The primary structure of a 1,4-β-glucan cellobiohydrolase from the fungus *Trichoderma reesei* QM 9414

Lars G. Fägerstam, L. Göran Pettersson* and J. Åke Engström+

Institute of Biochemistry, University of Uppsala, Biomedical Center, PO Box 576, S-75123 Uppsala and +*Institute of Medical Chemistry, University of Uppsala, Biomedical Center, PO Box 575, S-75123 Uppsala, Sweden*

Received 30 December 1983

The sequence of the approx. 490 amino acid residues of the main 1,4-β-glucan cellobiohydrolase (CBH I) (EC 3.2.1.91) from culture filtrates of the fungus *Trichoderma reesei* QM 9414 has been established by automatic liquid phase Edman degradation. Peptides obtained by chemical and enzymatic cleavage of the reduced and *S*-carboxymethylated protein were isolated by a combination of gel filtration and high-performance liquid chromatography. The amino-terminus of the single polypeptide chain is blocked by a pyroglutamyl residue. Most of the neutral carbohydrate present in the glycoprotein is bound within a short region near the carboxyl-terminus. Three attachment sites of glucosamine residues have also been established.

Cellulase *Cellobiohydrolase* Trichoderma reesei *Amino acid sequence*

1. INTRODUCTION

Culture filtrates of different species of the fungus *Trichoderma* have been shown to contain at least 3 types of cellulolytic enzymes; 1,4-β-glucan glucanohydrolase (EC 3.2.1.4), 1,4-β-glucan cellobiohydrolase (EC 3.2.1.91) and β-glucosidase (EC 3.2.1.21). The 1,4-β-glucan glucanohydrolase attacks internal glucosidic bonds in less ordered regions of the cellulose fibrils, thereby increasing the number of free chain ends that can be attacked by the 1,4-β-glucan cellobiohydrolase. The cellobiose split of by this enzyme is then converted to glucose by the β-glucosidase. The cellobiohydrolase activity has been shown to be present as at least two structurally unrelated enzymes [1,2]. Under optimal culturing conditions the most abundant cellobiohydrolase constitutes about 50% of the protein released into the culture medium [3]. This enzyme, which is readily purified by ion exchange chromatography on DEAE–Sepharose in large quantities [4,5], has previously been reported to be a glycoprotein consisting of a single polypeptide chain [4,6]. The enzyme is present in the filtrate as several subforms, differing in glycosylation [7] which can be resolved by isoelectric focusing [1,5].

Although the amino acid sequence determination described here was done without separation of the subforms of the enzyme, no heterogeneity has been observed in the sequence established so far.

2. MATERIALS AND METHODS

2.1. *1,4-β-Glucan cellobiohydrolase*

Freeze-dried culture filtrate of *Trichoderma reesei* QM 9414 was a gift from Dr M. Mandels, US Army Research and Development Command, Natick, MA. The 1,4-β-glucan cellobiohydrolase (EC 3.2.1.91) was isolated as in [2].

* To whom correspondence should be addressed

Abbreviations: CBH I, 1,4-β-glucan cellobiohydrolase; RCM, reduced and carboxymethylated

Published by Elsevier Science Publishers B.V.

2.2. *Enzymes*

Trypsin (DPCC-treated), chymotrypsin (TLCK-treated) and carboxypeptidases A and B were purchased from Sigma. *Staphylococcus aureus* V8 protease (S-protease) was obtained from Miles Laboratories. Calf liver pyroglutamyl aminopeptidase was purchased from Boehringer Mannheim GmbH. Thermolysin was obtained from Chugai Boyeki Co., Japan. [^{14}C]Formaldehyde and iodo[^{14}C]acetic acid were obtained from The Radionuclear Centre, Amersham, England.

2.3. *Methods*

The enzyme was reduced with dithiothreitol and alkylated with iodoacetic acid as in [8].

CNBr cleavage was done as in [9,10]. Since the reduced and *S*-carboxymethylated protein (RCM-protein) was not readily soluble in 70% formic acid, it was first dissolved in the acid which was then diluted to 70% with distilled water. This greatly increased the cleavage yield. The CNBr fragments were purified by gel filtration on Sephadex G-75 superfine (Pharmacia Fine Chemicals, Sweden) as described in the legends to fig.1,2.

The intact RCM-protein and the CNBr fragments were digested with trypsin (1 h, 37°C), chymotrypsin (1 h, 37°C), thermolysin (40 min, 37°C) or with S-protease (20 h, room temperature). All digestions were done in 0.25 M NH_4HCO_3 at an enzyme:substrate ratio of 1:100 (w/w). The resulting peptide mixtures were separated by gel filtration on Sephadex G-75 or G-50 superfine columns. When further purification was found necessary it was done by reversed-phase chromatography on Lichrosorb RP-8 (E. Merck) or Spherisorb S10 ODS (Phase Separations, Ltd).

Carboxypeptidase A and B digestions were done at 37°C in 0.1 M *N*-ethylmorpholine acetate buffer (pH 8.0), at an enzyme:substrate molar ratio of 1:400 using norleucine as internal standard. The reaction was terminated at times 0, 1, 2, 3, 4, 5, 10, 15, 30 and 60 min with glacial acetic acid. Released amino acids were determined by amino acid analysis.

Labelling of lysyl residues by reductive methylation [12] was done by adding a 5-fold molar excess of [^{14}C]formaldehyde (10 Ci/mol) over lysyl residues to 1 μmol of sample dissolved in 2 ml of 0.1 M *N*-ethylmorpholine acetate buffer (pH 8.5) containing 20 mM $NaCNBH_3$. After 6 h incubation at 37°C excess reagents were removed by gel filtration on Sephadex G-50.

Methionyl residues were alkylated by adding 1 μmol iodo[^{14}C]acetate (50 Ci/mol) to 0.5 μmol RCM-protein dissolved in 4 ml of 5% formic acid in 6 M guanidine hydrochloride. After incubation at 37°C for 24 h unlabeled iodoacetate was added to give a 50-fold molar excess over methionine and after another 24 h at 37°C the sample was desalted by dialysis against distilled water [13,14].

Amino acid analysis was performed on a Durrum D-500 analyzer after hydrolysis for 24 h at 110°C in 6 M HCl containing 2 mg/ml of phenol. The serine and threonine values were corrected using the recovery factors 0.90 and 0.96, respectively. Tryptophan was determined spectrophotometrically in conjunction with the amino acid analysis.

The carbohydrate content of effluents from gel filtration columns was estimated by the orcinol method [16].

Liquid scintillation counting was used to monitor the radioactivity in effluents from gel filtration and HPLC columns and to determine radioactivity in PTH-amino acid derivatives from the sequencer.

A Beckman 890 C Sequencer (Beckman Instruments, USA) was used in all sequence determinations. PTH-amino acid derivatives were identified by reversed-phase chromatography. For sequencing of short peptides, Polybren (Aldrich Co.) was added to the sequenator cup [17,18].

3. RESULTS AND DISCUSSION

3.1. *Isolation and characterization of CNBr fragments*

An attempt to determine the N-terminal sequence of the RCM-polypeptide by Edman degradation was unsuccessful, indicating that the α-amino group was blocked. By treatment with pyroglutamyl aminopeptidase [11] a free α-amino group became available enabling automatic sequencing of the first 23 residues following the N-terminal pyroglutamyl residue.

Treatment of the polypeptide with carboxypeptidase A for 1 h released one residue of leucine.

After CNBr treatment of the RCM-protein 7

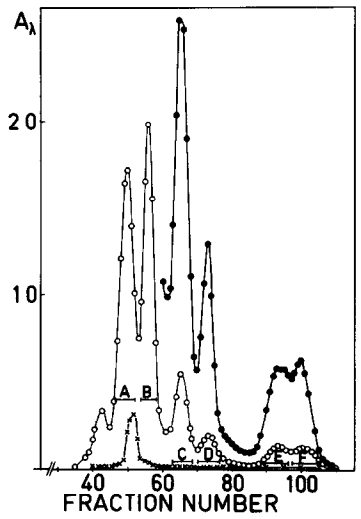


Fig.1. Gel filtration of 110 mg of CNBr-treated reduced and *S*-carboxymethylated *Trichoderma reesei* 1,4-β-glucan cellobiohydrolase on Sephadex G-75 superfine (2.6 × 98 cm) in 0.05 M ammonium bicarbonate. Fractions of 4.7 ml were collected at a flow rate of 18.8 ml/h. (○) $A_{280\ nm}$, (●) $A_{230\ nm}$, (×) $A_{505\ nm}$ (orcinol). Content of pools: A, CN I and CN II; B, CN III; C, CN IV; D, CN V; E, CN VI; F, CN VII.

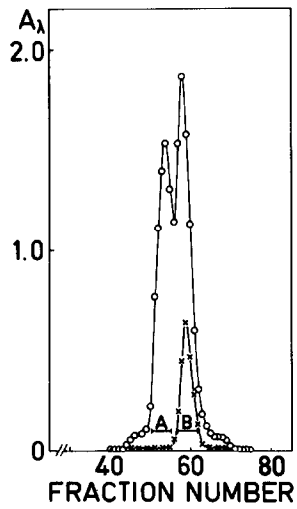


Fig.2. Rechromatography of pool A (fig.1) on Sephadex G-75 superfine (2.6 × 98 cm) in 0.05 M *N*-ethylmorpholine acetate buffer (pH 8.0). Fractions of 4.5 ml were collected at a flow rate of 20.0 ml/h. (○) $A_{280\ nm}$, (●) $A_{505\ nm}$ (orcinol). Content of pools: A, CN I; B, CN II.

fragments were isolated by two steps of gel filtration on Sephadex G-75 superfine. The first step, in 0.05 M NH_4CO_3 (pH 7.8), gave good separation of all but the two largest fragments (fig.1). Rechromatography of the material in pool A on a column of Sephadex G-75 superfine in 0.05 M *N*-ethylmorpholine acetate buffer (pH 8.0) separated the two fragments (fig.2). The reason for this change in behaviour might be a conformational change in one of the fragments, or weak interaction with the gel matrix. Fragments CN VI and CN VII were further purified by HPLC.

CNBr cleavage of the RCM-protein unexpectedly gave two fragments lacking homoserine, i.e., fragments CN II and CN V. However, from a digest of the RCM-protein with S-protease, the peptide G1 (fig.3) was isolated which corresponds to the sequence from residue 16 onwards in fragment CN V. This peptide contains a methionyl residue at position 23, indicating the cleavage of a peptide bond not involving methionine under the conditions used for CNBr cleavage. An attempt to obtain this split by treatment of the RCM-protein with 70% formic acid was unsuccessful. The peptide, containing homoserine, which would constitute the C-terminal part of the unsplit CN V fragment was not isolated.

Sequence analysis of fragment CN VI and of CN VII revealed that the latter constitutes the C-terminal 10 residues of fragment CN VI. This is due to a Met–Ser peptide bond which is known to give poor cleavage yields [20,21]. In the present case the yield was about 50%. The small fragment corresponding to the first 4 residues of CN VI was not isolated.

Analysis of the eluate from the two gel filtration steps by the orcinol method indicated that most of the carbohydrate bound to the enzyme is found in fragment CN II. Amino acid analysis further revealed the presence of glucosamine in fragments CN I, CN II, and CN III.

How the enzymatically derived peptides are used to establish the sequence of the individual CNBr fragments is shown in fig.3.

3.2. *Alignment of CNBr fragments*

From a digest of the RCM-protein with S-protease only two methionine-containing peptides were recovered. Sequence analysis of peptide G2 revealed the sequence: Met–Asp–Ile–Trp–Glu,

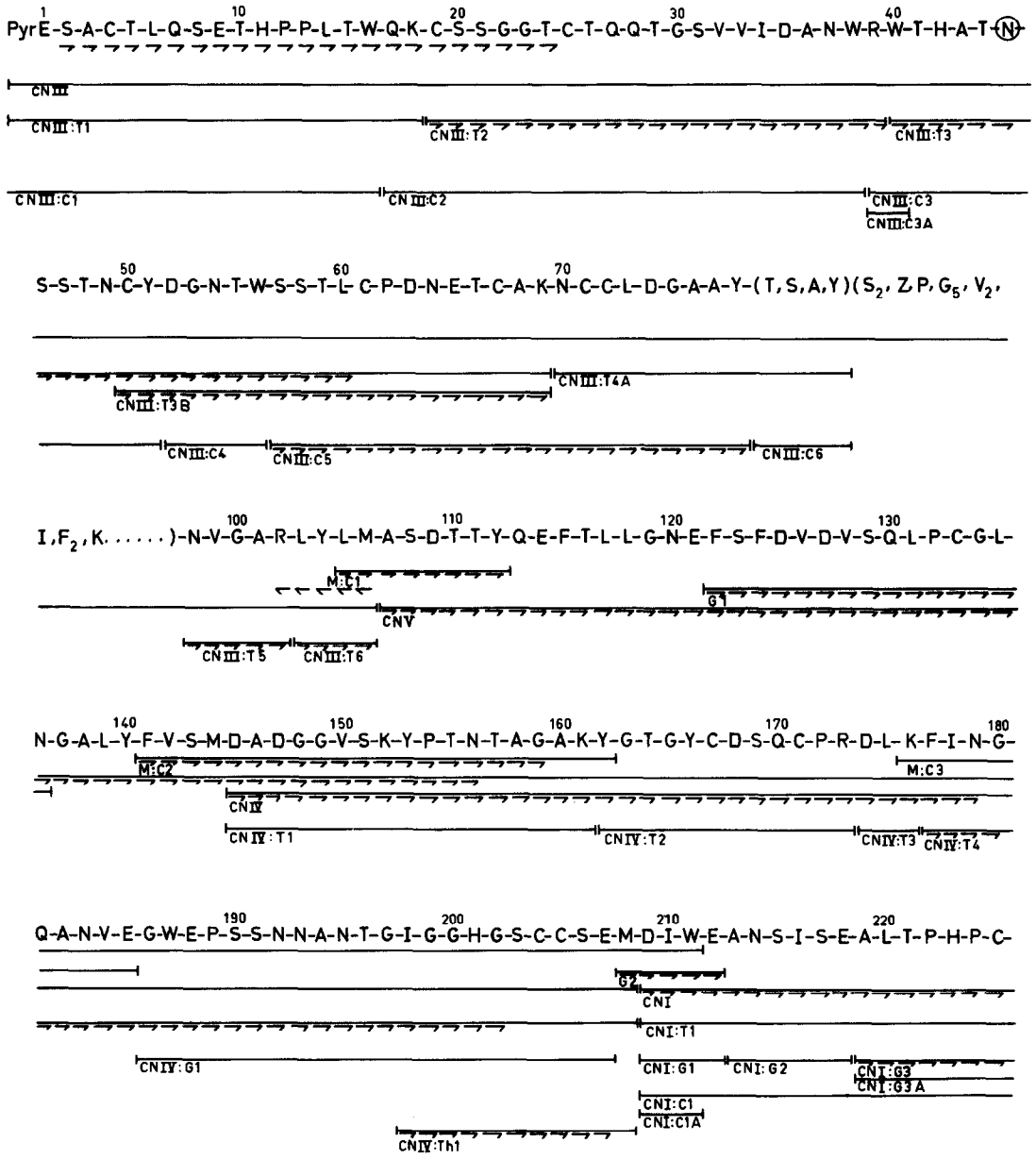
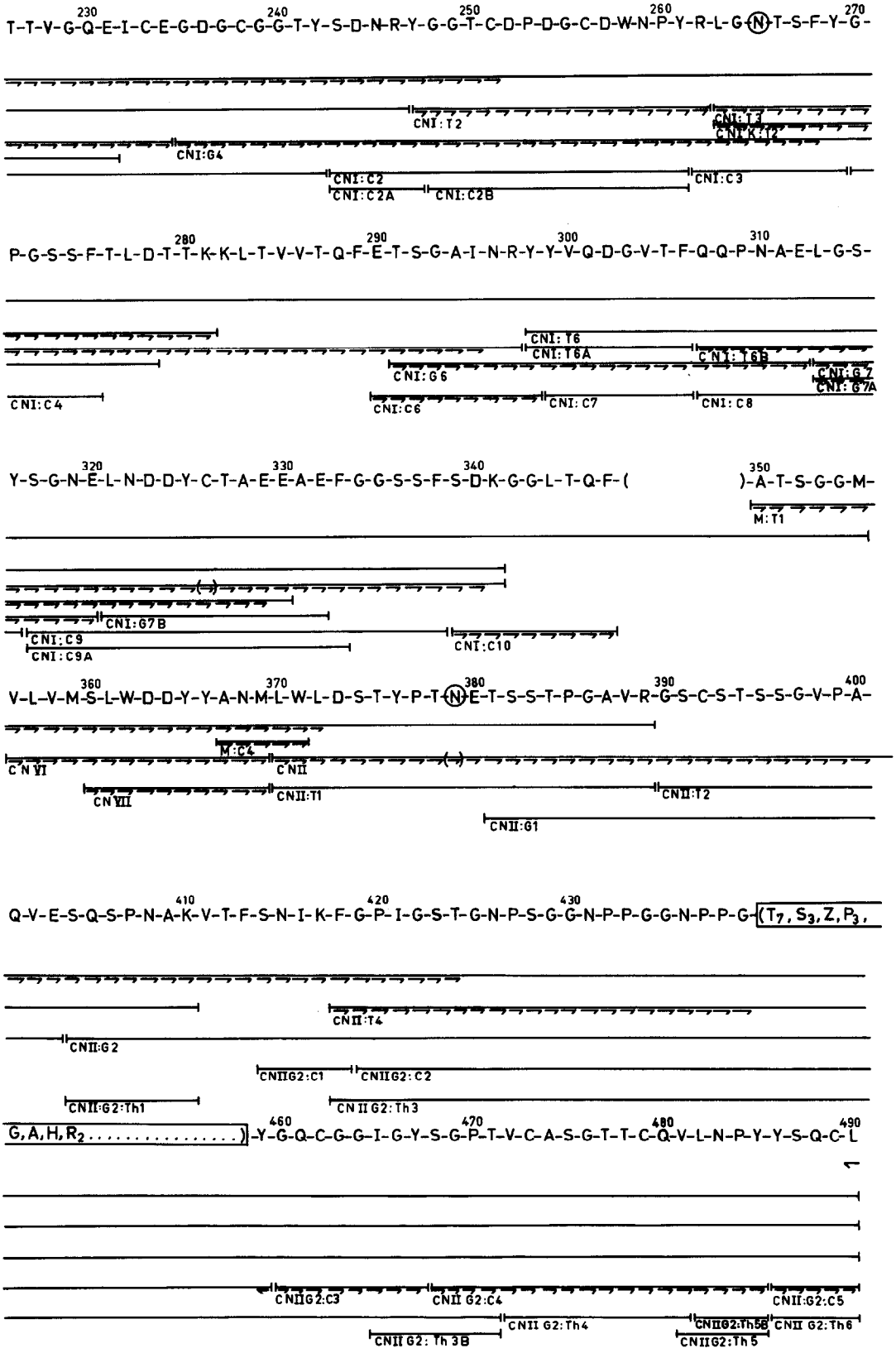


Fig.3. The complete sequence of cellobiohydrolase. Unsequenced fragments with known amino acid composition are in parentheses. Nomenclature of peptides: CNBr peptides from gel filtration are denoted CN I–CN VII. Arabic numbers denote the order of elution of peptides, of tryptic (T), chymotryptic (Ch), thermotryptic (Th) and S-protease (G) digestion, e.g., CN II: C3. Radioactive peptides containing [^{14}C]methionine are denoted by M.


T-T-V-G-Q-E-I-C-E-G-D-G-C-G-G-T-Y-S-D-N-R-Y-G-G-T-C-D-P-D-G-C-D-W-N-P-Y-R-L-G-N-T-S-F-Y-G-
CNI:T2
CNI:T3
CNI:K12
CNI:G4
CNI:C2
CNI:C3
CNI:C2A
CNI:C2B
P-G-S-S-F-T-L-D-T-T-K-K-L-T-V-V-T-Q-F-E-T-S-G-A-I-N-R-Y-Y-V-Q-D-G-V-T-F-Q-Q-P-N-A-E-L-G-S-
CNI:T6
CNI:T6A
CNI:T6B
CNI:G6
CNI:G7
CNI:G7A
CNI:C4
CNI:C6
CNI:C7
CNI:C8
Y-S-G-N-E-L-N-D-D-Y-C-T-A-E-E-A-E-F-G-G-S-S-F-S-D-K-G-G-L-T-Q-F-( )-A-T-S-G-G-M-
M:T1
CNI:G7B
CNI:C9
CNI:C9A
CNI:C10
V-L-V-M-S-L-W-D-D-Y-Y-A-N-M-L-W-L-D-S-T-Y-P-T-N-E-T-S-S-T-P-G-A-V-R-G-S-C-S-T-S-S-G-V-P-A-
M:C4
CNVI
CNII
CNVII
CNII:T1
CNII:T2
CNII:G1
Q-V-E-S-Q-S-P-N-A-K-V-T-F-S-N-I-K-F-G-P-I-G-S-T-G-N-P-S-G-G-N-P-P-G-G-N-P-P-G-(T7, S3, Z, P3,
CNII:T4
CNII:G2
CNIIG2:C1
CNIIG2:C2
CNII:G2:Th1
CNII G2:Th3
G, A, H, R2 . . . . .)-Y-G-Q-C-G-G-I-G-Y-S-G-P-T-V-C-A-S-G-T-T-C-Q-V-L-N-P-Y-Y-S-Q-C-L
CNIIG2:C3
CNII G2:C4
CNII:G2:C5
CNII G2:Th4
CNIIG2:Th5B
CNII G2:Th6
CNII G2:Th3B
CNIIG2:Th5

which assigns the peptide to the N-terminus of CN I. Since this peptide was generated by a split at a Glu(or Asp)–Met peptide bond all fragments but CN VI can be excluded in the alignment.

The second methionine-containing peptide G1 was, from its elution position on the Sephadex column, estimated to contain about 60 amino acid residues. Sequencing revealed that it consisted of a mixture of two peptides in the approximate ratio 2:1. Fortunately they were both derived from the same part of the polypeptide and could therefore be unambiguously sequenced without further purification. The dominant component resulted from cleavage of the Glu 15–Phe 16 bond of fragment CN V, whereas the other was due to a less typical split at the Asp 21–Val 22 bond in the same fragment. The sequence following the methionine residue identified in the 23rd cycle of the dominant peptide corresponded to the N-terminal sequence of CN IV, thereby establishing the alignment of CN V and CN IV.

From a chymotryptic digest of the RCM-protein labelled by *S*-carboxymethylation of methionyl residues with iodo[^{14}C]acetic acid, 4 radioactive peptides were isolated. The largest of these (M:C3) had a composition corresponding to the sequence from residue 32 in CN IV to residue 3 in CN I. Due to the size of the peptide it was not subjected to Edman degradation. The peptide M:C4 gave the sequence Ala–Asn–Met–Leu–Trp, which unambiguously places CN VI–CN VII at the N-terminus of CN II. Degradation of M:C1 resulted in the sequence Leu–Met–Ala–Ser–Asp–Thr–Thr–Tyr. The sequence after the methionyl residue is identical with the N-terminal sequence of CN V. Since the C-terminal sequence of CN III is Leu–Tyr–Leu–Met, the alignment CN III–CN V would give a chymotryptic peptide with the sequence obtained for M:C1. Degradation of M:C2 revealed that it was a mixture of two peptides. The dominant peptide gave the sequence Phe–Val–Ser–Met–Asp–Ala–Gly–Gly–Val... which supports the alignment CN V–CN IV. The contaminant, which gave the sequence Thr–Leu–Leu–Gly–Asn–Glu–Phe, corresponds to the sequence from residue 10 and onwards in CN V.

Tryptic digestion of the methionyl-*S*-[^{14}C]carboxymethylated protein gave a highly labelled peptide containing the last 6 residues of CN I, CN VI and the first 20 residues of CN II (M:T1).

4. CONCLUSION

The main cellobiohydrolase (CBH I) from *T. reesei* has previously been reported to consist of a single polypeptide chain containing about 370 amino acid residues [4,6]. However, the results of the primary structure study reported here indicate that the polypeptide chain contains about 490 residues which is in good agreement with a reported M_r of about 58000 [22].

The previously reported absence of a free N-terminus [7] has been shown to be due to the presence of a pyroglutamyl residue [2]. Structural studies on other cellulolytic enzymes from *T. reesei* have revealed the same type of blockage in a second cellobiohydrolase (CBH II) [2] and in the main 1,4-β-glucan glucanohydrolase [23].

In the sequence established so far glucosamine was found to be attached to 3 of the 5 asparagine residues that occurred in the 'restricted sequences' [24] of the Asn–X–Ser/Thr type, which usually serve as recognition sites for the glycosylating enzymes [25]. The reason why two of them remain unglycosylated (Asn 64 and 426) is not obvious but might reflect inaccessibility to the enzyme.

The main part of the neutral carbohydrate has been reported to be bound through an average of 16.7 *O*-glycosidic linkages to serine and threonine per molecule, based on a molecular size of 377 amino acid residues [4]. This would correspond to about 22 linkages for a molecular size of 490 residues.

The results of the present study indicate that the main part of the neutral carbohydrate is bound in a short region (residues 440–458) near the C-terminus of the polypeptide chain, in which 10 out of 19 residues are either serine or threonine. Noteworthy is the proline-rich repetitive sequence (residues 425–439) preceding the carbohydrate-containing region:

425
-Gly-Asn-Pro-Ser-Gly-Gly-Asn-Pro-Pro-Gly-

439
-Gly-Asn-Pro-Pro-Gly-

Recently the nucleotide sequence of the CBH I gene has been reported [26] and the agreement with our amino acid sequence is excellent. Only two

minor differences have been observed, in position 302 of the amino acid sequence Asp, in the nucleotide sequence Asn and in positions 437–438 Pro–Pro and in the nucleotide sequence Arg. The first difference might be due to a deamidation, while the second is more difficult to explain. It now, however, seems safe to conclude that through the combined efforts on the DNA and the protein side the complete amino acid sequence of a cellulase has been determined.

ACKNOWLEDGEMENTS

This study has been supported by grants from the Swedish Natural Science Research Council and the Swedish Medical Science Research Council. We also want to thank Dr Hans Bennich and Dr David Eaker for guidance and encouragement throughout the work.

REFERENCES

[1] Fägerstam, L.G. and Pettersson, L.G. (1979) FEBS Lett. 98, 363–367.
[2] Fägerstam, L.G. and Pettersson, L.G. (1980) FEBS Lett. 119, 97–100.
[3] Montenecourt, B.S., Kelleher, T.J., Eveleigh, D.E. and Pettersson, L.G. (1980) Biotech. Bioeng. Symp. no.10, 15–26.
[4] Gum, E.K. jr and Brown, R.D. jr (1976) Biochim. Biophys. Acta 446, 371–386.
[5] Fägerstam, L., Håkansson, U., Pettersson, G. and Andersson, L. (1977) in: Proc. Bioconvers. Symp. (Ghose, T.K. ed.) New Delhi, 165–178.
[6] Berghem, L.E.R., Pettersson, L.G. and Axiö-Fredriksson, U.-B. (1975) Eur. J. Biochem. 53, 55–62.
[7] Gum, E.K. jr and Brown, R.D. jr (1977) Biochim. Biophys. Acta 492, 225–231.
[8] Crestfield, A.M., Moore, S. and Stein, W.H. (1963) J. Biol. Chem. 238, 622–627.
[9] Gross, E. and Witkop, B. (1961) J. Am. Chem. Soc. 83, 1510–1511.
[10] Gross, E. and Witkop, B. (1962) J. Biol. Chem. 237, 1856–1860.
[11] Podell, D.N. and Abraham, G.N. (1978) Biochem. Biophys. Res. Commun. 81, 176–185.
[12] Jentoft, N. and Dearborn, D.G. (1979) J. Biol. Chem. 254, 4359–4365.
[13] Vithayathil, P.J. and Richards, F.M. (1960) J. Biol. Chem. 235, 2343–2351.
[14] Wilkinson, J.M. (1969) FEBS Lett. 4, 170–172.
[15] Edman, P. and Begg, G. (1967) Eur. J. Biochem. 1, 80–91.
[16] Vasseur, E. (1948) Acta Chem. Scand. 2, 693.
[17] Tarr, G.E., Beecher, J.F., Bell, M. and McKean, D.J. (1978) Anal. Biochem. 84, 622–627.
[18] Klapper, D.G., Wilde, C.E. iii and Capra, J.D. (1978) Anal. Biochem. 85, 126–131.
[19] Mendez, E. and Lai, C.Y. (1975) Anal. Biochem. 68, 47–53.
[20] Narita, K. and Titani, K. (1968) J. Biochem. (Tokyo) 63, 226–241.
[21] Schroeder, W.A., Shelton, J.B. and Shelton, J.R. (1969) Arch. Biochem. Biophys. 130, 551–556.
[22] Leach, B.S., Collawn, J.F. and Fish, W.W. (1980) Biochemistry 19, 5741–5747.
[23] Pettersson, G., Fägerstam, L., Bhikhabhai, R. and Leandoer, K. (1981) SPCI Int. Symp. Wood and Pulping Chem. (Stockholm) 3, 39–42.
[24] Hunt, L.T. and Dayhoff, M.O. (1970) Biochem. Biophys. Res. Commun. 39, 757–765.
[25] Pless, D.D. and Lenarz, W.J. (1977) Proc. Natl. Acad. Sci. USA 74, 134–138.
[26] Shoemaker, S., Schweickart, V., Ladner, M., Gelfand, D., Kwok, S., Myambo, K. and Imris, M. (1983) Biotechnol. Oct., 691–695.