be addressed

CD5 is a type-I transmembrane glycoprotein found on thymocytes, T-cells and a subset of B-cells. The extracellular region consists of three domains belonging to the scavenger receptor cysteine-rich (SRCR) superfamily for which the three-dimensional polypeptide fold is as yet unknown. Glycosylated CD5 domain 1 (CD5d1) has been obtained by expression by secretion from both Chinese hamster ovary (CHO) cells and Pichia pastoris. Recombinant CD5d1 expressed in this manner was shown to be correctly folded by binding to anti-CD5 L17F12/Leu1 monoclonal antibody. Preliminary nuclear magnetic resonance (NMR) spectra obtained for CD5d1 (residues 1–118) had spectral dispersion typical of a folded protein, but otherwise of such poor quality that NMR structural studies were not feasible. The analysis of glycoproteins by NMR is frustrated by sample heterogeneity and poor spectral quality associated with glycan resonance overlap and the potential for increased line-widths due to the large hydrodynamic volume. In order to pursue NMR structural studies of CD5d1 it was necessary to optimize the quality of NMR spectra of CD5d1. A range of constructs of varying length and carbohydrate content were expressed in CHO cells and in P. pastoris. In addition the P. pastoris CD5d1 proved susceptible to N-glycan cleavage with endoglycosidase H. The protein products were characterised using size exclusion chromatography, NMR measurement of transional self-diffusion coefficients and two-dimensional 1H nuclear Overhauser effect spectroscopy experiments. Removal of an eight residue O-glycosylated C-terminal peptide, in particular, resulted in significant improvements in the quality of the CD5d1 NMR data, while retaining native protein structure. Two-dimensional heteronuclear NMR spectroscopy of nitrogen-15 isotope labelled deglycosylated CD5d1 (residues 1–110) prepared from P. pastoris suggests that this protein product is now amenable to solution structure determination.

Keywords: CD5/CHO/glycosylation/NMR/Pichia pastoris

Introduction

CD5 is a type-I transmembrane glycoprotein found on all T lymphocytes and a subset of B lymphocytes. Functional data suggest a role for CD5 in the control of lymphocyte responses (e.g. Freeman et al., 1990; Tarakhovsky et al., 1995; Bikah et al., 1996). The extracellular region comprises three domains belonging to group B of the scavenger receptor cysteine-rich (SRCR) superfamily (Resnick et al., 1994; Barclay et al., 1997), for which no three-dimensional fold has been obtained to date. Domains 1 and 2 are N-glycosylated, and are separated by an O-glycosylated linker domain (McAlister et al., 1998). The N-glycosylation site in CD5 domain-1 is conserved throughout all five species homologues sequenced to date, as well as in the ligand binding domain of CD6. CD5 shares highest sequence homology and tissue distribution with the CD6 antigen which also has three SRCR domains (Aruffo et al., 1997), and with a soluble protein that binds to macrophages SPS (Gebe et al., 1997). SRCR domains are found in various cell surface antigens of leukocytes and other cell types, although it is an ancient superfamily including proteins from primitive invertebrates such as the speract receptor of sea urchins. The first SRCR domain cloned was from the scavenger receptor, giving the superfamily its name. The scavenger receptor has several macromolecular ligands, but its single SRCR domain has not yet been directly implicated in ligand binding (Krieger and Herz, 1994). Other SRCR containing proteins found on leukocytes include: MARCO, a macrophage antigen containing a single SRCR domain that has a topology similar to scavenger receptor (Pearson, 1996); CD163 (M130) antigen, a type I glycoprotein found on macrophages containing nine SRCR domains (Law et al., 1993); and the WC1 antigen found on found on γδ T cells in cattle that contains 11 SRCR domains (Wijngaard et al., 1992). The SRCR superfamily can be divided into two groups (A and B) based on sequence alignments and in particular on the number of cysteines (Resnick et al., 1994). SRCR group A (domains with six cysteines) currently comprises eight members and group B (eight cysteines) eleven members. Support for the hypothesis that these homologous SRCR sequences represent autonomous folding units arises on the basis of sequence analysis, the location of intron/exon boundaries, and the existence of type-I and type-II scavenger receptors that differ only in the presence of a single putative SRCR sequence repeat (Krieger and Herz, 1994; Aruffo et al., 1997; Barclay et al., 1997).

At present there is a lack of convincing evidence for a molecular ligand for CD5 (Aruffo et al., 1997; Barclay et al., 1997). However the related antigen CD6 has been shown by mutagenesis studies to bind, through the membrane proximal SRCR domain, to the N-terminal domain of CD166 (Bowen et al., 1995; Whitney et al., 1995; Bodian et al., 1997). CD5 has a comparatively large cytoplasmic domain (94 residues) with regulatory intracellular tyrosine activation motifs (ITAMs) suggesting a role in signalling in response to interactions of the extracellular region with proteins at the surface of other cells or soluble proteins (Beyers et al., 1992). Indeed signalling through CD5 was recently reported to activate a pathway involving phosphatidylinositol 3-kinase, Vav and Rac1 in mature human T-lymphocytes (Gringhuis et al., 1998).

Glycoproteins are under-represented in structural databases in comparison to non-glycosylated proteins, although the vast majority of secreted and cell surface proteins are glycosylated.

© Oxford University Press
There are only 30 X-ray and four solution structures of intact glycoproteins currently present in the Brookhaven Protein Databank (Wyss and Wagner, 1996). This under-representation may be due to the practical difficulties in handling glycoproteins. The expression of recombinant glycosylated proteins requires the use of eukaryotic hosts, and the compositional and conformational heterogeneity of the carbohydrate structure can hinder structural studies. Overexpression in bacterial cells has been used to produce non-glycosylated forms of some glycoproteins (e.g. rat CD2 domain 1; Driscoll et al., 1991). This strategy is not appropriate for proteins that require the presence of glycosylation in order to obtain or maintain structural integrity (e.g. human CD2 domain 1, Wyss et al., 1995; Davis and van der Merwe, 1996), and it is difficult to predict a priori whether any individual glycan is important for a particular protein. The importance of accounting for glycosylation in the structural analysis of glycoproteins by NMR was highlighted in the study human CD2 domain-1, where it was found that the N-glycans were essential for the production of structurally intact protein (Wyss et al., 1995). The growing number of proteins for which carbohydrates have been shown to have some stabilizing influence suggests that this may describe a general role e.g. RNAaase B (Joao and Dwek, 1993; Rudd et al., 1994); the protease inhibitor PMP-C (Mer et al., 1996); human chorionic gonadotropin α-chain (van Zuylen et al., 1997). For reviews on structural and functional properties of sugars attached to glycoproteins see Dwek et al. (1995) and Wyss et al. (1996).

The routine application of modern heteronuclear NMR spectroscopy to glycoproteins is complicated by the practical and economic difficulties of producing samples labelled with stable isotopes (e.g. nitrogen-15 and carbon-13) in eukaryotic expression hosts, as well as the increase in molecular size compared with non-glycosylated forms of the protein. Depending upon the degree of flexibility of the protein–glycan linkages, the larger molecular size could cause an increase in the effective hydrodynamic volume (radius of gyration) leading to a larger rotational correlation time (\( \tau_c \)) and a corresponding increase in spectral line width. The heterogeneity and spectral overlap associated with N- and O-linked glycan carbohydrate resonances further complicates NMR studies of glycoproteins.

We chose to study the N-terminal domain of human CD5 (CD5d1) as an example of the SRCR domain superfamily and as an antigen for which there are functional data to suggest an important role in T cell responses. Recombinant constructs of CD5d1 were produced in Chinese hamster ovary (CHO) cells and in the methylotrophic yeast P. pastoris as an example of the SRCR domain superfamily and to elucidate previously (McAlister et al., submitted). The sites of N-linked and O-linked glycosylation are indicated. Protein constructs described in this paper include CD5d1 residues 1–118 (cCD5d1c) and residues 1–110 (cCD5d1, pCD5d1, c' and p' refer to expression in CHO cells and in P. pastoris respectively.

were expressed in Chinese hamster ovary (CHO) cells, using the pEE14 expression vector with the CHO K-1 cell line as described previously (Cockett et al., 1990; Davis et al., 1990). Both cCD5d1 samples were purified from 20 l of culture supernatant using L17F12/Leu1 monoclonal antibody immuno-affinity chromatography and gel filtration. The pCD5110 construct was also expressed in the methylotrophic yeast P. pastoris strain GS115, using the expression vector pPIC9K, with multivcopy transcription selection (Scorer et al., 1994). Cell culture and protein expression were performed by Invitrogen Inc., and protein purification procedures were as above. Partially deglycosylated samples of pCD5110 from P. pastoris were prepared by digestion with endoglycosidase H (endo H), which cleaves only high mannose glycans to leave a single N-acetyl glucosamine residue (GlcNAc) attached to the asparagine side chain. Endo H (NEB Inc.) digestions were performed in 20 mM phosphate pH 6 at room temperature for 1–3 h with around 1000 U/endo H per milligram of pCD5110. Protein concentrations were estimated using either inhibition ELISA with L17F12/Leu1 or measurement of absorbance at 280 nm. The extinction coefficient at 280 nm (28360 M\(^{-1}\) cm\(^{-1}\)) was calculated using the program PEPTIDESORT (GCC Inc.). In this paper, the glycosylation state of each construct is indicated in parenthesis after the construct name e.g. cCD5118(NO) (N, N-glycans; O, O-glycans; c, from CHO cells; p, from P. pastoris).

\(^{15}\)N-isotope labelling of pCD5110 in P. pastoris

For \(^{15}\)N-isotope labelling, cultures were grown in shake-flasks rather than high cell density fermentations as there was less evidence of proteolysis of pCD5110. Single colonies were selected from BMG agar plates and inoculated into 5 ml BMG starter culture in a 50 ml Falcon tube and incubated for 24 h (media were prepared as described by Invitrogen). The culture conditions were optimized for pCD5110 in the following manner. Aliquots of the starter cultures were used to inoculate 2 l flasks capped with four layers of muslin, containing 500 ml of medium: 0.2% (w/v) glycerol, 0.8% (w/v) sorbitol, 1.3% (w/v) YNB without ammonium sulphate or amino acids (Difco), 0.4 µg/ml biotin, 3 g/l \(^{15}\)N ammonium sulphate (CIL,}

---

**Fig. 1.** The amino acid sequence of human CD5 domain 1 (CD5d1). Dashed lines connecting cysteine residues indicate the disulfide pairing pattern elucidated previously (McAlister et al., submitted). The sites of N-linked and O-linked glycosylation are indicated. Protein constructs described in this paper include CD5d1 residues 1–118 (cCD5d1c) and residues 1–110 (cCD5d1, pCD5d1, c' and p' refer to expression in CHO cells and in P. pastoris respectively.
Inc.). Growth of P. pastoris on a mixture of sorbitol and methanol as the carbon source allows induction of recombinant protein expression throughout the growth of the culture and more rapid growth than with methanol alone (Sreekrishna et al., 1997), since sorbitol does not cause repression of the AOX1 gene. In contrast glycerol, which is conventionally used as carbon source during the growth phase, is a repressor of the AOX1 gene (Wegner and Harder, 1986) preventing recombinant protein expression. Sorbitol was therefore used to optimize yields of \(^{15}\text{N}\)-labelled pCD5\(_{110}\) and also to simplify the culture procedure, since it is no longer necessary to separate growth and protein expression phases of the culture. This procedure yielded more reproducible yields of pCD5\(_{110}\). Cultures were grown until a wet cell weight of around 10 g/l, before adding methanol to 0.5% (v/v). Cultures were then fed twice daily with methanol to 0.5% (v/v) for 48 h, before harvesting the cultures at a wet cell mass of around 40-50 g/l. Cultures were grown in an enriched oxygen environment to ensure efficient aeration (refreshed twice daily with pure O\(_2\)), in a shaking incubator that was compatible with enriched oxygen levels (Gallenkamp). The ammonium sulfate concentration was minimized to allow economic \(^{15}\text{N}\)-labelling of pCD5\(_{110}\). Levels of secreted pCD5\(_{110}\) were not limited under the culture conditions described, until the concentration of \(^{15}\text{NH}_4\text{SO}_4\) was below 0.05% (w/v). Ammonium sulfate concentrations below this value led to significant proteolytic degradation of pCD5\(_{110}\) during purification, probably due to the up-regulation of proteases under nitrogen starvation (Jones, 1991). The uniform \(^{15}\text{N}\)-labelling method described here yielded around 2-3 mg/l of pCD5\(_{110}\), similar to the yield per gram of biomass obtained in oxygenated unlabelled cultures performed in a fermenter. Purification of \(^{15}\text{N}\) labelled pCD5\(_{110}\) was as described for the unlabelled protein.

**Analytical size exclusion chromatography**

Analytical SEC experiments were performed using a BioSelect SEC 125-5 semi-analytical HPLC column (Biorad) controlled by a BioCAD Sprint (Perceptive Biosystems). The column was run in PBS (50 mM potassium phosphate pH 6.0, 140 mM NaCl) at 1.0 ml/min, using 20 µl sample volumes containing 100 µg of CD5\(_{118}\) samples. The column was calibrated with a range of non-glycosylated molecular weight standards to allow estimation of apparent molecular mass (for a review see Prentea, 1990).

**NMR measurements**

\(^1\text{H}\)NMR experiments were performed on either a Varian UnityPlus 600 or 600 MHz NMR spectrometer at 25°C in 100% D\(_2\)O for translational diffusion experiments or a 90% H\(_2\)O/10% D\(_2\)O mixture for NOESY spectra. Samples were 1 mM protein in 20 mM potassium phosphate buffer pH 6.0, with a sample volume of 600 µl. 2D \(^1\text{H}\) NOESY spectra were acquired using ‘jump-return’ water suppression (Driscoll et al., 1989) with a spectral width of 7400 Hz in both dimensions. A NOE mixing time of 100 ms was used. NOESY spectra were acquired with 2048 points in the direct dimension and 256 complex points in the indirect dimension. 64 transients were acquired for each \(\tau_n\) increment.

Translational self-diffusion coefficients (D\(_T\)) were determined using the pulsed field gradient longitudinal encode–decode (PFGLLED) pulse sequence (Altieri et al., 1995; Dingley et al., 1995). A diffusion delay of 40 ms between the encode and decode gradient sequences was used for all samples except cCD5\(_{118}\)(NO), where a delay of 80 ms was used. Encode/decode gradients were of 10 ms duration in all experiments. Values of D\(_T\) were determined from a least-squares fit of PFGLLED spectra intensity integrated between 2.5 and 0.5 p.p.m. with varying gradient power levels. A two-parameter model was used to fit the data, where I is the observed signal intensity for a given value of G, the magnitude of the applied gradient pulse, \(I_0\) the signal intensity with zero applied gradient, D\(_T\) the translational diffusion coefficient, \(\gamma\) the gyromagnetic ratio, \(\delta\) the duration of the encode/decode gradient pulses and \(\Delta\) the delay between the encode and decode pulses:

\[
I = I_0 \exp\left[-D(\gamma G)^2(\Delta-\delta/3)\right]
\]

The apparent molecular mass of a species, \(M^{\text{app}}\), is proportional to 1/D\(_T^3\) (Dingley et al., 1995). Values of 1/D\(_T^3\) were calculated, and normalized with respect to the value obtained for pCD5\(_{110}\) in order to avoid assumptions concerning sample viscosity and molecular shape (Dingley et al., 1995).

Two-dimensional \(^1\text{H},^{15}\text{N}\)-heteronuclear single quantum coherence spectra were recorded at 25°C and pH 4.7. Water flip-back pulses and gradient sensitivity enhancement procedures were incorporated into the pulse scheme according to Zhang et al. (1994).

**Results and discussion**

We have targeted human CD5d1 as an archetypal SRCR domain for structure determination, since sequence analysis suggests this may be an autonomously folding unit. So far, attempts to produce native CD5d1 by refolding of material expressed in *Escherichia coli* have proved unsuccessful. Expression of CD5d1 in mammalian and yeast hosts has been extensively investigated, since eukaryotic cells are competent at expression and secretion of multiple disulphide-bonded glycoproteins. Preliminary \(^1\text{H}\)NMR spectroscopic examination of a 118 residue construct expressed by secretion from Chinese hamster ovary (CHO) cells (cCD5\(_{118}\)) yielded spectra with unexpectedly broad resonances, suggesting the possibility of self-association of the molecules or elongation of the rotational correlation time due to a large radius of gyration, possibly effected by glycosylation present in this construct. This observation prompted a more systematic examination of the spectroscopic properties of a set of different CD5d1 protein products of differing protein and carbohydrate composition.

Chinese hamster ovary cells (CHO) were used to express two recombinant constructs of the N-terminal domain of CD5, residues 1–118 (cCD5\(_{118}\)) and 1–110 (cCD5\(_{110}\)) (Figure 1). The shorter 110 residue construct omits a C-terminal tail region which we have shown is a site of CD5-directed secretion. Further details of CD5\(_{110}\) protein expression,
disulphide bond mapping and glycosylation are described elsewhere (McAlister et al., 1998).

The apparent molecular masses ($M_r^{app}$) of the four different CD5d1 samples was estimated from the elution volumes in analytical SEC (Figure 2; Table 1). The elution volume of a molecular species on a SEC column is inversely proportional to the logarithm of the apparent molecular weight of the species (for a review see Preneta, 1990). $M_r^{app}$ can be therefore be estimated by comparison of elution volume with a range of standard non-glycosylated proteins of known molecular weight. The prefix 'apparent' is an important qualifier, since this qualitative analysis disregards the potential for differential density and shape effects. The irregular elution peak shapes obtained with these measurements indicates that the hydrodynamic size of the CD5d1 constructs is in the order:

cCD5$_{118}$(NO) > pCD5$_{110}$(N) = cCD5$_{110}$(N) > pCD5$_{110}$

Clearly this trend reflects both the size of the proteins in terms of amino acid residues and the degree of N- and O-linked glycosylation. We note that chemical cross-linking experiments failed to demonstrate the presence of self-associated CD5d1 that would be consistent with the formation of homo-oligomers (McAlister et al., unpublished data).

The one-dimensional $^1$H NMR spectrum of each of the four CD5d1 samples contained essentially identical patterns of amino acid resonances and glycan resonances, as shown in Figure 3. Disregarding differential density, solvation and molecular shape effects, the molecular mass of a species is proportional to $1/D_T^3$ (Dingley et al., 1995). Relative values of $1/D_T^3$ and $M_r^{app}$ were obtained by normalization to the value obtained for pCD5$_{110}$, the sample with the lowest apparent molecular mass (Table I). Inspection of Table I shows that consistent values of relative SEC $M_r^{app}$ and PFGLED $1/D_T^3$ were obtained with both techniques for the four CD5d1 samples investigated. The pattern of results obtained with these measurements indicates that the hydrodynamic size of the CD5d1 constructs is in the order:

cCD5$_{118}$(NO) > pCD5$_{110}$(N) = cCD5$_{110}$(N) > pCD5$_{110}$

for cCD5$_{118}$ in Figure 3. Disregarding differential density, solvation and molecular shape effects, the molecular mass of a species is proportional to $1/D_T^3$ (Dingley et al., 1995). Relative values of $1/D_T^3$ and $M_r^{app}$ were obtained by normalization to the value obtained for pCD5$_{110}$, the sample with the lowest apparent molecular mass (Table I). Inspection of Table I shows that consistent values of relative SEC $M_r^{app}$ and PFGLED $1/D_T^3$ were obtained with both techniques for the four CD5d1 samples investigated. The pattern of results obtained with these measurements indicates that the hydrodynamic size of the CD5d1 constructs is in the order:

cCD5$_{118}$(NO) > pCD5$_{110}$(N) = cCD5$_{110}$(N) > pCD5$_{110}$

Clearly this trend reflects both the size of the proteins in terms of amino acid residues and the degree of N- and O-linked glycosylation. We note that chemical cross-linking experiments failed to demonstrate the presence of self-associated CD5d1 that would be consistent with the formation of homo-oligomers (McAlister et al., unpublished data).

The one-dimensional $^1$H NMR spectrum of each of the four CD5d1 samples contained essentially identical patterns of chemical shift dispersion, indicating that all samples possess the same polypeptide fold. Two dimensional $^1$H NOESY spectra were acquired under the same conditions (Figure 4). The dispersion of cross-peaks is extremely similar for each construct. Small chemical shift differences are attributed to the different peptide and/or glycan chemical composition. The overall similarity of the spectra confirm that neither the presence of the O-glycosylated C-terminal residues or the bulk of the N-glycans are essential for maintenance of native structure, consistent with the observation that all four protein

---

**Table I.** Apparent molecular weights and translational self-diffusion coefficients of constructs of CD5d1

<table>
<thead>
<tr>
<th>Construct</th>
<th>SEC elution volumes (ml)</th>
<th>$M_r^{app}$ (kDa)</th>
<th>rel. $M_r$</th>
<th>$D_T^2$ (10$^{-10}$m$^2$/s)</th>
<th>rel. $D_T^3$ (see note c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cCD5$_{118}$(NO)</td>
<td>7.72</td>
<td>31.6</td>
<td>2.1</td>
<td>0.60 ± 0.01</td>
<td>2.8</td>
</tr>
<tr>
<td>pCD5$_{110}$(N)</td>
<td>8.43</td>
<td>22.9</td>
<td>1.5</td>
<td>0.74 ± 0.02</td>
<td>1.5</td>
</tr>
<tr>
<td>cCD5$_{110}$(N)</td>
<td>8.58</td>
<td>21.9</td>
<td>1.5</td>
<td>0.75 ± 0.03</td>
<td>1.5</td>
</tr>
<tr>
<td>pCD5$_{110}$</td>
<td>9.18</td>
<td>14.8</td>
<td>1.0</td>
<td>0.85 ± 0.02</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Notes:
- The prefixes ‘c’ and ‘p’ denote protein samples expressed from CHO cells and E. coli respectively; ‘(N)’ and ‘(O)’ refer to the presence of attached N- and O-glycans. pCD5$_{110}$ refers to the endo H-treated form of pCD5$_{110}$(N) which has a single GlcNAc residue attached at Asn92.
- The $M_r^{app}$ values were estimated by comparison of their SEC elution volumes to a standard curve of log(molecular mass) versus elution volume of protein standards: thyroglobulin, 670 kDa, 5.68 ml; immunoglobulin G, 158 kDa, 6.33 ml; ovalbumin, 44 kDa, 7.35 ml; myoglobin, 17 kDa, 9.03 ml; cytochrome c, 12.2 kDa, 9.72 ml; aprotinin, 6 kDa, 10.67 ml.
- Relative values of $M_r^{app}$ and $D_T^3$, respectively; ‘(N)’ and ‘(O)’ refer to the presence of attached N- and O-glycans.
- $D_T^2$ values were obtained by one-dimensional $^1$H PFGLED NMR as described in the Materials and methods section.
products are recognized by a conformation-dependent anti-CD5 monoclonal antibody. In spite of the overall similarity of the chemical shift dispersion of the spectra, it is clear that the quality of the spectrum varies across the set of proteins, with the cCD5<sub>118</sub>(NO) protein in particular yielding spectra with broader resonances and fewer discernible cross-peaks (Figure 4A). The endo H treated pCD5<sub>110</sub> yields a NOESY spectrum (Figure 4D) which is again somewhat superior (sharper resonances and more intense cross-peaks) than obtained for the forms with intact N-glycans pCD5<sub>110</sub>(N) and cCD5<sub>110</sub>(N) (Figure 4B and C).

The hydrodynamic measurements suggest that the 118 residue CD5d1 construct with both N- and O-linked glycans present behaves as a disproportionately large molecule. The presence of the extra eight amino acid residues and the O-linked glycans at the C-terminus of cCD5<sub>118</sub>(NO) caused an increase in apparent size in comparison to cCD5<sub>110</sub>(N) corresponding to an extra ~10 kDa mass. In contrast the removal of the N-linked glycan by endo H treatment for pCD5<sub>110</sub> leads to a reduction in apparent mass corresponding to ~7 kDa (Table I). The disproportionate effect of the O-glycosylated C-terminal tail is mirrored in the much poorer quality of the NMR spectra obtained with this construct (Figure 4A). O-Glycosylation is thought to rigidify peptides causing formation of extended structures, as has been found for mucins (Shogren et al., 1989) and leukosialin (Cyster et al., 1991). It may therefore be possible that the eight residue C-terminal tail of CD5d1, which contains three potential sites of O-glycosylation and three Pro residues contributes substantially to the radius of gyration of the molecule, in a manner that is not observed for more flexible terminal peptide appendages. Endo H trimming of pCD5<sub>110</sub>(N) results in a significant increase in elution volume in SEC experiments. The actual mass of the major N-glycan removed is predicted to be ~2 kDa, but carbohydrates are known to form extended structures in comparison to folded protein and again add disproportionately to

Fig. 4. The downfield region of NOESY spectra of (A) cCD5<sub>118</sub>(NO), (B) cCD5<sub>110</sub>(N), (C) pCD5<sub>110</sub>(N) and (D) pCD5<sub>110</sub>. All spectra were recorded at 25°C in 90% H<sub>2</sub>O, 10% D<sub>2</sub>O with a 100 ms NOE mixing time. Removal of the O-linked glycans (panels A and B) improves spectral quality and line-widths. Trimming of the N-linked glycans has limited effects on the line-widths (panels C and D).
the overall molecular size. SEC data indicate that the $M_\text{rapp}$ of pCD5$_{110}$(N) is 2–3 kDa larger than cCD5$_{110}$(N), possibly reflecting the different glycan composition.

The marked effect on the NMR spectral quality and line-widths on removing the O-glycosylated peptide (Figure 4A and B) is consistent with the significant decrease in $M_\text{rapp}$ observed by SEC and PFGLED. The effect of endo H trimming the N-glycans (Figure 4C and D) is less pronounced than the effect of removing the O-glycosylated peptide (Figure 4A and B), again consistent with SEC and PFGLED data. Following endo H treatment, NMR spectra of the pCD5$_{110}$ construct are similar to those obtained for pCD5$_{110}$(N), and are of sufficient quality to warrant further NMR studies of both the glycosylated and deglycosylated species. We note that removal of the N-linked glycans from pCD5$_{110}$(N) decreases the degree of cross peak overlap in the spectra of CD5d1, most obviously in the aliphatic region of the NOESY spectra (data not shown), and will therefore provide a useful method for simplification of the spectrum.

A culture protocol was developed to allow economic production of uniformly $^{15}\text{N}$ labelled samples from P.pastoris for study of pCD5$_{110}$ by $^{15}\text{N}$-$^{1}\text{H}$ heteronuclear NMR experiments. The conventional procedure for growth and expression of P.pastoris in minimal medium (Invitrogen, Inc.) was modified to optimize expression/secretion of pCD5$_{110}$ for economic $^{15}\text{N}$-labelling and to minimize adventitious proteolysis of the protein product during purification (see Material and methods). A two-dimensional heteronuclear single quantum coherence (HSQC) spectrum of $^{15}\text{N}$-labelled endo H treated pCD5$_{110}$ is shown in Figure 5. The spectrum reveals good $^{1}\text{H}$ and $^{15}\text{N}$ chemical shift dispersion consistent with a folded globular protein and provides a sound basis for further multi-dimensional heteronuclear NMR analysis towards the solution structure determination of this archetypal SRCR domain.

N-glycans in proteins produced in mammalian or insect cells are often refractory to facile enzymatic deglycosylation due to steric hindrance caused by core fucosylation of GlcNAc-1 (Maley et al., 1989). This characteristic prevents application of deglycosylation strategies to the NMR analysis of CD5d1 constructs produced in CHO cells. Mutant strains of CHO cells which produce glycoproteins with only high mannosyl glycans without core fucosylation are available (Davis et al., 1993) and high mannosyl glycans can be readily trimmed using endo H. However, stable isotopic labelling for heteronuclear NMR studies in CHO cells is only economically viable for proteins expressed at very high levels (Lustbader et al., 1996). In contrast, expression of glycoproteins in the methylotrophic yeast P.pastoris allows the potential for economic isotopic labelling with nitrogen-15 ammonium salts and carbon-13 sources glucose and methanol (Laroche et al., 1994; Thomsen et al., 1996; Wiles et al., 1997; Gronwald et al., 1998). Pichia pastoris usually performs N-deglycosylation at the native sites in recombinant mammalian proteins with the addition of only high mannosyl glycans (Grinna and Tschopp, 1989), and therefore allows facile trimming with endo H to leave a homogenous product with a single Asn-attached GlcNAc moiety. A single GlcNAc residue is often sufficient for the stabilization of proteins which require glycosylation for structural integrity (Wyss and Wagner, 1996) and which can not be produced in a folded form from bacteria. The N-terminal domain of the T- and B-cell antigen CD5 is an example of a disulphide-rich, glycoprotein domain which has not yet been successfully produced in folded form by bacterial expression. We have shown in this work that CD5 domain 1 can be successfully produced in eukaryotic expression systems and that optimization of the expression construct and enzymatic treatment of the attached N-glycan provides for a stably folded protein that is suitable for further structural studies by NMR spectroscopy. The homogeneity of the pCD5$_{110}$ sample should also provide a suitable product for crystallization trials. In addition, comparative NMR analysis of pCD5$_{110}$(N) before and after endo H treatment will allow study of the structural role of the bulk of the N-glycan, while comparative NMR analysis of pCD5$_{110}$(N) with cCD5$_{110}$(N) may allow elucidation of the structural properties of CD5d1 with different N-glycan structures attached.

Acknowledgements

The authors are grateful to Dr Helen R. Mott for help in the early stages of the project. M.S.B.M has been supported by the Human Frontiers Science Program Organisation and the BBSRC. B.D. is supported by the Ludwig Institute for Cancer Research. M.F. is supported by an long-term EMBO Fellowship. P.C.D. is a Royal Society University Research Fellow.

References
