The cellulolytic enzymes of *Trichoderma reesei* as a system of homologous proteins

Cyanogen bromide peptides and partial sequence of endoglucanase II

Ramagauri Bhikhabhai and L. Göran Pettersson

Institute of Biochemistry, University of Uppsala, Biomedical Center, PO Box 576, S-75123 Uppsala, Sweden

Received 30 December 1983

The partial amino acid sequences of cyanogen bromide peptides of endoglucanase II, 1,4-β-glucan glucanohydrolase (from *Trichoderma reesei*) were determined by automated Edman degradation with liquid and solid phase techniques. Computer programs for finding distant relationships between the two amino acid sequences were used to compare the partial sequences of endoglucanase II with the complete sequence of cellobiohydrolase I. Homologous regions were found for all the cyanogen bromide peptides.

Cellulase *Endoglucanase* Trichoderma reesei *Amino acid sequence* *Homology*

1. INTRODUCTION

The fungus *Trichoderma reesei* is probably the most efficient cellulolytic organism known so far. This is due to its ability to produce large quantities of cellulases, mainly of two types, 1,4-β-glucan glucanohydrolases and 1,4-β-glucan cellobiohydrolases. This classification is based on experiments using different cellulose derivatives as substrates and it is doubtful whether it defines their specificity on native cellulose, where synergistic effects between endoglucanases and cellobiohydrolases and between two different cellobiohydrolases is a dominating phenomenon. The difference between the two categories of enzymes is nevertheless evident. The endoglucanases attack cellulose derivatives and have a low affinity for crystalline cellulose, while the opposite is true for the cellobiohydrolases. This work has been undertaken as a first approach to elucidate the structural differences between the two classes of cellulases. We have thus compared the partial amino acid sequences of the CNBr fragments of an endoglucanase (endo II) with the complete sequence of a cellobiohydrolase (CBH I) using a computer program to evaluate homologies.

Due to the technical potentialities of cellulases many groups have started cloning experiments [2,3] and it is our hope that partial amino acid sequences will facilitate this research and that the complete amino acid sequence of the enzyme will be determined in the near future from the nucleotide sequence of the gene. We are convinced that the combination of the conventional technique and the recombinant DNA technique will be the method of choice for sequencing the highly glycosylated cellulases.

Abbreviations: endo II, 1,4-β-glucan glucanohydrolase (EC 3.2.1.4); CBH I, 1,4-β-glucan cellobiohydrolases (EC 3.2.1.91); HPLC, high-performance liquid chromatography; RCM, reduced and carboxymethylated

2. MATERIALS AND METHODS

2.1. *Materials*

Freeze-dried culture filtrates of *T. reesei* QM 9414 were a gift from Dr M. Mandels, US Army Natick Research and Development Command,

Published by Elsevier Science Publishers B.V.
00145793/84/$3.00

Natick, MA. Endo II was isolated and further purified by ion-exchange chromatography as in [4,5].

All chemicals and proteolytic enzymes were commercially obtained in the most pure form.

2.2. *Methods*

Reduction and carboxymethylation, CNBr cleavage [6,7], staphylococcal protease and carboxypeptidase A digestion were performed as in [1]. Deblocking of N-terminal amino acid by pyroglutamyl aminopeptidase was performed as in [8].

2.3. *Separation of peptides by reversed-phase HPLC*

Peptide separations were performed on an Altex model 110 chromatography system (two model 110 A pumps, model 410 solvent programmer, a Rheodyne model 7120 sample injector with a 230 μl sample loop, LDC Spectromonitor III and Gilson minicol TDC fraction collector).

Peptides were chromatographed on either a 4.5 × 250 mm Lichrosorb RP 8, 7μ or Spherisorb ODS 10μ column, equilibrated in 5 mM ammonium acetate buffer (pH 6.4). Peptides were eluted with a linear gradient of one of the following organic solvents: 70% ethanol, 70% 1-propanol or a mixture of 1-propanol:methoxyethanol:water (70:5:25). After pooling the appropriate fractions spectra were recorded and aliquots were taken for amino acid analysis.

2.4. *Amino acid analysis*

Samples were hydrolyzed in evacuated and sealed tubes for 24 h at 110°C with 6 M HCl containing 2 mg/ml phenol. After evaporation, the hydrolysates were analyzed with a Durrum-500 analyzer. Serine and threonine values were calculated using the standard recovery factors of 0.90 and 0.96, respectively.

2.5. *Sequence determination*

Manual Edman degradation was performed as in [9] but phenylthiohydantoin (PTH) derivatives of the released amino acid were identified by HPLC as described below.

Solid-phase sequencing was carried out in an LKB 4020 sequencer. Peptides were coupled by either the homoserine lactone method [10] or the carbodiimide method [11]. For liquid-phase sequencing, a Beckman 890 C sequencer (Beckman Instruments, Palo Alto) with Beckman program 122974 was used.

The PTH derivatives of amino acids were analyzed on a reversed-phase column (0.4 × 30 cm) of Nucleosil C18 (5μ) using a linear gradient elution method with acetonitrile as organic solvent and 5 mM sodium acetate (pH 5) as the aqueous medium.

Concentration of aqueous medium, gradient, time and temperature were adjusted to obtain better separation of all PTH amino acids. The conditions most often used for a complete run of 27 min were: a 12-min linear gradient from 20 to 48% acetonitrile at 35°C.

2.6. *Homology calculations*

Alignment of sequences and scores for best alignment were obtained with a computer program based on the algorithm in [12]. Weights (scores) for amino acid pairs were taken from the table in [13] giving the relatedness odds matrix for 256 PAM (accepted point mutations per 100 residues) suitable for the detection of distant relationships. Each value was increased by 78 in order to avoid negative scores, as required by the algorithm. For each alignment a total score was calculated. The above chance similarity of two sequences, also called the alignment score, was the deviation of the total score for the comparison of the two real sequences from the mean score of the comparisons of 50 random sequences of the same composition [12]. The alignment score was expressed in standard deviation units (SD).

To find the best fit of a short sequence when compared to a much longer sequence, the following procedure was used. The short sequence was compared to segments of equal length along the entire longer sequence and values of total score were plotted against residue number of the longer sequence. For the short stretches of the longer sequence where only composition was available, composite sequence was used. In this each position contained all residues occurring in the non-sequenced part. The weight for a position in the comparison was then calculated as the mean of all possible pairs. Finally an alignment score was calculated for the alignment with highest total scores in the study.

Alignment score, expressed in SD units, estimates the relatedness of two sequences: >10 SD units indicates strong homology; 3–10 SD units indicates a distant but still significant evolutionary relationship; 2–3 SD units indicates a potential homology [14].

3. RESULTS

3.1. *Amino acid composition, N-terminal and C-terminal residues*

The M_r of the peptide chain was estimated to be 47000 by normalizing the amino acid composition (table 1) of the RCM protein to 9 lysine residues. The M_r of the whole glycoprotein was difficult to determine, since the carbohydrate composition is unknown. Carbohydrate analysis indicated the presence of mannose, glucose and glucosamine. Since no N-terminal amino acid was detected by the Edman method, the N-terminus seemed to be blocked. After treatment with calf liver pyroglutamyl aminopeptidase, automated sequence analysis on the RCM protein gave 24 residues from the N-terminal. Incubation with carboxypeptidase A gave leucine as the C-terminal.

3.2. *Preparation and analysis of peptides*

The RCM protein was subjected to CNBr cleavage. The digest was separated into 7 peaks by gel filtration (fig.1). The peaks were denoted

Table 1

Amino acid composition of reduced and carboxymethylated protein, cyanogen bromide peptides and two peptides of S-protease digest

	CNBr peptides of endo II										Endo II			S-protease peptides	
	CN II	CN III	CN IV	V B	V C	V D	VI A	VI B	VII A	VII B	C	B	A	GL VII D	GL IX H
CM-Cys	3.1(3)	2.2(2)	3.7(4)	3.5(4)	3.9(4)	2.2(2)	–	0.9(1)	–	–	20	20	5.0	3.2(3)	–
Asp	13.6(14)	7.9(8)	7.9(8)	8.6(9)	3.1(3)	3.3(3)	2.2(2)	4.2(4)	1.2(1)	–	52	55	12.7	7.0(7)	1.2(1)
Thr	16.1(16)	6.9(8)	6.5(6)	4.3(4)	4.0(4)	6.8(7)	–	–	–	–	44	44	10.2	3.8(4)	–
Ser	19.0(19)	6.9(7)	5.0(5)	2.8(3)	4.1(4)	4.0(4)	6.7(7)	4.1(4)	–	2.1(2)	55	58	13.5	1.7(2)	–
Glu	5.8(6)	3.7(4)	2.6(3)	7.0(7)	3.6(4)	4.5(4)	1.4(1)	4.1(4)	1.0(1)	–	34	30	7.0	2.2(2)	2.1(2)
Pro	8.0(8)	3.3(3)	1.0(1)	1.2(1)	1.1(1)	3.1(3)	1.8(2)	1.0(1)	–	–	20	24	5.6	1.1(1)	–
Gly	12.4(12)	7.6(7)	6.0(6)	7.3(7)	6.3(6)	4.2(4)	3.6(3)	3.0(3)	–	2.3(2)	50	47	11.0	3.4(3)	1.2(1)
Ala	5.0(5)	4.3(4)	3.8(4)	4.0(4)	–	1.8(2)	–	1.1(1)	–	1.1(1)	21	25	5.9	2.0(2)	–
Val	2.1(2)	3.0(3)	3.6(4)	1.2(1)	–	3.9(4)	2.0(2)	1.1(1)	–	–	17	19	4.5	1.9(2)	1.0(1)
Met	–	–	–	–	–	–	–	–	–	–	–	9	2.2	0.3(1)	0.7(1)
Iso	4.0(4)	2.1(2)	1.2(1)	–	1.0(1)	–	–	–	–	–	8	12	2.9	1.0(1)	–
Leu	3.3(3)	2.5(2)	2.0(2)	1.2(1)	1.3(1)	2.4(2)	3.1(3)	6.8(7)	–	1.1(1)	22	23	5.4	1.2(1)	2.1(2)
Tyr	3.9(4)	3.2(3)	1.8(2)	2.7(3)	3.7(4)	1.6(2)	2.8(3)	1.1(1)	1.0(1)	–	23	23	5.3	2.0(2)	1.4(1)
Phe	2.2(2)	1.5(2)	0.8(1)	0.8(1)	–	–	–	1.1(1)	–	–	7	9	2.2	0.9(1)	–
His	2.3(2)	0.9(1)	1.0(1)	0.8(1)	–	1.0(1)	–	–	–	–	6	6	2.0	0.9(1)	–
Lys	1.2(1)	2.0(2)	1.0(1)	–	0.9(1)	3.0(3)	–	1.0(1)	–	1.0(1)	10	9	2.1	1.0(1)	1.0(1)
Trp[a]	–	–	–	–	–	–	–	–	–	–	–	7	1.6	2.8(3)	–
Arg	2.5(2)	0.6(1)	–	0.8(1)	–	–	0.9(1)	–	–	–	5	7	1.8	1.0(1)	–
Hser	(1)	(1)	(1)	(1)	0	–	(1)	(1)	(1)	(1)	8	–	–	–	–
Glucos-amine	+	+	+	+	–	–	–	–	–	–	+	+	+	0.7(1)	–
Total	104	59	50	48	33	41	25	30	4	8	402	420	100	38	10

[a] Determined spectrophotometrically

Results are given as residues per molecule. Values in parentheses are nearest integer. A, mole percent amino acid composition of endo II; B, amino acid composition of endo II normalized to lysine = 9; C, sum of all CNBr fragments

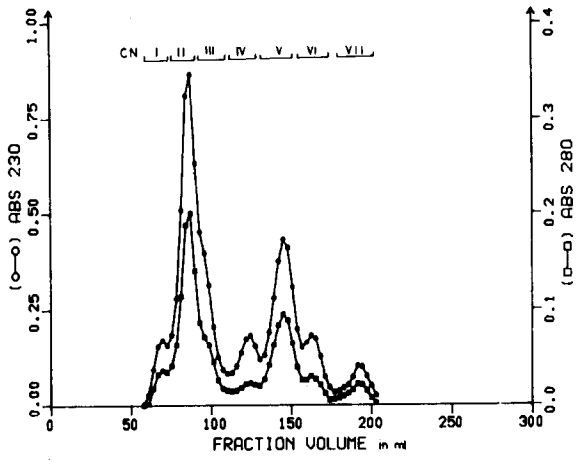

Fig.1. Elution profile of the CNBr-generated peptides of endo II. The column of Sephadex G-50 superfine (150 × 1.5 cm) was in 0.05 M HCl acid. Fractions of 2.9 ml (95 drops) were collected at a flow rate of 12 ml/h. (●) A_{230}, (■) A_{280}.

CN I–CN VII in order of elution. CN I and CN II were further purified on a Sephadex G-75 column. CN III was rechromatographed on G-50 and showed the same amino acid composition as CN I. The identity of the two peptides was further supported by the fact that CN III and CN I had the same elution positions when chromatographed in an alkaline medium. CN IV–CN VII were further fractionated by HPLC. The subfractions obtained were denoted by capital letters. In many cases poorly resolved double peaks with identical amino acid compositions were obtained, which may indicate heterogeneity in glycosylation or incomplete conversion of homoserine lactone to homoserine. A total of 10 peptides were recovered. When several peptides had the same amino acid composition, only one of them is reported (table 1). Two peptides, CN V C and CN V D did not contain homoserine. One of them is presumed to be the C-terminal peptide and the other must be the result of an atypical cleavage. Three peptides, namely, CN II, CN IV and CN V B contained carbohydrate as indicated by the presence of

Table 2

Alignment scores of CNBr peptides of endo II, which gave high total score values in the computer runs, and residue numbers of CBH I which aligned with these peptides

Peptides of endo II	Number of amino acids compared	Residue no. of CBH I	Total score	Alignment score
CN II	25	15– 39	2282	3.0
		38– 62	2325	2.6
		370–393	2350	3.3
CN III	20	49– 69	1683	3.2
		209–228	1932	6.4
GL VII D	17	38– 55	1679	2.1
		258–274	1703	3.1
	11	38– 49		3.0
CN V B	20	145–167	2287	6.5
		318–338	1829	2.3
CN V C	27	235–265	2911	2.5
(C-terminal)		460–486	3822	8.2
CN VI	51	90–141	5483	11.0
CN V D	26	1– 26	2529	3.7
(N-terminal)		147–172	2301	2.2
		301–326	2313	3.5

glucosamine. To obtain overlapping sequences, an S-protease digest was fractionated and purified in the same way as the CNBr peptides. Only two methionine-containing peptides are reported here (GL VII D and GL IX H). Their amino acid compositions are shown in table 1.

3.3. *Partial sequences of CNBr peptides*

Fig.2 shows the partial amino acid sequences of the 10 CNBr peptides resulting from cleavage at the 9 methionine residues present in the intact protein. From its amino acid composition, CN V D was assigned to the N-terminus of the molecule. By coupling this peptide to aminopropyl glass with the carbodiimide method and treatment with trypsin, the sequence after Lys-20 could be determined. In many cases, repetitive yields were low, probably due to the fact that the peptides are glycosylated. The sequence of S-protease peptide GL VII D provided the overlap on the N-terminal side of CN IV

Fig.2. Sequence determination of CNBr and S-protease peptides of endo II. Residues are identified by liquid-phase sequencer (——→), solid-phase sequencer (——→) and manual Edman degradation (– – –).

peptide, whereas GL IX H bridged CN VI A and CN VI B.

3.4. *Homology calculations*

The partial sequences of the CNBr peptides of endo II were compared with the sequence of CBH I. Positions for the best alignments and their total scores and SD values are reported in table 2. Fig.3 shows the alignment of endo II peptides along the complete sequence of CBH I. CN VI A and CN VI B were treated as a single peptide (CN VI) since they could be aligned. The best alignment score was used to allocate the peptides. For two peptides (CN V C and CN VI) overwhelming evidence for homology was found (alignment score: 8.2 and 11.0). In two cases (CN II and GL VII D), alternative alignments with comparable score were obtained. In these cases circumstantial evidence was used to decide the location. CN V D was thus assumed to be the N-terminal peptide. CN II had one possible location in the N-terminal region which, however, was already occupied by CN V D. The second alternative (370–393) was thus assumed for CN II. GL VII D was located on the basis of the amino acid compositions of the total set of S-protease peptides and CNBr peptides and partial sequences of some of S-protease peptides (unpublished). Here it should be observed that the occurrence of two possible alignments with rather high score for a peptide may reflect the presence of internal duplication.

4. DISCUSSION

Endo II and CBH I are extracellular enzymes. Both have 10–12 disulphide bridges, which give the protein stability and a compact structure. Both have pyroglutamic acid as the N-terminal amino acid and leucine as C-terminal. By visual inspection of the partial sequences of the CNBr peptides of endo II, strong homology was observed in one region beginning around residue 100 and also in the C-terminal region. Although the rest of the peptides did not show any apparent homology they could nevertheless be aligned along the CBH I sequence using the computer programs. According to our results the two enzymes do not seem to have any common antigenic determinants [5]. Recently, however, results have been presented which con-

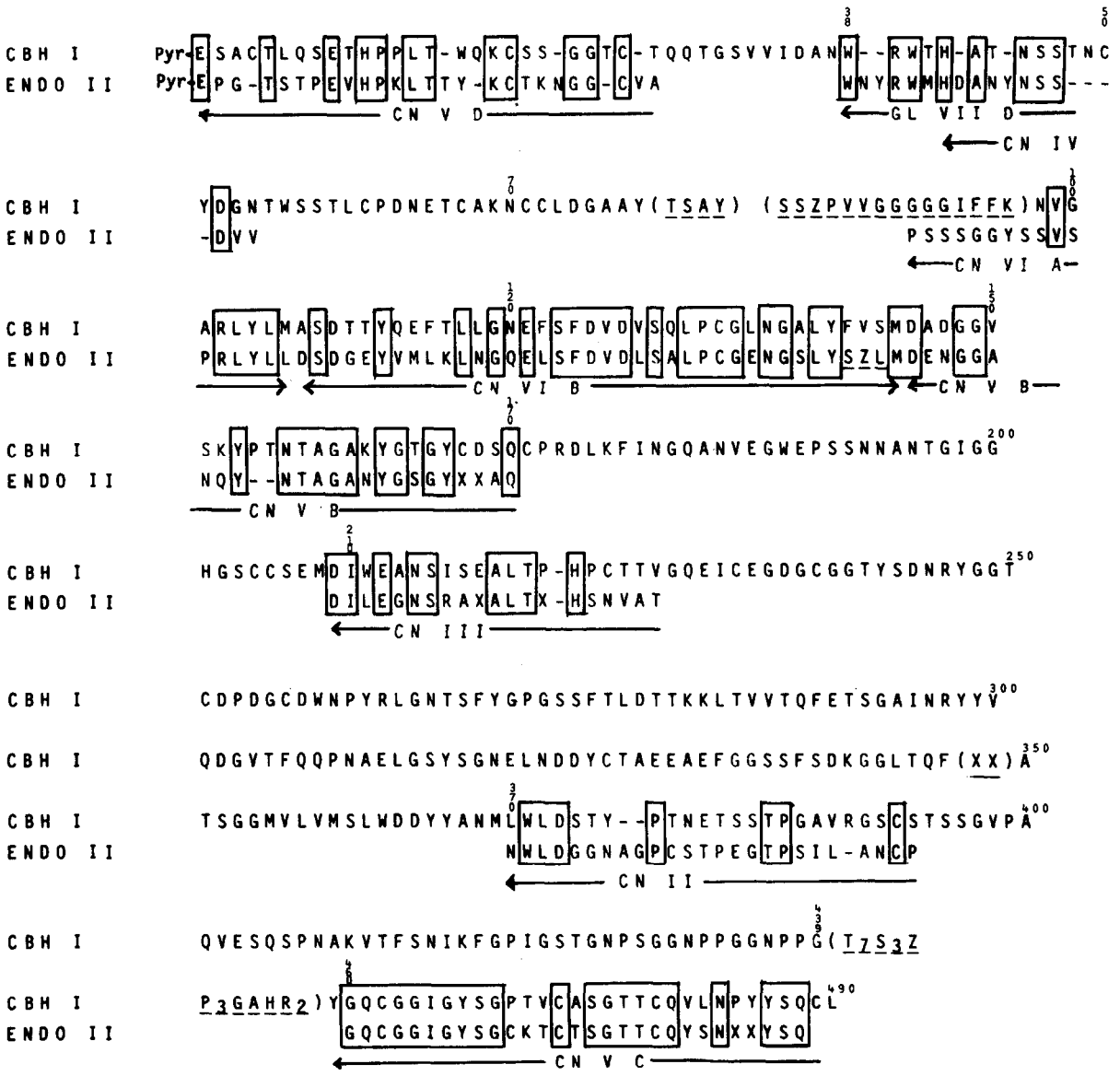

Fig.3. Alignment of CNBr peptides of endo II with complete sequence of CBH I. Sequence identities are boxed. A gap (−) is inserted to obtain homology. Underlined (---) are unsequenced fragments with known amino acid composition.

tradict this statement [15]. Although we only have partial sequences of endo II it may be concluded that the extent of homology between the enzymes is so evident that the two enzymes certainly have a common origin and that the chain folding is similar.

The available partial sequences of other saccharifying enzymes from different organisms were also compared with the cellulases. Comparison was done in the same manner as previously described. Fig.4 shows some sequences where high total scores and alignment scores were obtained. No amino acid sequence data for the β-glucosidases from *T. reesei* are available, but the amino acid sequence of a tryptic glycopeptide (63 residues) from the active site of a β-glucosidase

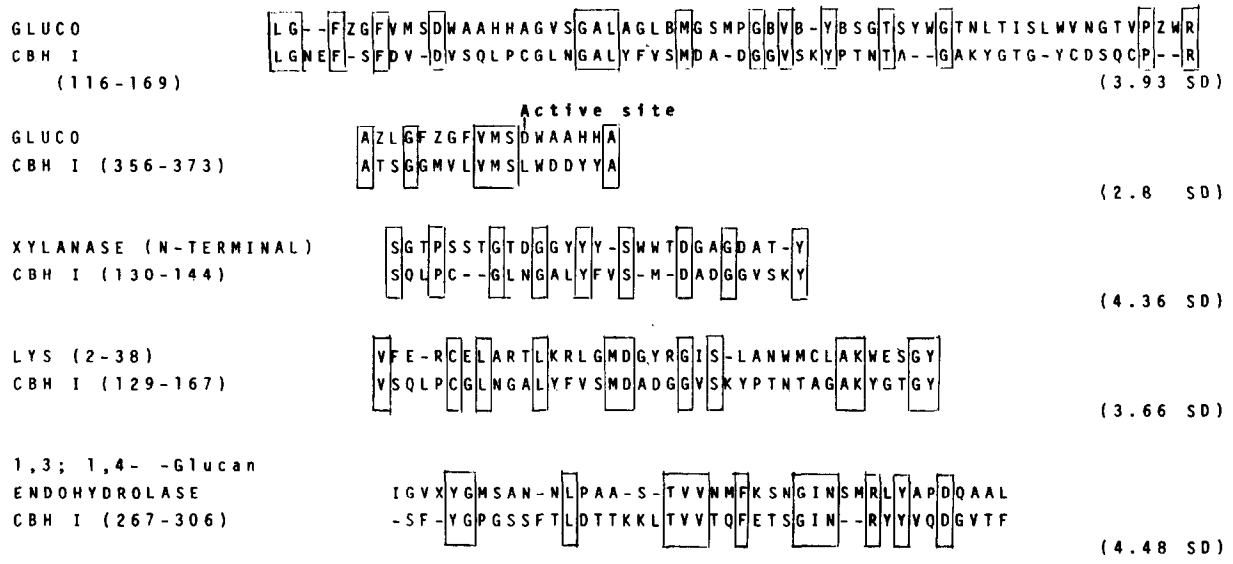

Fig.4. Alignment of sequences from different saccharifying enzymes with CBH I and SD values in parentheses. Sequence identities are boxed. GLUCO, β-glucosidase; LYS, human lysozyme.

from *Aspergillus wentii* [16] gave high total scores with two regions in CBH I. Only one of them (CBH I 116–169) gave a high alignment score. Visual inspection of the second alternative (350–412), revealed that the first 18 residues from the N-terminus, including a catalytically active aspartic acid residue, showed similarity (fig.4). It should be observed that the catalytically active aspartic acid residue in the β-glucosidase is replaced by a leucine residue in CBH I but two aspartic acid residues are only two positions ahead in the CBH I sequence. The question is whether one of them might be catalytically active.

Endo II has been reported to show xylanase activity and hence it was compared with the available N-terminal sequence of a true xylanase from *Schizophyllum commune* [17], and a potential homology was found in the region from position 130 and onwards of CBH I. Lysozyme was the first saccharifying enzyme to be studied in detail. The N-terminal region of lysozyme showed similarity to a region in CBH I from position 129 onwards. It is interesting to note that β-glucosidase, xylanase and lysozyme show homology around the same region of both CBH I and endo II. The complete DNA sequence of the CBH I gene has been published [2] and an intron was found between residue 132 and 133. More sequence data are necessary to make any comment on the relation between the presence of an intron and similarity in amino acid sequence in the region of the exon-intron junction. The N-terminal of 1,3;1,4 β-glucan endohydrolase from germinating barley [18] showed significant evolutionary relationship with CBH I in the region 267–306. The sequence from Glu-33 to Tyr-51 of endo-β-1,4 glucanase from *S. commune* [19] has been reported. It shows homology with the active site sequences of various hen egg-white type lysozymes. No homology of this type was observed with CBH I or endo II.

Sequence work both on the protein and DNA side is in progress on cellulases and this work indicates that lots of interesting findings regarding structure and genetic relations among cellulases and other saccharifying enzymes are to be expected.

ACKNOWLEDGEMENTS

This work has been supported by grants from the Swedish Natural Science Research Council. We thank the staff of the protein laboratory of Wallenberg Laboratory, and Dr Åke Engström for sequencing the peptides on a liquid sequencer. We also thank Dr Lars Rydén for guidance and

valuable discussions in homology section and Dr David Eaker for linguistic revision of the manuscript.

REFERENCES

[1] Fägerstam, L.G., Pettersson, L.G. and Engström, J.Å. (1984) FEBS Lett. 167, 309–315.
[2] Shoemaker, S., Schweickart, V., Ladner, M., Gelfand, D., Kwok, S., Myambo, K. and Innis, M. (1983) Biotechnology Oct. 1983, 691–695.
[3] Teeri, T., Salovuori, I. and Knowles, J. (1983) Biotechnology Oct. 1983, 696–699.
[4] Håkansson, U., Fägerstam, L.G. and Pettersson, L.G. (1979) Biochem. J. 179, 141–149.
[5] Fägerstam, L.G. and Pettersson, L.G. (1979) FEBS Lett. 98, 363–367.
[6] Gross, E. and Witkop, B. (1961) J. Am. Chem. Soc. 83, 1510–1511.
[7] Gross, E. and Witkop, B. (1962) J. Biol. Chem. 237, 1856–1860.
[8] Podell, D.N. and Abraham, G.N. (1978) Biochem. Biophys. Res. Commun. 81, 176–185.
[9] Edman, P. and Henschen, A. (1975) in: Protein Sequence Determination (Needleman, S.B. ed.) 2nd edn, pp.232–279, Springer-Verlag, Berlin.
[10] Horn, M.J. and Laursen, R.A. (1973) FEBS Lett. 36, 284–288.
[11] Salnikow, J., Lehmann, A. and Wittmann-Liebold, B. (1981) Anal. Biochem. 117, 433–442.
[12] Needleman, S.B. and Wünsch, C.D. (1970) J. Mol. Biol. 48, 443–453.
[13] Dayhoff, M.O. (1972) Atlas of Protein Sequencing and Structure (5th edn) pD-3, National Biomedical Research Foundation, Washington, DC.
[14] Barker, W.C., Ketcham, L.K. and Dayhoff, M.G. (1978) J. Mol. Evol. 10, 265–281.
[15] Shoemaker, S., Watt, K., Tsitovsky and Cox, R. (1983) Biotechnology Oct. 1983, 687–690.
[16] Bausse, E. and Legler, G. (1980) Biochim. Biophys. Acta 626, 459–465.
[17] Paice, M.G., Jurasek, L., Carpenter, M.R. and Smillie, L.P. (1978) Appl. Environ. Microbiol. 6, 802–808.
[18] Woodward, J.R., Morgan, F.J. and Fincher, G.B. (1982) FEBS Lett. 138, 198–200.
[19] Yaguchi, M., Roy, C., Rolin, C.F., Paice, M.G. and Jurasek, L. (1983) Biochem. Biophys. Res. Commun. 116, 408–411.