Volume 321, number 2,3, 135–139 FEBS 12340 April 1993

Amino acid sequence similarities between low molecular weight endo-1,4-β-xylanases and family H cellulases revealed by clustering analysis

Anneli Törrönen[a,*], Christian P. Kubicek[a] and Bernard Henrissat[b]

[a]*Institut für Biochemische Technologie und Mikrobiologie, Technische Universität Wien, Getreidemarkt 9, A-1060 Wien, Austria* and [b]*Centre de Recherches sur les Macromolécules Végétales, CNRS, BP 53X, F-38041 Grenoble, France*

Received 17 February 1993; revised version received 10 March 1993

The amino acid sequences of seventeen family G xylanases and the two known family H cellulases have been compared by hydrophobic cluster analysis. A weak but significant similarity was demonstrated between these two families suggesting that these enzymes share the same molecular mechanism and catalytic residues and that they have related 3D folds. The major differences were found in the N-terminal regions.

Cellulase; Hydrophobic cluster analysis; Sequence similarity; Structural similarity; Catalytic residue

1. INTRODUCTION

During the last few years microbial polysaccharidases have been the subject of intense research because of many potential industrial applications for this class of enzymes [1]. Especially cellulases and xylanases are now characterised at both the enzymological as well as at the molecular level [2,3] and their grouping into families of structurally related proteins has been established on the basis of amino acid sequence similarities [4–7].

Wong et al. [8] have divided xylanases into two categories: (a) the low molecular weight, basic xylanases and (b) high molecular weight, acid xylanases. The low molecular weight basic xylanases are usually endo-1,4-β-xylanases with specific activity on xylan only, whereas the other group seems to contain endo-xylanases with cellulase activity. These groups roughly correspond to β-glycanase families G and F, respectively [4–7]. Some xylanases have also been reported to consist of several domains [6].

Saarilahti et al. [9] have determined the amino acid sequence of the cellulase CelS of *Erwinia carotovora* and found it was not related to other cellulases. Amino acid comparison of CelS with other enzymes revealed a very weak and local similarity with three family G xylanases from *Bacillus* sp. This similarity was, however, too limited to draw a definite conclusion on a possible relatedness between these enzymes [9]. The sequence of a cellulase from *Aspergillus aculeatus* was later determined [10] and found to be clearly related to *E. caratovora* CelS and led to the definition of family H of cellulases [6]. With the much larger number of sequences now available, distant relationships can be more easily inferred. We report here a detailed comparison by hydrophobic cluster analysis (HCA) [11] of 17 low molecular weight endo-1,4-β-xylanases of family G with the two known family H cellulases, and which reveals a distant but significant relationship between the two families.

Correspondence address: B. Henrissat, CERMAV-CNRS, BP 53X, F-38041 Grenoble, France. Fax: (33) (76) 54 72 03.

**Present address:* Cultor Ltd., Technology Center, SF-02460 Kantvik, Finland.

2. MATERIALS AND METHODS

The 17 xylanases and 2 cellulase sequences were from published literature (Table I). HCA plots with automatic hydrophobic cluster contouring were drawn using the plot program from Doriane S.A. (France) operating from a Macintosh computer. In the plots, residues VILMFWY are considered hydrophobic and are used to build the clusters. Amino acids are represented by their standard one-letter codes, except the following amino acids which represented by symbols: ★ for proline, ♦ for glycine, ▣ for serine, □ for threonine. HCA scores were calculated using the SUNHCA program [12] operating from a SUN Sparc station 2. The PCOMPARE program from PC/Gene (Genofit, Switzerland) on a PC-486 compatible microcomputer was used for calculation of alignment scores (S.D. = standard deviation) with the method of Needleman and Wunsch [13] using 50 random runs with Dayhoff's matrix [14]. The phylogenetic tree was calculated with the neighbour-joining method [15] implemented in the MSA program [16] and operated from a SUN Sparc station 2.

3. RESULTS

The HCA plots of 17 different xylanases belonging to family G were compared to those of two cellulases of family H to detect similarities in cluster shape, orientation and distribution over the sequences [11,17]. Because the hydrolytic mechanism of family G xylanases is known to proceed by a general acid catalysis promoted by two Glu residues [18,19], these residues were

Table I

Enzyme sequences for the comparison

Family	Enzyme	Origin	SWISS-PROT accession number or reference
G	XYNA	*Bacillus circulans*	P09850
G	XYNA	*Bacillus pumilus*	P00694
G	XYNA	*Bacillus subtilis*	P18429
G	XYNB	*Streptomyces lividans*	P26515
G	XYNC	*Streptomyces lividans*	P26220
G	XYN	*Streptomyces* sp.	[27]
G	XYN	*Clostridium acetobutylicum*	P17137
G	XYN	*Aspergillus kawachii*	[28]
G	XYN	*Aspergillus niger* var. *awamori*	[24]
G	XYN	*Schizophillum commune*	[27]
G	XYN	*Trichoderma viride*	[29]
G	XYN	*Trichoderma harzianum*	[30]
G	XYN1	*Trichoderma reesei*	[26]
G	XYN2	*Trichoderma reesei*	[26]
G	XYN[a]	*Ruminococcus flavefaciens*	P29126
G	XYN[a]	*Neocallimastix patriciarum*	P29127
G	XYN[b]	*Neocallimastix patriciarum*	P29127
H	FI-CMCase	*Aspergillus aculeatus*	P22669
H	CELS	*Erwinia carotovora*	P16630

[a]N-terminal domain; [b]C-terminal domain.

Table II

Comparison scores between selected family G xylanases and family H cellulases

Sequences	XYN1.tr	XYN2.tr	XYN.bp	XYN.ak	CEL.aa	CEL.ec
XYN1.tr	100	84	83	86	69	67
	(100)	(51)	(41)	(52)	(14)	(14)
XYN2.tr		100	82	84	63	64
		(100)	(49)	(45)	(15)	(11)
XYN.bp			100	81	64	64
			(100)	(38)	(18)	(14)
XYN.ak				100	72	68
				(100)	(20)	(19)
CEL.aa					100	85
					(100)	(29)
CEL.ec						100
						(100)

XYN1.tr = *T. reesei* xylanase 1; XYN2.tr = *T. reesei* xylanase 2; XYN.bp = *B. pumilus* xylanase A; XYN.ak = *A. kawachii* xylanase; CEL.ec = *E. carotovora* endoglucanase S; CEL.aa = *A. aculeatus* FI-carboxymethylcellulase. For each entry, the HCA score is given on top and the sequence identity rate (%) is given below in parentheses. The HCA score has been calculated [11,17] as follows for each cluster:

$$\text{HCA score} = 2CR/(RC_1 + RC_2) \times 100\%$$

where RC_1 and RC_2 are the number of hydrophobic residues in cluster 1 and 2, respectively. CR is number of hydrophobic residues in cluster 1 that are in correspondence with hydrophobic residues in cluster 2. The mean value obtained for all the clusters along the sequences compared gave the final HCA scores.

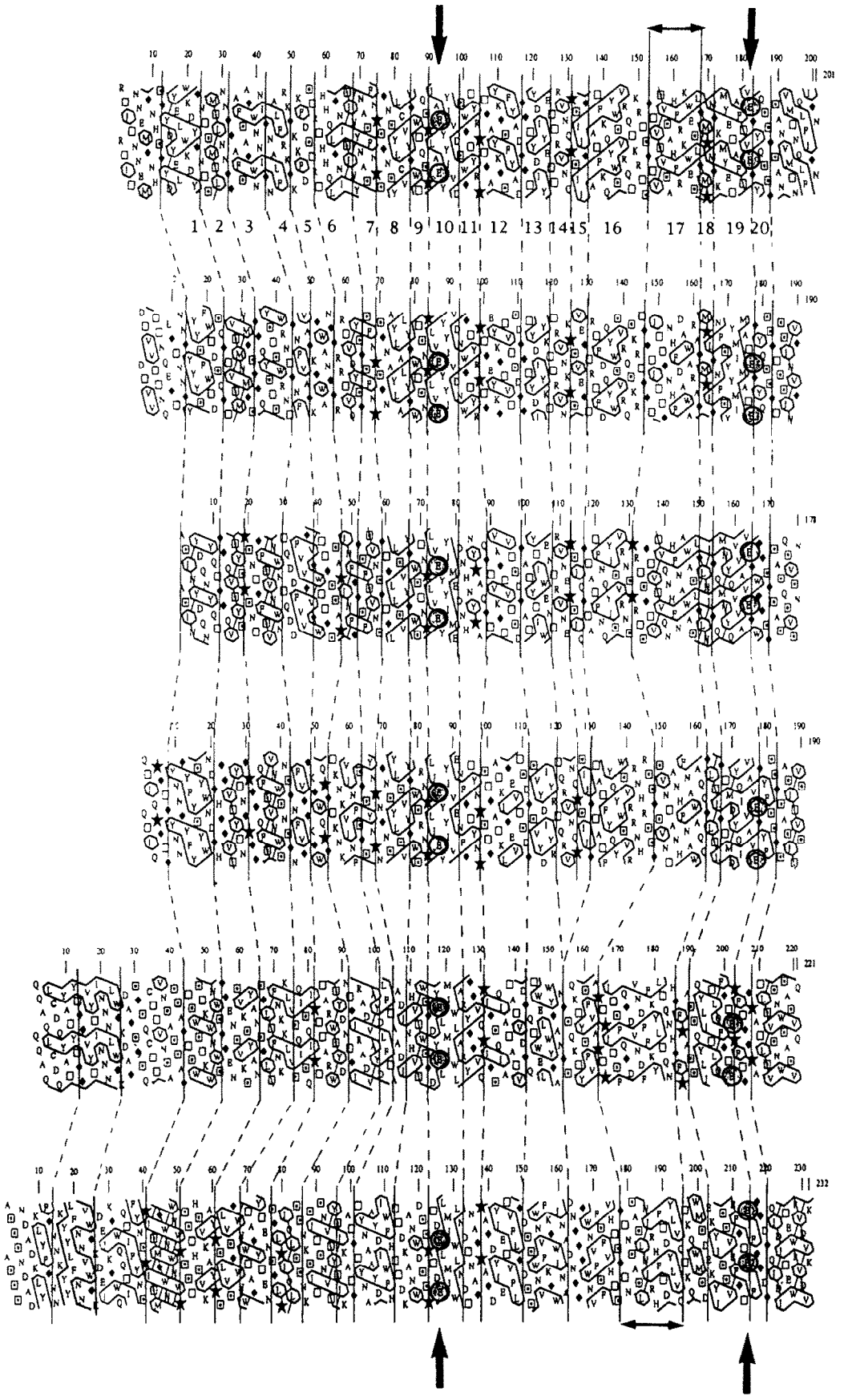
taken as anchor points for the HCA comparison. Fig. 1 shows the HCA plots of four selected xylanases and of the two family H cellulases.

The careful comparison of the 17 HCA plots of family F xylanases with those 2 of family G cellulases revealed a similar distribution of short hydrophobic clusters over a length of more than 100 residues suggesting that the two families might be related. A first assessment of the significance of this possible relatedness was done by calculation of the pairwise HCA homology scores [11,17] (Table II). HCA score values greater than 80% are commonly found in proteins with superimposable 3D structures (r.m.s. deviation between polypeptide chains less than 1.5 Å) while values of 65% are found among proteins having the same overall fold albeit with significant structural divergence (r.m.s. deviation between polypeptide chains ~ 2–3 Å).

A second assessment was done after progressive removal of the less conserved N-termini by the classical comparison (PCOMPARE) of the alignment score with those of computed after randomisation of one sequence (Table III). As expected, the scores were well above the commonly accepted level of significance (3 S.D.) within each family. The scores between members of different families were much lower, e.g. mostly in the 1.0–2.0 S.D. range, sometimes around 3 S.D. (Table III). However, significant scores (4.0 to 4.5 S.D.) were consistently obtained between truncated forms of *E. caro-*

→

Fig. 1. Hydrophobic cluster analysis (HCA) plots of four selected family G xylanases and two family H cellulases. In these plots, the amino acid sequence of the protein is written on a duplicated helical net and the clusters formed by contiguous hydrophobic residues are drawn [11]. The different clusters (or their constitutive parts) are delineated by vertical bars and are numbered (1–20). The α-helix region is indicated with a horizontal arrow. I, *Bacillus pumilus* xylanase; II, *Streptomyces lividans* xylanase B; III, *Trichoderma reesei* xylanase 1; IV, *Trichoderma reesei* xylanase 2; V, *Aspergillus aculeatus* FI-carboxymethylcellulase; VI, *Erwinia carotovora* endoglucanase S. Hydrophobic cluster analysis was conducted following published guidelines [11,17]. Typically the analysis starts by the observation of the segmentation: all the plots are dominated by short vertical and mosaic (zig-zag) clusters indicative of a high content in β-strands. The two catalytic Glu residues which have been determined for the *B. pumilus* xylanase [18,19] are next taken as anchor points (vertical arrows at the top and bottom of figure) for the comparison because they are expected to be better conserved during evolution. The topologically conserved residues in the other sequences (circled and shaded) are then identified by searching for Glu residues associated to hydrophobic clusters the shapes and positions of which are reminiscent of those in the *B. pumilus* enzyme. For instance, the typical shape of cluster 10 is particularly useful. The correspondences between the sequences are then iteratively extended to the rest of the plots using dotted lines and considering the parts of the clusters that are best conserved in terms of shape and position and not necessarily in terms of sequence identity [11,17].


I
II
III
IV
V
VI
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

tovora endoglucanase and xylanase B of *Streptomyces lividans* and validated the structural relatedness between the two families following the principle that if *A* is significantly related to *B* and *C*, then *B* is related to *C*. A phylogenetic tree was constructed according to the method of Saitou and Nei [15] to show the relative distances between the two families (Fig. 2). The two families seem to have diverged before speciation. Although family G members are more strongly related, some divergence between bacterial and fungal xylanases can also be seen. No similarity was detected between the enzymes under investigation (i.e. family G xylanases and family H cellulases) and the other families of cellulases and xylanases.

4. DISCUSSION

A careful sequence analysis performed on 19 sequences demonstrates that family G endo-1,4-β-xylanases and family H cellulases are distantly, but significantly related. Thus the earlier reported assumption [9],

Table III

Significance scores (SD) for the alignments between selected family G xylanases and family H cellulases

Alignments	Region used				
	Whole sequence	From cluster 2	From cluster 3	From cluster 5	From cluster 7
CEL.aa vs.					
CEL.ec	8.7	8.5	8.5	8.9	9.0
XYN.bp	1.6	2.5	1.4	1.8	1.4
XYNB.sl	2.4	1.6	1.4	1.2	1.1
XYN1.tr	0.8	2.0	1.3	2.2	1.8
XYN2.tr	0.7	2.0	0.7	2.0	2.1
CEL.ec vs.					
XYN.bp	1.1	2.2	3.0	2.2	3.1
XYNB.sl	1.3	4.3	3.8	4.0	4.5
XYN1.tr	0.01	1.0	0.9	0.5	0.1
XYN2.tr	0.9	1.2	0.8	0.03	0.4
XYN.bp vs.					
XYNB.sl	25	25	23	23	29
XYN1.tr	22	20	22	17	21
XYN2.tr	23	24	24	17	21
XYNB.sl vs.					
XYN1.tr	22	21	21	19	23
XYN2.tr	26	28	29	27	26
XYN1.tr vs.					
XYN2.tr	24	23	26	21	22

XYN1.tr = *T. reesei* xylanase 1; XYN2.tr = *T. reesei* xylanase 2; XYN.bp = *B. pumilus* xylanase A; XYNB.sl = *S. lividans* xylanase B; CEL.ec = *E. carotovora* endoglucanase S; CEL.aa = *A. aculeatus* FI-carboxymethylcellulase.

Scores (S.D. units) were calculated with PCOMPARE using Dayhoff's matrix [14] and 50 random runs. The values are the average of 3 independent calculations.

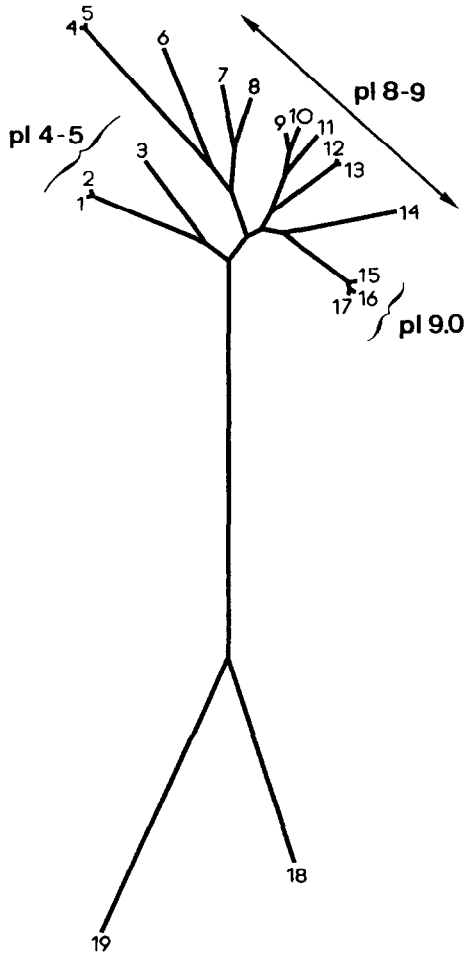


Fig. 2. Phylogenetic tree for family G xylanases and family H cellulases. 1, *Aspergillus awamori* xylanase; 2, *Aspergillus kawachii* xylanase; 3, *Trichoderma reesei* xylanase 1; 4, *Neocallimastix patriciarum* xylanase[a]; 5, *Neocallimastix patriciarum* xylanase[b]; 6, *Ruminococcus flavefaciens* xylanase[a]; 7, *Clostridium acetobutylicum* xylanase; 8, *Bacillus pumilus* xylanase; 9, *Streptomyces lividans* xylanase C xylanase; 10, *Streptomyces* sp. xylanase; 11, *Streptomyces lividans* xylanase B; 12, *Bacillus circulans* xylanase; 13, *Bacillus subtilis* xylanase; 14, *Schizophillum commune* xylanase; 15, *Trichoderma harzianum* xylanase; 16, *Trichoderma viride* xylanase; 17, *Trichoderma reesei* xylanase 2; 18, *Aspergillus aculeatus* FI-carboxymethylcellulase; 19, *Erwinia carotovora* endoglucanase S. [a]N-terminal domain; [b]C-terminal domain.

that CelS of *E. carotovora* and low molecular weight xylanases might be related, has proven to be correct. The sequence similarity between the two families suggests a common ancestry as well as related folds. The hydrolysis reaction catalysed by family G xylanases has recently been shown to proceed with retention of anomeric configuration at the newly formed reducing end [20] and the same mechanism of action can therefore be predicted for family H cellulases. The two catalytic Glu residues in *Bacillus pumilus* xylanase have been identified from 3D-structure data [18] and site-directed mutagenesis [19]. By similarity, it can be predicted that the two corresponding Glu residues (Glu-126 and -214 for

E. carotovora CelS; Glu-118 and 202 for *A. aculeatus* cellulase) in family G cellulases will also be essential for catalysis.

Although the sequence identity between these two families is below 20%, the HCA scores are about 70%, a value which is found among proteins sharing related 3D folds [11,17]. This confirms that, although time-consuming, HCA is a very sensitive tool to uncover distant relatedness between proteins for which standard alignments methods are unreliable. The increased sensitivity of HCA over other methods has recently been shown to originate from the good correspondence between hydrophobic clusters and the secondary structure elements in proteins [21]. The 3D structure of the *B. pumilus* xylanase has been solved [18] and was shown to consist of 21 β-strands and one α-helix. Quite interestingly the number of hydrophobic clusters detected by HCA was 20 (Fig. 1).

Family G xylanases are very specific towards xylan and do not have any activity on carboxymethylcellulose (CMC). Although CMC was found to be the best substrate for CelS of *E. carotovora*, a notable activity was also reported on xylan [9]. The molecular weight difference between the catalytic domain of family G xylanases (ca. 20 kDa) and the catalytic domain of family H cellulases (ca. 25 kDa) is significant and it can be thus speculated whether the strict specificity of xylanases for xylan is due to a more compact 3D fold which could prevent the binding of larger cellulose molecules in the enzyme active site. Also, the HCA plots in Fig. 1 show that the two cellulases have an N-terminal extension of ~ 40 residues and which is missing or shorter in the xylanases. The role of this region is not known.

The divergence of high pI and acid pI xylanases is clear (Fig. 2). The family G acid xylanases reported so far are all from fungal origin. There are many indications in the literature suggesting that e.g. *Aspergillus* and *Trichoderma* spp. produce both alkaline and acid xylanases [22–26]. By far the most carefully studied organism in this respect is *Trichoderma reesei* with its low and high pI xylanases [26]. Acid xylanases tend to have slightly lower pH optima (~ pH 3–4) than the alkaline ones (~ pH 4–5). They both have the highest activity against substituted xylan but the specific catalytic activities of alkaline xylanases are about two times higher than that of acid ones [25]. The biological significance of these two very similar but different enzymes in their natural environment remains to be explained.

Acknowledgements: The help of the Association of Finnish Chemical Societies and the Finnish Academy for a travel grant (to A.T.) is gratefully acknowledged.

REFERENCES

[1] Wong, K.K.Y. and Saddler, J.N., in: Xylans and Xylanases (J. Visser, Ed.), Elsevier, Amsterdam, 1992, pp. 171–187.
[2] Knowles, J., Lehtovaara, P. and Teeri, T. (1987) Trends Biotechnol. 5, 255–261.
[3] Kubicek, C.P. (1992) Adv. Biochem. Eng. 45, 1–27.
[4] Henrissat, B., Claeyssens, M., Tomme, P., Lemesle, L. and Mornon, J.-P. (1989) Gene 81, 83–95.
[5] Béguin, P. (1990) Annu. Rev. Microbiol. 44, 219–248.
[6] Gilkes, N.R., Henrissat, B., Kilburn, D.G., Miller, R.C. and Warren, R.A.J. (1991) Microbiol. Rev. 55, 303–315.
[7] Henrissat, B. (1991) Biochem. J. 280, 309–316.
[8] Wong, K.K.Y., Tan, L.U.L. and Saddler, J.N. (1988) Microbiol. Rev. 52, 305–317.
[9] Saarilahti, H.T., Henrissat, B. and Palva, E.T. (1990) Gene 90, 9–14.
[10] Ooi, T., Shinmyo, A., Okada, H., Hara, S., Ikenaka, T., Murao, S. and Arai, M. (1990) Curr. Genet. 18, 217–222.
[11] Gaboriaud, C., Bissery, V., Benchetrit, T. and Mornon, J.P. (1987) FEBS Lett. 224, 149–155.
[12] Lemesle-Varloot, L., Gaboriaud, C., Pantel, G., Morgat, A., Mornon, J.P., Lavaitte, S., Lestang, F. and Henrissat, B. (1993) Comput. Appl. Biosci., in press.
[13] Needleman, S.B. and Wunsh, C.D. (1970) J. Mol. Biol. 48, 443–453.
[14] Dayhoff, M.O., in: Atlas of Protein Sequence and Structure (M.O. Dayhoff, Ed.), Vol. 5, Suppl. 3., National Biomedical Research Foundation, Washington, DC, 1978, pp. 1–8.
[15] Saitou, N. and Nei, M. (1987) Mol. Biol. Evol. 4, 406–425.
[16] Lipman, D.J., Altschul, S.F. and Kececioglu, J.D. (1989) Proc. Natl. Acad. Sci. USA 86, 4412–4415.
[17] Lemesle-Varloot, L., Henrissat, B., Gaboriaud, C., Bissery, V., Morgat, A. and Mornon, J.-P. (1990) Biochimie 72, 555–574.
[18] Okada, H. (1989) Adv. Prot. Des. 12, 81–86.
[19] Ko, E.P., Akatsuka, H., Moriyama, H., Shinmyo, A., Hata, Y., Katsube, Y., Urabe, I. and Okada, H. (1992) Biochem. J. 288, 117–121.
[20] Gebler, J.C., Gilkes, N.R., Claeyssens, M., Wilson, D.B., Béguin, P., Wakarchuk, W.W., Kilburn, D.G., Miller, R.C., Warren, R.A.J. and Withers, S.G. (1992) J. Biol. Chem. 267, 12559–12561.
[21] Woodcock, S., Mornon, J.-P. and Henrissat, B. (1992) Prot. Eng. 5, 629–635.
[22] Frederick, M.M., Kiang, C.-H., Frederick, J.R. and Reilly, P.J. (1985) Biotechnol. Bioeng. 27, 525–532.
[23] Fournier, R., Frederick, M.M., Frederick, J.R. and Reilly, P.J. (1985) Biotechnol. Bioeng. 27, 539–546.
[24] Maat, J., Roza, M., Verbakel, J., Stam, H., Santos da Silva, M.J., Bosse, M., Egmond, M.R. and Hagemans, M.L.D., in: Xylans and Xylanases (J. Visser, Ed.), Elsevier, Amsterdam, 1992, pp. 349–360.
[25] Tenkanen, M., Puls, J. and Poutanen, K. (1992) Enzyme Microb. Technol. 14, 566–574.
[26] Törrönen, A., Mach, R.L., Messner, R., Gonzalez, R., Kalkkinen, N., Harkki, A. and Kubicek, C.P. (1992) Bio/Technology 10, 1461–1465.
[27] Shareck, F., Roy, C., Yaguchi, M., Morosoli, R. and Kluepfel, D. (1991) Gene 107, 75–82.
[28] Ito, K., Iwashita, K. and Iwano, K. (1992) Biosci. Biotech. Biochem. 56, 1338–1340.
[29] Yaguchi, M., Roy, C., Ujiee, M., Watson, D.C. and Wakarchuk, W., in: Xylans and Xylanases (J. Visser, Ed.), Elsevier, Amsterdam, 1992, pp. 149–154.
[30] Yaguchi, M., Roy, C., Watson, D.C., Rollin, F., Tan, L.U.L., Senior, D.J. and Saddler, J.N., in: Xylans and Xylanases (J. Visser, Ed.), Elsevier, Amsterdam, 1992, pp. 435–438.