Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

A branched N-linked glycan at atomic resolution in the 1.12 Å structure of rhamnogalacturonan acetylesterase

Anne Mølgaard and Sine Larsen*

Centre for Crystallographic Studies, University of Copenhagen, Universitetsparken 5, DK-2100 Copenhagen, Denmark

Correspondence e-mail: sine@ccs.ki.ku.dk

Received 17 September 2001
Accepted 31 October 2001

PDB Reference: rhamnogalacturonan acetylesterase, 1k7c.

The crystal structure of the glycoprotein rhamnogalacturonan acetylesterase from *Aspergillus aculeatus* has been refined to a resolution of 1.12 Å using synchrotron data collected at 263 K. Both of the two putative N-glycosylation sites at Asn104 and Asn182 are glycosylated and, owing to crystal contacts, the glycan structure at Asn182 is exceptionally well defined in the electron-density maps, showing the six-carbohydrate structure Manα1-6(Manα1-3)Manα1-6Manβ1-4GlcNAcβ1-4GlcNAcβ-Asn182. Equivalent carbohydrate residues were restrained to have similar geometries, but were refined without target values. The refined bond lengths and angles were compared with the values obtained from small-molecule studies that form the basis for the dictionaries used for glycoprotein refinement.

1. Introduction

Protein glycosylation is one of the most prevalent post-translational modifications of eukaryotic proteins synthesized in the endoplasmatic reticulum. The carbohydrate can be attached to either the O atoms of hydroxyl groups (O-linked) or the N atom of Asn side chains (N-linked). In contrast to the linear nature of proteins and nucleic acids, glycans have the possibility of branching. Furthermore, the individual carbohydrate residues can be in furanose or pyranose forms and can adopt different anomeric configurations, so that the number of isomeric structures is vast (Laine, 1994). When it is considered that each given post-translational modification of a protein may have a unique functional role, it is clear that the genomic sequence does not reflect the functional complexity of proteins (Vosseller *et al.*, 2001). This has made the role of carbohydrates as hardware in biological information transfer and storage a subject of increasing interest (Feizi, 2000; Gabius, 2000; Solis *et al.*, 2001).

Several important biological functions have been attributed to oligosaccharides. They have been shown to affect the physical properties of the protein to which they are bound; they can also provide lectin-recognition sites and shield the backbone against protease degradation. They have also been shown to play a role in transport and secretion and to affect protein folding in the endoplasmatic reticulum. Recent studies have indicated that the glycan structure can exert a stabilizing effect on the structure of large parts of the protein backbone spatially far away from the glycan attachment site (Wormald & Dwek, 1999) and O-glycosylation may also be a regulatory modification analogous to phosphorylation (Vosseller *et al.*, 2001). The structural diversity of the covalently attached carbohydrates of glycoproteins contributes to the difficulties

in assigning a unique biological role to glycoconjugates. This is expressed well by Varki (1993) in his extensive review on the subject. It is concluded there that all the advanced theories are correct, though exceptions can be found in each case.

From an analysis of the SWISS-PROT database (Bairoch & Apweiler, 1999), Apweiler *et al.* (1999) estimated that more than half of all proteins in nature are glycosylated, taking into account that not all N*X*S/T consensus sequences are occupied and that some proteins (∼10%) are solely O-glycosylated. This high frequency of glycoproteins is not reflected in the structures in the Protein Data Bank (PDB; Berman *et al.*, 2000), where only 3.6% of the protein structures contain covalently linked glycosylation. The notorious difficulty in crystallizing glycoproteins can explain some of this discrepancy, which can be attributed to heterogeneity with respect to chemical composition and to the conformational flexibility of the glycans. For these reasons, deglycosylated proteins have become the favoured targets for structure determination.

The number of glycoprotein structures is however continuously growing, with about 40 new structures being deposited each year. There are examples in the literature of glycoprotein structures where the resolved carbohydrate part of the structure accounts for as much as 9% (glucoamylase from *A. awamorii*; Aleshin *et al.*, 1996) and 11% (rhamnogalacturonase A from *A. aculeatus*; Petersen *et al.*, 1997) of the resolved covalently bound structure. However, a recent survey of proteins containing N- and O-glycan structures in the PDB showed that in about 50% of the cases only the first one or two residues of the glycan structure had been identified (Petrescu *et al.*, 1999).

The available models for the description of the carbohydrate moieties have not been developed to the same perfection as for the amino acids in the polypeptide chain and the heavily glycosylated structures would benefit from improved models for the carbohydrate structures. Among the 102 non-redundant (sequence identity < 90%; Hobohm *et al.*, 1993) protein structures in the PDB determined by X-ray diffraction methods to a resolution of 1.2 Å or better, only three are glycoproteins. The 1.12 Å structure of heparin-binding protein (PDB entry 1a7s; Karlsen *et al.*, 1998) contains three *N*-acetyl glucosamine residues. Two O-glycosylation sites each equipped with one mannose residue were identified in the 0.95 Å structure of penicillopepsin (PDB entry 1bxo; Khan *et al.*, 1998). The 1.2 Å structure of myrosinase represents the protein with the highest degree of glycosylation at atomic resolution (Burmeister *et al.*, 2000). It contains two branched N-linked glycan structures, one with five and the other with seven carbohydrate residues. All three structures were refined using data from a cryocooled crystal. These few results show that the structural knowledge of protein glycosylation at atomic resolution is still very limited despite its biological significance.

Rhamnogalacturonan acetylesterase (RGAE) from *A. aculeatus* is an excellent candidate for structural investigation of N-linked glycan structures. The consensus sequence for N-glycosylation is found at Asn104 and Asn182 and although the recombinant protein expressed in *A. oryzae* has been shown to be heterogeneously glycosylated, RGAE crystallizes readily, forming well diffracting crystals (Mølgaard *et al.*, 1998).

The structure determination to 1.55 Å resolution showed that RGAE belongs to a new family of esterases, the SGNH-hydrolase family (Mølgaard *et al.*, 2000), and that it adopts an α/β fold with a central five-stranded β-sheet sandwiched between α-helices, with a catalytic Ser9-His195-Asp192 triad oriented perpendicular to the central β-sheet. The SGNH hydrolase family also includes a serine esterase from *Streptomyces scabies* (SsEst; Wei *et al.*, 1995), a platelet-activating factor from *Bos taurus* (PAF-AH; Ho *et al.*, 1997) and haemagglutinin-esterase fusion glycoprotein from influenza C virus (HEF; Rosenthal *et al.*, 1998).

We present here the structure of RGAE based on data to 1.12 Å resolution collected with synchrotron radiation at 263 K. This structure refinement has identified a discrepancy with respect to the previous model of the structure in the C-terminal Leu233 and has enabled us to perform a thorough analysis of the seven covalently linked carbohydrate residues, one at Asn104 and six at Asn182.

2. Methods

2.1. Data collection and processing

Orthorhombic crystals were grown using Li_2SO_4 as a precipitant as described previously (Mølgaard *et al.*, 1998). Attempts to cryocool the crystals increased the mosaicity to more than 1° and diffraction data were therefore collected at 263 K using synchrotron radiation at the BW7B beamline at the EMBL Outstation, Hamburg. Using the wavelength λ = 1.1024 Å three data sets were collected from the same crystal. Two data sets were collected in the resolution range 12–1.12 Å, the second data set being collected to improve the completeness of the data. The lower resolution limit was chosen to exclude reflections within the beamstop shadow. The third data set was collected within the resolution range 38–2.05 Å. This low-resolution data set was collected with a dose which was 1.8% of the dose used for the high-resolution data in order to measure accurately the high-intensity reflections. The crystal belongs to the space group $P2_12_12_1$, with unit-cell parameters a = 52.17, b = 56.92, c = 71.69 Å. The data were integrated and merged using *DENZO* and *SCALEPACK* (Otwinowski & Minor, 1997). The resolution limit for the high-resolution data set was chosen so that 50% of the data have $I/\sigma(I) > 2$ in the outermost shell. Overloaded reflections were discarded. Combined merging of the three data sets resulted in an overall $R_{\rm merge}$ of 0.060. Data-collection statistics for the merged data set are listed in Table 1. For initial refinement with *X-PLOR* (Brünger, 1992*b*), the intensities were reduced to structure-factor amplitudes and brought to an approximate absolute scale using the scale factor estimated from the Wilson plot using the program *TRUNCATE* from the *CCP*4 suite (Collaborative Computational Project, Number 4, 1994). The subsequent refinement with *SHELXL*97 (Sheldrick & Schneider, 1997) was based on $|F|^2$

Table 1
Data collection and processing statistics.

Values in parentheses are for the outermost shell.

Resolution (Å)	38–1.12 (1.14–1.12)
Completeness (%)	97.7 (82.1)
Total No. of reflections	644629
No. of rejections†	6415
No. of unique reflections‡	80624
R_{merge}	0.060 (0.370)

† Reflections failing the merging procedure in *SCALEPACK*. ‡ Including systematically absent reflections.

and the reflection file from *SCALEPACK* was converted to *SHELXL*97 format using the program *SHELXPRO*.

2.2. Structure refinement

The 1.55 Å structure of RGAE (PDB code 1deo) was used as a starting point for the refinement including all 233 amino-acid residues, seven carbohydrate residues, two SO_4^{2-} ions and 153 water molecules. This model included no disordered residues. The set of 10% of the reflections chosen randomly that was used for the calculation of the R_{free} value (Brünger, 1992*a*), was extended to also cover the high-resolution reflections. An initial round of rigid-body refinement was carried out, followed by positional and *B*-factor refinement, all using the program *X-PLOR* (Brünger, 1992*b*). The resolution range for the *X-PLOR* refinement was 36–1.12 Å. The *R* and R_{free} values were 0.213 and 0.234, respectively, at this stage. The model resulting from this refinement was then taken into *SHELXL*97 (Sheldrick & Schneider, 1997), which was used for the remaining part of the refinement. For the refinement with *SHELXL*97, 5% of the data were used for calculation of the R_{free} value. These reflections were chosen randomly and were not identical to the reflections used for calculations of the R_{free} value in the *X-PLOR* refinement. The R_{free} value must therefore only be used to monitor the progress of the *SHELXL*97 refinement and not as an absolute value.

The strategy of the *SHELXL*97 refinement was similar to that suggested by Sheldrick & Schneider (1997). The refinement was carried out using conjugate-gradient refinement (CGLS) until the final rounds of refinement, where blocked least-squares refinement was used in order to obtain estimated standard deviations (e.s.d.s). At the beginning of the refinement, all automatically defined restraints were applied with default e.s.d.s. Additional restraints were set up for the sulfate ions and the carbohydrate residues, where the geometry was restrained to be similar within chemically equivalent groups but without specific target values. The N-linked glycan structure at Asn182 has the connectivity Manα1-6(Manα1-3)-Manα1-6Manβ1-4GlcNAcβ1-4GlcNAcβ-Asn182, as shown schematically in Fig. 1. The anomeric effect was not taken into account and all four mannose residues were thus restrained to be similar. In order to be able to assess the hydrogen-bonding pattern in the catalytic triad through a comparison of the CG OD1 and CG OD2 distances in Asp192, these bonds in the aspartate residues were not restrained. Diffuse solvent was modelled using Babinet's principle (Moews & Kretsinger, 1975). After refining the model from *X-PLOR* using data in the range 10–1.12 Å, the *R* and R_{free} values were 0.187 and 0.204, respectively. At this stage, anisotropic displacement parameters were introduced, which caused the *R* factor to drop to 0.136 and the R_{free} to 0.162. In the following cycles, more water molecules were introduced and several residues were modelled in double conformations using the program *O* (Jones *et al.*, 1991). After eight rounds of refinement, all data in the range 38–1.12 Å were employed. Dauter *et al.* (1997) have showed that although inclusion of the low-resolution data causes a slight increase in the *R* factor (in this case from 0.124 to 0.128) and R_{free} (from 0.153 to 0.158), the inclusion of this data is very important for obtaining the best solvent model and the best map quality. The second largest decrease in both the *R* factor (from 0.128 to 0.117) and R_{free} (from 0.158 to 0.145) after the large drop following the introduction of anisotropic displacement parameters occurred when H atoms were introduced into the model.

The quality of the model was checked using *PROCHECK* (Laskowski *et al.*, 1993) and *WHATCHECK* (Hooft *et al.*, 1996). The comparison of the structure with the previously published 1.55 Å structure of RGAE was performed using the program *LSQMAN* (Kleywegt, 1999).

The refinement parameters and statistics of the final model of RGAE are shown in Table 2. The model consists of 233 amino-acid residues, four sulfate ions, seven carbohydrate residues and 242 fully occupied and 87 partly occupied water molecules. 14 amino-acid residues were modelled in double conformations and two of the sulfate ions had partial occupancies (see Table 3).

3. Results and discussion

3.1. Quality of the model

As was observed in a study of eight high-resolution protein crystal structures by Wilson *et al.* (1998), there is a tight clustering of residues in the Ramachandran plot, with no residues in the lower right section of the most favoured α-helix region and the upper left of the β-strand section as defined by *PROCHECK* (Laskowski *et al.*, 1993). Recently, Kleywegt & Jones (1996) used 403 protein models at a resolution of 2.0 Å or better as a sample for re-evaluation of the core region of the

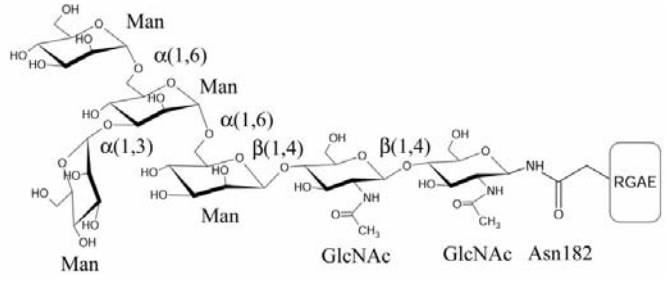

Figure 1
The N-linked glycan structure at Asn182. Shown are the residues that can be seen in the electron-density maps. At Asn104, only the first GlcNAc residue can be seen.

Table 2
Refinement parameters and statistics for the final model.

Resolution range (Å)	1.12–38.0
No. of reflections used in refinement	80568
No. of parameters	19970
No. of restraints	24756
R factor for $F_o > 4\sigma(F_o)$	0.103
R factor for all data	0.110
$R_{\rm free}$ for $F_o > 4\sigma(F_o)$	0.132
$R_{\rm free}$ for all data	0.139
Goodness of fit	1.34
No. of non-H atoms in	
Protein†	1735
Carbohydrate	86
Sulfate ions	20
Water molecules	329
$\langle B \rangle$ (Å^2)	
Protein main chain	12.5
Protein side chain‡	16.0
Carbohydrate	26.4
Sulfate ions	48.1
Water molecules	32.6

† Excluding disordered residues. ‡ Including disordered residues.

Table 3
The disordered residues in the 1.12 Å structure of RGAE.

The side chains of these amino-acid residues have been modelled in double conformations with the total occupancy constrained to 1; the occupancy of the major conformation is indicated here. For the sulfate ions, the occupancy has been refined for one conformation only.

Residue No.	Accessibility	Occupancy of side chain	Atoms involved
Ser32	57.44	0.73	CB OG
Val40	38.95	0.64	CB CG1 CG2
Asp59	82.37	0.59	CB CG OD1 OD2
Thr62	79.58	0.78	CB OG1 CG2
Ser80	76.07	0.71	CB OG
Asp101	100.75	0.65	CB CG OD1 OD2
Lys119	120.97	0.55	CD CE NZ
Leu120	85.90	0.53	CB CG CD1 CD2
Ser147	61.15	0.61	CB OG
Glu153	123.99	0.67	CB CG CD OE1 OE2
Thr215	45.86	0.63	CB OG1 CG2
Lys220	77.83	0.56	CB CG CD CE NZ
Leu223	1.83	0.63	CB CG CD1 CD2
Glu229	96.41	0.65	CB CG CD OE1 OE2
Sulf1		0.80	S O1 O2 O3 O4
Sulf4		0.51	S O1 O2 O3 O4

Ramachandran plot. 98% of all non-glycine residues were found to occupy a core region consisting of 19.7% of the entire plot area, which is consistent with the experimental distributions derived from the eight high-resolution structures. This version of the Ramachandran plot contains only two areas: core and non-core. According to this definition of the Ramachandran plot, the final high-resolution model of RGAE has 99.5% of all non-glycine residues in the core region, with Asp8 as the only outlier (Fig. 2).

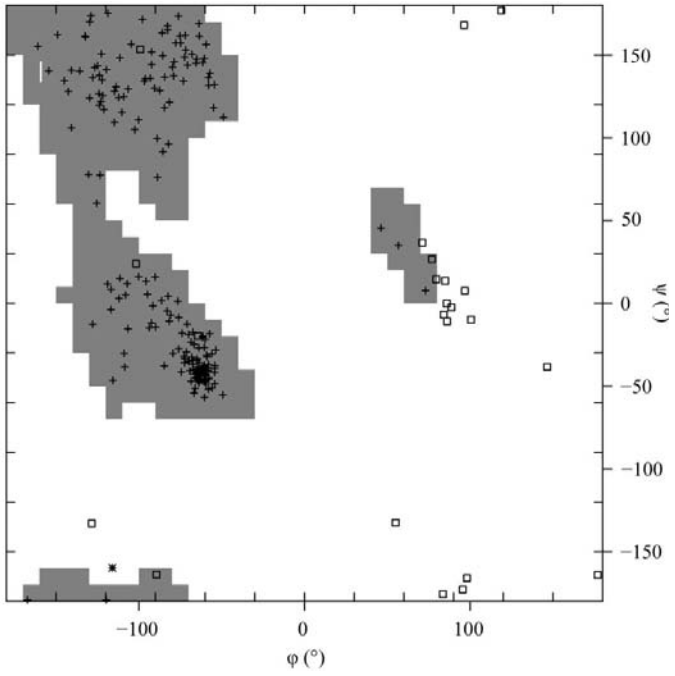

Figure 2
Ramachandran plot as defined by Kleywegt & Jones (1996). The outlier at (φ, ψ) = (−115.5, −159.9°) corresponds to Asp8. Glycine residues are shown as open squares.

3.2. The role of Asp8

The aspartate residue Asp8 next to the catalytic nucleophile (Ser9) is part of a type I β-turn, which has been shown to be characteristic of structures that adopt the SGNH-hydrolase fold (Mølgaard *et al.*, 2000). In the three structures that were included in the original analysis (PAF-AH, SsEst and RGAE), the conformation of this Asp is conserved and it is found in the same position in the Ramachandran plot. For PAF-AH the (φ, ψ) angles are (−98.4, −158.9°), for SsEst the (φ, ψ) angles are (−105.3, −163.6°) and for RGAE the angles are (−115.5, −159.9°). In the structure of HEF, the (φ, ψ) angles (−97.6, −109.0°) are in the same general region as in the other SGNH-hydrolase structures. It must be noted, however, that the resolution of this structure is 3.2 Å and the percentage of outliers in the Ramachandran plot as defined by Kleywegt & Jones (1996) is 11%, where a normal (better than 2 Å) structure is expected to have less than 5% outliers. This gives reason to expect that a structure of HEF to higher resolution would be more similar to that found for the other SGNH-hydrolases.

The role of this highly conserved Asp is thought to be stabilization of the flexible part of the β-turn by the formation of a hydrogen bond from the Asp side-chain O atom to the Nu + 1 backbone amide (Mølgaard *et al.*, 2000). In the otherwise structurally unrelated serine protease subtilisin (Kuhn *et al.*, 1998), the nucleophile is located on a type I β-turn as in the SGNH-hydrolases. Subtilisin does not have an Asp in the same position as the SGNH hydrolases, but a similar hydrogen bond is formed between the Nu − 1 backbone carboxyl group and the Nu + 1 amide group. The net effect of immobilizing the nucleophile is the same as in RGAE. When the nucleophilic β-turn in RGAE is superimposed upon the corresponding β-turn in subtilisin (Fig. 3), the atoms of the side chain of Asp8 coincide almost perfectly

Table 4
Carbohydrate bond lengths (Å).

Bond	RGAE	E.s.d.†	No. of bonds‡	*X-PLOR*§	Ref. 1¶	Ref. 2††	E.s.d.†	Ref. 3‡‡	E.s.d.†
C—C (ring)	1.523	0.003	28	1.526	1.526	1.527	0.001		
C—C (exocyclic)	1.507	0.007	7	1.516	1.516	1.511	0.002		
C—O (exocyclic)	1.429	0.004	18	1.420	1.420	1.418	0.001		
Axial glycosidic bond α-D-$^4C_1$									
C1—O5	1.410	0.008	3	1.419	1.419				
C5—O5	1.473	0.008	3	1.438	1.434				
C1—O1	1.409	0.009	3	1.398	1.398				
Equatorial glycosidic bond β-D-$^4C_1$									
C1—O5	1.477	0.009	4	1.428	1.428	1.424	0.003		
C5—O5	1.453	0.009	4	1.438	1.426	1.439	0.003		
C1—O1	1.373	0.014	2	1.385	1.385	1.396	0.003		
The *N*-acetyl moiety									
C2—N2	1.550	0.012	3	1.450				1.452	0.002
N2—C7	1.270	0.017	3	1.329				1.326	0.002
C7—O7	1.273	0.017	3	1.231				1.241	0.002
C7—C8	1.593	0.017	3	1.520				1.505	0.003

† $(1/n)[\sum_n \sigma^2(x)]^{1/2}$. ‡ The number of bond lengths that contribute to the average value for RGAE. § From param3.cho and parhcsdx.pro. ¶ Jeffrey *et al.* (1990). †† Hirotsu *et al.* (1974). ‡‡ Mo & Jensen (1975), Gilardi & Flippen (1974), Neuman *et al.* (1975*a*,*b*); average values.

Table 5
Bond angles (°).

References and definitions as in Table 4.

Angles	RGAE	E.s.d.	No. of angles	*X-PLOR*	Ref. 1	Ref. 2	E.s.d.	Ref. 3	E.s.d.
C—C—C (ring)	110.1	0.3	21	110.4	110.4	110.6	0.1		
C—C—O (ring)	109.4	0.4	14	110.0	110.0	109.3	0.1		
C—C—C (exocyclic)	111.6	0.6	7	112.5	112.5	112.9	0.1		
C—C—O (exocyclic)†	108.3	0.3	28	109.7	109.7	110.1	0.1		
Axial glycosidic bond α-D-$^4C_1$									
C5—O5—C1	115.8	0.7	3	114.0	114				
O5—C1—O1	109.6	0.7	3	112.1	112.1				
Equatorial glycosidic bond β-D-$^4C_1$									
C5—O5—C1	109.0	0.7	4	112.0	112	112.7	0.1		
O5—C1—O1	106.4	1.2	2	108.0	108	107.5	0.1		
The *N*-acetyl moiety									
C1—C2—N2	106.0	1.0	3	110.0				110.9	0.2
C3—C2—N2	102.5	1.0	3	110.0				111.3	0.2
C2—N2—C7	119.7	1.6	3	120.0				123.5	0.2
N2—C7—O7	119.7	1.7	3	123.0				123.1	0.2
N2—C7—C8	116.3	1.9	3	117.5				116.4	0.2
O7—C7—C8	121.0	1.7	3	121.5				120.6	0.2

† In the C—C—O exocyclic angles, only those angles that are not involved in glycosidic bonds are included, *i.e.* C—C—O(H).

with the N, C, O and C^α atoms of the backbone of subtilisin. This mimicking of the backbone can explain the unfavourable backbone conformation of the Asp residue.

3.3. Comparison of the 1.12 and 1.55 Å models of RGAE

The overall high-resolution structure is essentially identical to the 1.55 Å structure determined at room temperature. The two structures superimpose well, with an average distance of 0.16 Å based on C^α atoms. The distance between corresponding C^α atoms after superimposing the two structures is plotted against residue number in Fig. 4. The largest deviation corresponds to the C-terminal residue. This residue is a leucine and in the high-resolution model it became apparent from the electron-density maps that the side chain and the COO^- moieties should be interchanged by rotating around the φ and ψ angles of the residue (Fig. 5). All remaining distances above 0.30 Å correspond to disordered residues or residues adjacent to disordered residues.

The water molecules in the two structures were compared using a distance cutoff of 1.5 Å for matching water molecules. Of the 329 water molecules in the high-resolution structure, 147 (out of 153 possible) were also found in the 1.55 Å model. The average distance between matching water molecules is 0.32 Å and the average ΔB is 5.2 Å^2.

Of the two sulfate ions that are shared in the 1.55 and the 1.12 Å structures, one is located in the active site mimicking the tetrahedral intermediate of the substrate (Mølgaard *et al.*, 2000) and one is hydrogen bonded to Arg150. One of the additional sulfate ions in the 1.12 Å model is hydrogen bonded to the main-chain amino group of Gly77 and to Arg150 and the fourth to His193 and Asn136.

The 1.12 Å structure of RGAE displays a glycan structure at both of the two putative N-glycosylation sites. At Asn104 only the first GlcNAc residue is resolved in the electron density, but at Asn182 the glycan structure participates in crystal contacts to three neighbouring symmetry-related molecules and a total of six carbohydrate residues could be modelled into the electron density, as shown in Fig. 6. Fig. 7 illustrates the quality of the electron-density maps at the two N-glycosylation sites.

3.4. Carbohydrate parameters compared with parameters obtained from small-molecule studies

Owing to the limited observation-to-parameter ratio in most protein structures, it is necessary to apply a great number of geometrical restraints on the model. Only at near-atomic resolution (1.2 Å or better) is it possible to loosen the restraints on the bond lengths and angles, but even in the 0.97 Å resolution structure of dethiobiotin synthetase it was

Table 6
Conformation of glycosidic linkages (°) in RGAE compared with average torsion angles for the distinct conformers that were found in the database (Petrescu *et al.*, 1999).

The nomenclature used for the torsion angles is φ = O5—C1—O—C(x)′ and ψ = C1—O—C(x)′—C(x−1)′ for 1–2, 1–3 and 1–4 linkages (x = 2, 3 or 4); φ = O5—C1—O—C6′, ψ = C1—O—C6′—C5′ and ω = O—C6′—C5′—C4′ for 1–6 linkages.

	φ_{RGAE}	$\langle\varphi\rangle$	ψ_{RGAE}	$\langle\psi\rangle$	ω_{RGAE}	$\langle\omega\rangle$
GlcNAcβ1-4GlcNAc	−70.5	−73.7 ± 8.4	131.4	116.8 ± 15.6		
Manβ1-4GlcNAc	−85.2	−88.0 ± 10.8	110.3	107.9 ± 20.3		
Manα1-6Man	64.3	66.5 ± 10.8	166.9	180.7 ± 15.1	179.2	185.0 ± 11.2
	57.2	65.4 ± 9.0	183.8	182.6 ± 5.1	57.8	66.4 ± 10.2
Manα1-3Man	81.7	72.5 ± 11.0	−97.9	−112.3 ± 22.5		

not possible to perform completely unrestrained refinement despite a relatively high ratio of observations to parameters (5.6:1) (Sandalova *et al.*, 1999). In the 0.78 Å structure of *Bacillus lentus* subtilisin, an unrestrained refinement (10:1 ratio of observations to parameters) was carried out without problems however (Kuhn *et al.*, 1998). The target values for the geometrical restraints are usually taken from an analysis of small-molecule X-ray structures from the Cambridge Structural Database performed by Engh & Huber (1991).

The corresponding library of carbohydrate geometry parameters is much more weakly founded than the parameters for amino-acid geometries. Whereas the Engh & Huber bond-length and bond-angle parameters were derived from a statistical survey of appropriate chemical fragments from the Cambridge Structural Database of more than 80 000 small-molecule structures, the parameters used in *X-PLOR* (Brünger, 1992*b*) and *CNS* (Brunger *et al.*, 1998) for carbohydrate geometries are based on studies of 13 pyranoses and methyl pyranosides whose structures were determined by neutron diffraction (Jeffrey & Taylor, 1980; Jeffrey, 1990) and a study comparing the geometries of four disaccharides with β-(1,4) linkages determined by X-ray crystallography (Hirotsu & Shimada, 1974). These studies do not include GlcNAc residues, so in the *X-PLOR* library the parameters used for the *N*-acetyl group of GlcNAc are 'as extended-atom carbon-N'.

During the refinement of the 1.12 Å structure of RGAE, the bond lengths and angles of the carbohydrate residues were restrained to be similar within chemically equivalent residues but without target values. This should provide an unbiased model of the protein-bound glycan structure and the parameters obtained can be compared with the parameters that are normally used as a carbohydrate dictionary in the refinement of medium- and low-resolution glycoprotein structures. The seven carbohydrate residues in the structure of RGAE can become an important part of an analysis of the validity of the existing carbohydrate

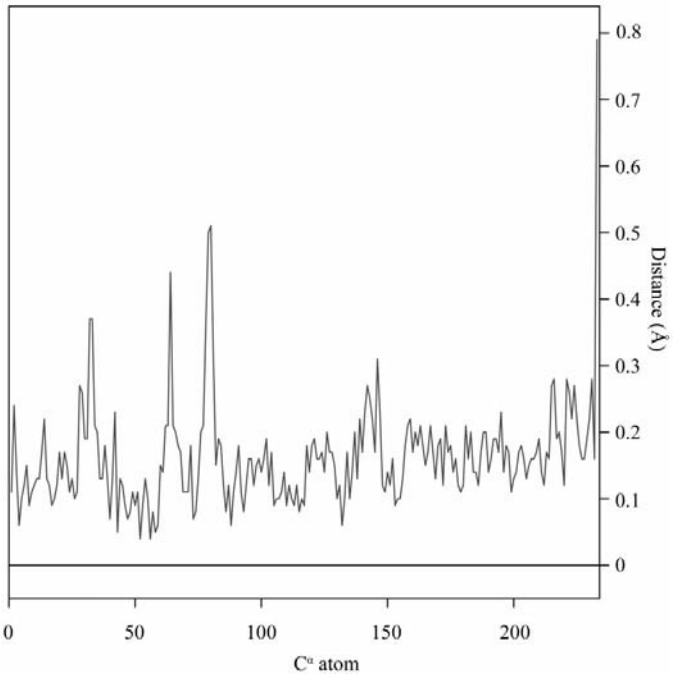

Figure 4
The distance between the corresponding C$^{\alpha}$ atoms (1–233) of the 1.55 and 1.12 Å models of RGAE.

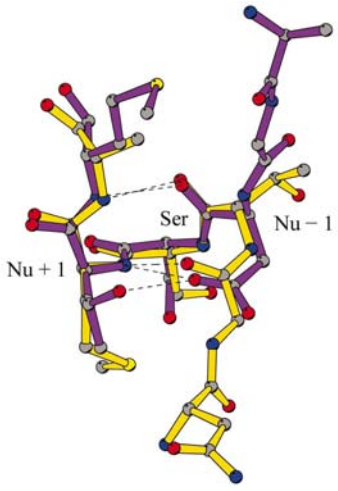

Figure 3
Superpositioning of the nucleophilic β-turn in RGAE (purple) and subtilisin (Kuhn *et al.*, 1998) (gold).

Figure 5
The C-terminal Leu233 in the 1.55 Å (gold) and the 1.12 Å (purple) structures of RGAE. The electron-density maps were calculated using the coordinates of the 1.55 Å model. The $2mF_o - DF_c$ map is contoured at 0.826 e Å^{-3} (1.6σ) and is shown in grey and the $mF_o - DF_c$ map is contoured at 0.28 e Å^{-3} (4.0σ) and is shown in green.

dictionaries as more atomic resolution structures with protein-bound glycans become available.

The average bond lengths and angles from the seven carbohydrate residues in the structure of RGAE were compared with the parameters used in the *X-PLOR* and *CNS* library taken from the parameter files param3.cho and parhcsdx.pro, as shown in Tables 4 and 5. Also included in the tables are the corresponding parameters from the studies of Jeffrey (1990) and Hirotsu & Shimada (1974). The bond lengths and angles of the N-acetyl group of the GlcNAc moiety were compared with average values obtained from small-molecule X-ray crystallographic studies of *N*-acetyl-α-D-glucosamine (Mo & Jensen, 1975), *N*-acetyl-α-D-galactosamine (Gilardi & Flippen, 1974; Neuman *et al.*, 1975*a*) and *N*-acetyl-β-D-mannosamine monohydrate (Neuman *et al.*, 1975*b*).

3.4.1. The pyranose-ring parameters. The parameters that were determined best in the structure of RGAE are the bond lengths and angles that relate to the pyranose rings. This is the the most rigid part of the carbohydrate structure and generally the most well defined part of the electron density. The parameters from RGAE which correspond most closely to the *X-PLOR* parameters are the ring bonds and angles; those that differ the most are the exocyclic parameters.

3.4.2. The anomeric effect. The geometry around the anomeric C1 atom in the pyranoses depends upon the configuration of the carbohydrate residue. The anomeric C atom is bonded to two more electronegative atoms, the O5 in the pyranose ring and the glycosidic O1 atom. The polarization in the C—O bonds causes a reduction in the σ electron density and an increase in the π bonding in the C—O bonds. This is best achieved when the orbitals have maximum overlap, which leads to a configurational preference for an electronegative substituent at the axial position, to bond-length differences and to a preferred *gauche* conformation about the glycosidic bond (Lemieux *et al.*, 1979). The anomeric effect has been investigated both crystallographically and through quantum-mechanical calculations (Jeffrey *et al.*, 1974, 1978).

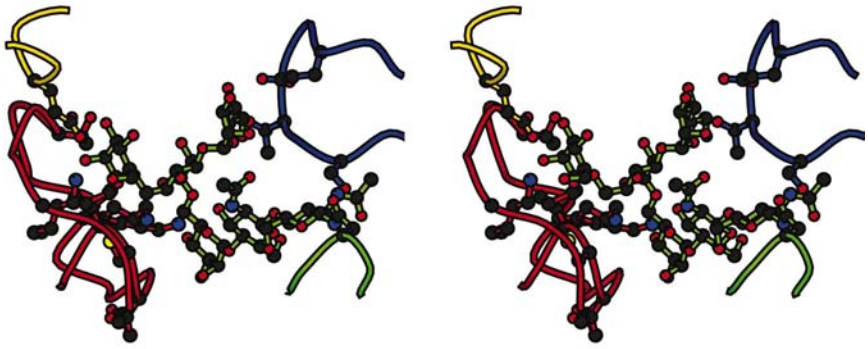

Figure 6
A stereoview of the glycan structure at Asn182 connecting four symmetry-related molecules (illustrated in different colours). The amino-acid residues that are involved in crystal contacts to the glycan moiety are shown in ball-and-stick representation.

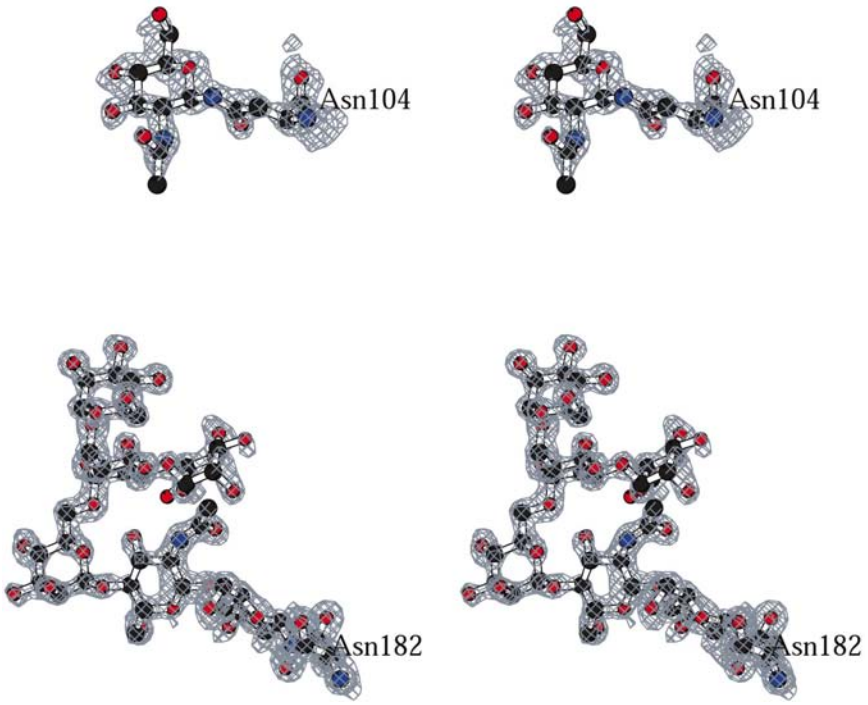

Figure 7
Stereoviews of the electron density at the two N-glycosylation sites contoured at 0.840 e Å^{-3} (1.6σ).

In the refinement of the high-resolution structure of RGAE, the bond lengths and angles of the individual GlcNAc and Man residues were restrained to be similar using the SAME restraint in *SHELXL*97. Bond lengths and angles of the α-glycosidic linkages were restrained to be similar using SADI restraints and similar restraints were imposed on the β-linkages. The restraints on the bond lengths were given an e.s.d. of 0.02, whereas the 1,3 distances which represent the bond angles were restrained with an e.s.d. of 0.04, allowing a greater variability of the angles. The anomeric effect is thus taken into consideration for the glycosidic linkages, but not in the C1—O5 and the C5—O5 bonds. The relatively large e.s.d. on the SAME restraints however, should be sufficiently large to allow for the natural variation in these parameters. This is also reflected in the large difference between the C1—O5 bond lengths in the axial and equatorial glycosidic bonds.

With the exceptions of the axial C5—O5 and the equatorial C1—O5 bonds, all of the bond lengths in RGAE involving the anomeric C atom and the ring O atom are

within the standard deviations comparable to the bond lengths found in the small-molecule studies. The anomeric effect which leads to greater bond-length differences in the axially substituted residues than in the residues involved in equatorial glycosidic linkages is seen in both the small-molecule structures and in RGAE.

3.4.3. The *N*-acetyl moiety. The parameters in *X-PLOR* for the *N*-acetyl moiety are based on idealized values and do not differ significantly from the average values derived from the four small-molecule structures described earlier. The bond lengths and angles determined from the structure of RGAE deviate quite a lot from the small-molecule structures compared with their e.s.d.s, but this may be explained by the fact that the electron-density maps are not as well resolved in these parts of the glycan structures as in the rings.

3.4.4. Conformation of the glycosidic linkages. Recently, a statistical analysis of N- and O-linked glycan conformations was performed using crystallographic data (Petrescu *et al.*, 1999). A small number of distinct conformers were identified for all linkages. The information from this database of linkage structures can be used analogously to the library of side-chain rotamers both in the model-building phase of protein structure determination and in the validation phase, where outliers can be identified and subjected to further examination.

The glycan linkages in the structure of RGAE were compared with the average torsion angles for the distinct conformers that were found in the database. All glycan-linkage torsion angles in RGAE fall within the standard deviations of the average torsion angles in the database (see Table 6).

4. Conclusions

The 1.12 Å structure of RGAE has been refined using anisotropic displacement parameters. The model includes six residues of N-glycan structure at one of the two glycosylation sites. The glycan structure was refined without restraints toward target values and may thus, as more atomic resolution structures with protein-bound glycans appear, contribute to an evaluation of the presently used target values obtained from small-molecule studies that are used for medium- and low-resolution glycoprotein structures. The use of dictionaries of carbohydrate parameters are presently only used in the refinement stage of glycoprotein structure determination, but high-quality dictionaries would also be very useful in the model building, validation and analysis stages. An obvious start would be to implement rotamer libraries of glycosidic conformations in model-building programs using the dictionary created by Petrescu *et al.* (1999).

Support from the Danish National Research Foundation is gratefully acknowledged. We are grateful to the EMBL Hamburg Outstation c/o DESY for beam time and for support from the European Community Access to Research Infrastructure Action of the Improving Human Potential Programme to the EMBL Hamburg Outstation (Contract Number HPRI-1999-CT-00017). We thank Anders Kadziola, Paul Rowland and Flemming Hansen for help with the data collection, and Pernille Harris and Leila Lo Leggio for critical reading of the manuscript.

References

Aleshin, A. E., Stoffer, B., Firsov, L. M., Svensson, B. & Honzatko, R. B. (1996). *Biochemistry*, **35**, 8319–8328.
Apweiler, R., Hermjakob, H. & Sharon, N. (1999). *Biochim. Biophys. Acta*, **1473**, 4–8.
Bairoch, A. & Apweiler, R. (1999). *Nucleic Acids Res.* **27**, 49–54.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
Brünger, A. T. (1992*a*). *Nature (London)*, **355**, 472–475.
Brünger, A. T. (1992*b*). *X-PLOR Version 3.1. A System for X-ray Crystallography and NMR*. New Haven, CT, USA: Yale University Press.
Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.
Burmeister, W. P., Cottaz, S., Rollin, P., Vasella, A. & Henrissat, B. (2000). *J. Biol. Chem.* **275**, 39385–39393.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Dauter, Z., Lamzin, V. S. & Wilson, K. S. (1997). *Curr. Opin. Struct. Biol.* **7**, 681–688.
Engh, R. A. & Huber, R. (1991). *Acta Cryst.* A**47**, 392–400.
Feizi, T. (2000). *Glycoconj. J.* **17**, 553–565.
Gabius, H.-J. (2000). *Naturwissenschaften*, **87**, 108–121.
Gilardi, R. D. & Flippen, J. L. (1974). *Acta Cryst.* B**30**, 2931–2933.
Hirotsu, K. & Shimada, A. (1974). *Bull. Chem. Soc. Jpn*, **47**, 1872–1879.
Ho, Y. S., Swenson, L., Derewenda, U., Serre, L., Wei, Y., Dauter, Z., Hattori, M., Adachi, T., Aoki, J., Arai, H., Inoue, K. & Derewenda, Z. S. (1997). *Nature (London)*, **385**, 89–93.
Hobohm, U., Scharf, M. & Schneider, R. (1993). *Protein Sci.* **1**, 409–417.
Hooft, R. W. W., Vriend, G., Sander, C. & Abola, E. E. (1996). *Nature (London)*, **381**, 272.
Jeffrey, G. A. (1990). *Acta Cryst.* B**46**, 89–103.
Jeffrey, G. A., Pople, J. A., Binkley, J. S. & Vishveshwara, S. (1978). *J. Am. Chem. Soc.* **100**, 373–379.
Jeffrey, G. A., Pople, J. A. & Radom, L. (1974). *Carbohydr. Res.* **38**, 81–95.
Jeffrey, G. A. & Taylor, R. (1980). *J. Comput. Chem.* **1**, 99–109.
Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.
Karlsen, S., Iversen, L. F., Larsen, I. K., Flodgaard, H. J. & Kastrup, J. S. (1998). *Acta Cryst.* D**54**, 598–609.
Khan, A. R., Parrish, J. C., Fraser, M. E., Smith, W. W., Bartlett, P. A. & James, M. N. (1998). *Biochemistry*, **37**, 16839–16845.
Kleywegt, G. J. (1999). *Acta Cryst.* D**55**, 1878–1884.
Kleywegt, G. J. & Jones, T. A. (1996). *Structure*, **4**, 1395–1400.
Kuhn, P., Knapp, M., Soltis, S. M., Ganshaw, G., Thoene, M. & Bott, R. (1998). *Biochemistry*, **37**, 13446–13452.
Laine, R. A. (1994). *Glycobiology*, **4**, 759–767.
Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.
Lemieux, R. U., Koto, S. & Vorsin, D. (1979). *Anomeric Effect: Origin and Consequences*, p. 87. Washington, DC: American Chemical Society.
Mo, F. & Jensen, L. H. (1975). *Acta Cryst.* B**31**, 2867–2873.
Moews, P. C. & Kretsinger, R. H. (1975). *J. Mol. Biol.* **91**, 201–225.

Mølgaard, A., Kauppinen, S. & Larsen, S. (2000). *Structure*, **8**, 373–383.
Mølgaard, A., Petersen, J. F. W., Kauppinen, S., Dalbøge, H., Johnsen, A. H., Poulsen, J.-C. N. & Larsen, S. (1998). *Acta Cryst.* D**54**, 1026–1029.
Neuman, A., Gillier-Pandraud, H. & Longchambon, F. (1975*a*). *Acta Cryst.* B**31**, 474–477.
Neuman, A., Gillier-Pandraud, H. & Longchambon, F. (1975*b*). *Acta Cryst.* B**31**, 2628–2631.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Petersen, T. N., Kauppinen, S. & Larsen, S. (1997). *Structure*, **5**, 533–544.
Petrescu, A. J., Petrescu, S. M., Dwek, R. A. & Wormald, M. R. (1999). *Glycobiology*, **9**, 343–352.
Rosenthal, P. B., Zhang, X., Formanowski, F., Fitz, W., Wong, C.-H., Meier-Ewert, H., Skehell, J. J. & Wiley, D. C. (1998). *Nature (London)*, **396**, 92–96.
Sandalova, T., Schneider, G., Käck, H. & Lindqvist, Y. (1999). *Acta Cryst.* D**55**, 610–624.
Sheldrick, G. M. & Schneider, T. R. (1997). *Methods Enzymol.* **277**, 319–343.
Solis, D., Jimenez-Barbero, J., Kaltner, H., Romero, A., Siebert, H.-C., von der Lieth, C.-W. & Gabius, H.-J. (2001). *Cells Tissues Organs*, **168**, 5–23.
Varki, A. (1993). *Glycobiology*, **3**, 97–130.
Vosseller, K., Wells, L. & Hart, G. W. (2001). *Biochimie*, **83**, 575–581.
Wei, Y., Schottel, J. L., Derewenda, U., Swenson, L., Patkar, S. & Derewenda, Z. S. (1995). *Nature Struct. Biol.* **2**, 218–223.
Wilson, K. S., Butterworth, S., Dauter, Z., Lamzin, V. S., Walsh, M., Wodak, S., Pontius, J., Richelle, J., Vaguine, A., Sander, C., Hooft, R. W. W., Vriend, G., Thornton, J. M., Laskowski, R. A., MacArthur, M. W, Dodson, E. J., Murshudov, G., Oldfield, T. J., Kaptein, R. & Rullmann, J. A. C. (1998). *J. Mol. Biol.* **276**, 417–436.
Wormald, M. R. & Dwek, R. A. (1999). *Structure*, **7**, R155–R160.