Three-Dimensional Structure of Cellobiohydrolase II from *Trichoderma reesei*

J. ROUVINEN, T. BERGFORS, T. TEERI, J. K. C. KNOWLES,* T. A. JONES†

The enzymatic degradation of cellulose is an important process, both ecologically and commercially. The three-dimensional structure of a cellulase, the enzymatic core of CBHII from the fungus *Trichoderma reesei* reveals an α-β protein with a fold similar to but different from the widely occurring barrel topology first observed in triose phosphate isomerase. The active site of CBHII is located at the carboxyl-terminal end of a parallel β barrel, in an enclosed tunnel through which the cellulose threads. Two aspartic acid residues, located in the center of the tunnel are the probable catalytic residues.

CELLULOSE, ONE OF THE MOST ABUNDANT ORGANIC COMpounds on Earth, is the major polysaccharide in plants where it is part of the cell walls. Cellulose-degrading enzymes participate in the natural, ecological recycling of plant material; they are now acquiring increasing commercial significance because of their use in detergents and in the manufacture of food and paper. The filamentous fungus *Trichoderma reesei* is an efficient producer of cellulases (*1, 2*) that have been defined according to their product specificity as endoglucanases or cellobiohydrolases (also called exoglucanases). Two of these enzymes, cellobiohydrolases I and II (CBHI and CBHII) participate in the hydrolysis of cellulose, acting synergistically to degrade both crystalline and amorphous cellulose (*3–5*). Because of the complex nature of the solid substrate, the difficulty of purifying individual enzymes, and the lack of three-dimensional structure information, the precise mechanism of action of these enzymes and the nature of the synergistic interactions between them is not understood.

The genes encoding CBHI and CBHII and the endoglucanases EGI and EGII (formally called EGIII) have been isolated and characterized (*6–12*). The predicted amino acid sequences of CBHI and CBHII show no overall homology. However, a 36–amino acid region of high homology occurs in all four enzymes. In CBHII and EGII, this region of homology occurs at the amino terminus, while in CBHI and EGI it is at the carboxyl terminus. This so-called A domain has no catalytic activity in CBHI and CBHII, but it is required for full activity on cellulose and has some cellulose binding activity (*13–15*). It is a compact, wedge-shaped molecule whose structure has been determined by two-dimensional nuclear magnetic resonance (*16*). It is always linked to the enzymatically active core protein by a polypeptide rich in serine and threonine residues that are O-glycosylated in CBHI and CBHII (*14, 17*). The intact molecules have been shown by low angle x-ray diffraction to be tadpole-shaped (*18, 19*). Proteolytic cleavage produces a catalytic domain (~45 kD) that has full activity on small synthetic substrates, but only partial activity to natural solid cellulose (*14*).

Both CBHI and CBHII yield, on hydrolysis, the disaccharide cellobiose. CBHII has greater substrate specificity because it requires at least three contiguous β(1 → 4)-linked glucosyl units to hydrolyze a glycosidic bond (*20, 21*). It has been proposed on the basis of competitive binding studies that CBHII has an extended active site with four subsites (*20*). CBHII inverts the configuration at the glycosidic bond so that the first-formed cellobiose is the α anomer, while CBHI procedes with retention, and produces cellobiose as the β anomer (*22*).

An analysis of all available cellulase sequences suggests that cellulases can be divided into seven families (*23*). The other members of the CBHII family are three bacterial endoglucanases, one of which, encoded by the *cenA* gene in the bacterium *Cellulomonas fimi*, is also an inverting enzyme (*24*) and is built up from domains linked by a glycosylated region (*25*).

The crystallization of endoglucanase D from *Clostridium thermocellum* has been reported (*26*). Attempts to crystallize other, intact cellulases have failed, probably because of the inherent flexibility at the linker region between separated domains. However, we have been able to crystallize the enzymatically active core protein of CBHII (*27*) corresponding to amino acid residues 83 to 447, and now report its three-dimensional structure.

Structure determination. Crystals were obtained that diffract to better than 2 Å resolution and belong to space group $P2_1$ with a = 49.1 Å, b = 75.8 Å, c = 92.9 Å and β = 103.2° (*27*). All data were collected on a Xentronics area detector and evaluated with software as described (*28*). The unit cell volume, molecular size, and a large non-origin peak in the native Patterson function suggested that there were two molecules in the asymmetric unit with a noncrystallographic translation of (0.50, ±0.37, 0.50) (*27*).

An initial search for derivatives revealed differences in the mercurials (Table 1). The location of the heavy atom binding sites was obtained by standard Patterson and difference Fourier techniques without the use of the noncrystallographic symmetry. Pairs of peaks could be identified in the top sites separated by an average translation operator (0.530, 0.623, 0.497). The multiple isomorphous replacement (MIR) map calculated with anomalous data showed the two, clearly separated molecules, built up from extensive α-β structure. These phases were also good enough to locate the binding site of an inhibitor, *o*-iodobenzyl-1-thio-β-cellobioside, but were not good enough to trace the chain. After model building the

J. Rouvinen, T. Bergfors, and T. A. Jones are in the Department of Molecular Biology, BMC, Box 590, S-75124, Uppsala, Sweden. T. Teeri and J. C. Knowles are at the Biotechnical Laboratory, VTT, Tietotie 2, SF-02150 Espoo, Finland.

*Present address: Glaxo Institute for Molecular Biology, Route des Acacias, 46, 1211, Geneva 24, Switzerland.
†To whom all correspondence should be addressed.

Table 1. Data collection and heavy atom refinement statistics. The heavy atom derivatives were refined with their anomalous data. The overall figure of merit was 0.55 to 2.8 Å resolution. ($Hg(OAc)_2$ is mercury acetate; HgPh, *o*-chloromercury phenol; inhibitor, *o*-iodobenzyl-1-thio-β-cellobioside.) The merging *R* factor is defined as *R* merge = $[\Sigma_N\Sigma_n(I-\langle I\rangle)/\Sigma_N n\langle I\rangle] \times 100$ percent where *N* is the number of unique measurements, *n* is the number of multiple measurements of a particular reflection; *I* is the measured and $\langle I\rangle$ the mean intensity of a reflection. The rejection *R* merge for native data used a PROTEIN rejection ratio of 0.6, the other compounds were merged with a rejection ratio of 0.4. The anomalous merging *R* factor is defined as *R* merge = $[\Sigma_N(\langle I+\rangle-\langle I-\rangle)/\Sigma_N(\langle I+\rangle+\langle I-\rangle)] \times 100$ percent where $I+$, $I-$ refer to Friedel pairs; rms Fh/residual is defined as $[(\Sigma f_H^2)/(\Sigma(F_{DER}-F_{PH})^2]^{1/2}$ where f_H is the heavy atom form factor, and F_{DER} and F_{PH} are the derivative structure factor and calculated structure factor. $R_{Cullis} = \Sigma(|F_{DER}-F_{PH}|)/\Sigma(|F_{DER}-F_{NAT}|)$ where F_{NAT} is the native structure factor and the summation is taken over the centric reflections.

Item	Native	$Hg(OAc)_2$	HgPh	$EuCl_2$	Inhibitor
Crystals	2	2	1	1	1
Soaking time (hours)		16/48	24	19	48
Soaking conc. (mM)		0.5	5.0	1.0	2.0
Measured reflections	88337	25822	25468	36893	39063
Unique reflections	32934	9819	10993	13815	15983
R merge	7.64	8.62	8.50	7.72	5.94
R merge, reject	7.26	8.08	7.72	7.20	5.53
R merge anomalous	4.73	6.46	6.73	5.97	4.71
R merge to native		16.62	21.13	13.71	9.89
Data to 2.8 Å (%)	98	60	67	84	86
Data to 2.5 Å	97				69
Data to 2.0 Å	74				
Number of sites		6	6	2	10
Rms Fh/residual		2.26	1.70	1.45	1.22
R_{Cullis}		0.64	0.73	0.76	0.79

inhibitor, we used it as a heavy atom derivative by treating each ring as a single "atom," and then including the sulfur and iodine atoms separately. A fourth derivative was obtained by soaking the crystals in europium dichloride. The final phasing statistics after heavy-atom refinement (*29*) are given in Table 1. The transformation operator relating the two molecules was improved by first translating the main-chain skeletonized electron density of molecule A into the density of molecule B, and then refining its fit by real space optimization (*30*). These new coordinates could then be used to calculate a new transformation between the A and B molecules. The resulting averaged map showed improved continuity.

The molecule chain tracing was made from skeletonized, nonaveraged and averaged maps, and the initial model was built with fragments from a library of well-refined protein structures (*31*). Deciding the correctness of this model was facilitated by the use of a new computer graphics program, O (*32*), which, among many other features, allows the simultaneous display of overlayed noncrystallographically related molecules and their respective densities. The initial model was improved by repeated cycles of crystallographic refinement and manual model rebuilding, at increasingly higher resolutions. Each cycle consisted of molecular dynamics refinement by the method of simulated annealing (*33*). Our model (M25 in our nomenclature) has a crystallographic *R* factor of 15.5 percent for all measured reflections in the resolution range 8.0 to 2.0 Å. This model contains coordinates and individual temperature factors for two protein molecules, each corresponding to residues 85 to 447 (the first two residues are not observed in our maps and are presumed to be disordered), and a total of 616 water molecules (0.85 water molecule per amino acid). The model is tightly restrained with overall root-mean-square (rms) derivations of 0.011 Å in bond lengths, 2.7° in bond angles, and 1.2° in fixed improper dihedral angles. The fit of the model to the electron density is shown in Fig. 1A (*34*), together with a representative volume of electron density in Fig. 1B.

The structures of three complexes of enzyme and ligand have been

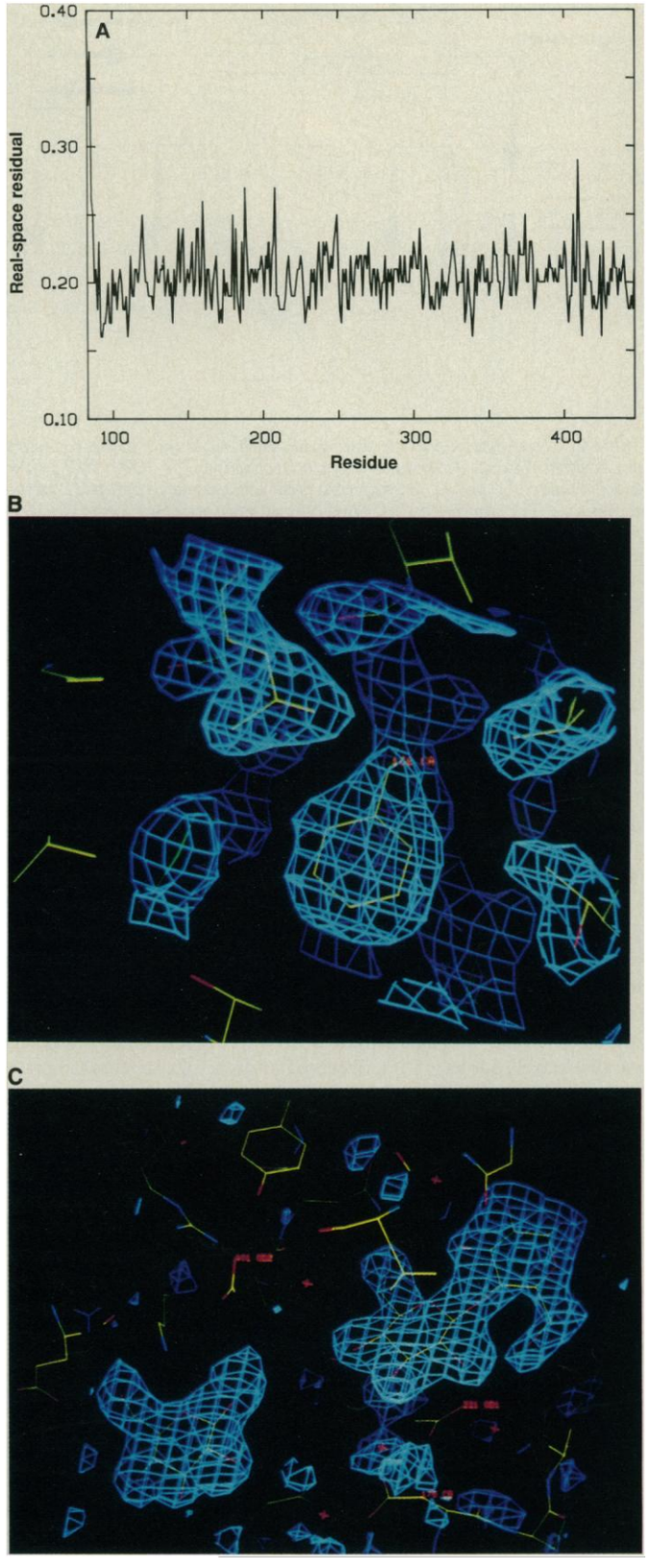

Fig. 1. (**A**) Main chain and side chain electron density fit (*34*) for model A25. (**B**) Representative electron density for the native protein. The map was calculated at 2 Å resolution with $(2|F_{obs}| - |F_c|)$ amplitudes and model phases. The contours are drawn at a level of one standard deviation, centered on Phe^{166}. (**C**) Electron density glucose-cellobiose inhibitor complex. The map was calculated at 2.0 Å resolution with $(|F_{obs}| - |F_c|)$ amplitudes and model phases. The ligands were not included in the phase calculation.

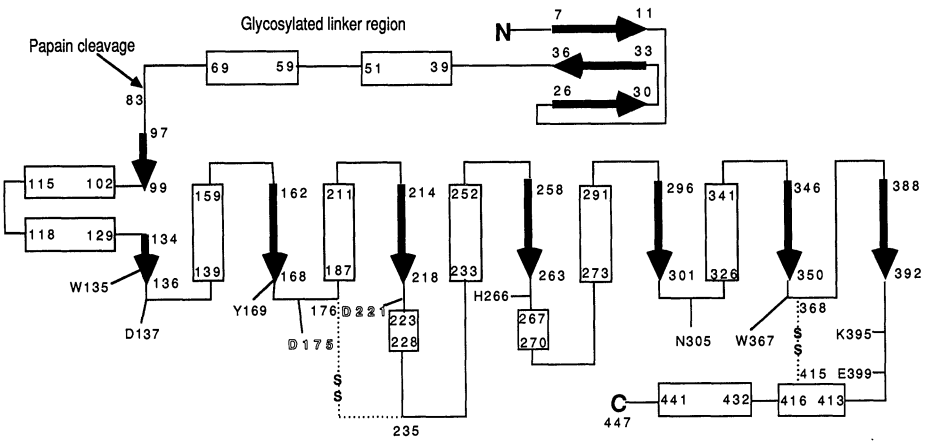

Fig. 2. Secondary structure schematic of the intact CBHII molecule. The NH_2-terminal domain (residues 3 to 38) is based on the two-dimensional-NMR structure of Kraulis *et al.* (*16*). Arrows indicate β strands, and rectangles indicate α helices. No three-dimensional structure for the linker region is available but the two α helices shown signify sequence homologies to the first α helix in the core domain (*39*). N, amino terminal; C, carboxyl terminal.

Table 2. Complexes of enzyme and ligand. A, B, C, D refer to the four sites described in the text. *R* is the crystallographic residual $R = 100 \times \Sigma(|F_o - F_c|)/\Sigma(F_o)$, where F_o and F_c are the observed and calculated structure factor amplitudes. The summation is carried over all measurements in the specified resolution range with a lower resolution cutoff of 8 Å. Glc refers to a glucosyl, *R*1 to the *o*-iodobenzyl, and *R*2 to the 4-methylumbelliferyl rings, respectively. The dash –, indicates a covalent linkage between the rings.

Ligand	A	B	C	D	*R* (%)	Res (Å)
o-Iodobenzyl-1-thio-β-cellobioside	Glc–	Glc–	*R*1		13.6	2.5
4-Methylumbelliferyl-β-D-cellobioside	Glc–	Glc–	*R*2		14.0	2.5
Glucose-celloboise	Glc		Glc–	Glc	15.9	2.0

determined and refined independently (Table 2) and the difference Fourier electron density of one of the complexes is shown in Fig. 1C.

Molecular fold. The enzymatic core of CBHII has an unusual fold (Figs. 2 and 3). The molecule is a large, single domain α-β protein with a central β barrel made up of seven parallel strands. The first six strands are connected by α helices, whereas the connection between the sixth and seventh strands is irregular. This fold is therefore similar to but different from a frequently observed $(\beta\alpha)_8$ parallel barrel topology [first observed in triose phosphate isomerase (*35*), and hence called the TIM barrel]. In this family of structures, the first and eighth strands form a hydrogen bonding ladder to complete the barrel. The barrels may vary in their precise shape, having different eccentricities so that some are almost circular in cross section—for example, glycolate oxidase (*35*)—while others such as TIM are almost elliptical (*36*). Neighboring strands are commonly tilted to each other by about −35°, and have a shear of 8 in encircling the barrel (*37, 38*). Because of the β-strand H-bonding patterns, regular seven- or nine-stranded barrels with odd shear numbers are not possible (because of the alternating directions of the hydrogen bonds along the chain). They would require shear numbers of 6 or 8 for seven-stranded barrels and 8 or 10 for nine-stranded barrels. It has been suggested that such barrels would result in poor interior side-chain packing (*38*). Other deviations from the eight-stranded TIM barrel structure have been considered and rejected. Six-stranded barrels would have a small internal radius while ten-stranded barrels would be larger. Both are likely to produce poor internal side-chain packing (*38*).

The shear in the CBHII barrel is 8 and the hydrogen-bonding interactions going around the barrel are as extensive as in the TIM family, except for the barrel closure between the first and the last strands. In CBHII, there is only one hydrogen bond between these strands, which are rather separated (Fig. 3). The side-chain packing inside the barrel results in no volume accessible to solvent. However, some of the helices packing onto the barrel are long, and cause a number of enclosed volumes that contain solvent. The region 97 to 99 is not a missing eighth strand; it is outside the barrel but helps complete the hydrogen bonding to strand 2.

Two of the loops at the carboxyl-terminal end of the barrel are extensive (172 to 189 and 394 to 429) and each contains a disulfide bridge. Side chains from these loops and from the barrel form an almost perfectly enclosed tunnel, about 20 Å long (Figs. 3 and 4). In the native structure, the tunnel contains numerous water molecules. In its center a small tube (containing water molecules) leads to the surface. The only other protrusion in the accessible surface of the tunnel leads to Asp^{263}. The loops making up the tunnel are well

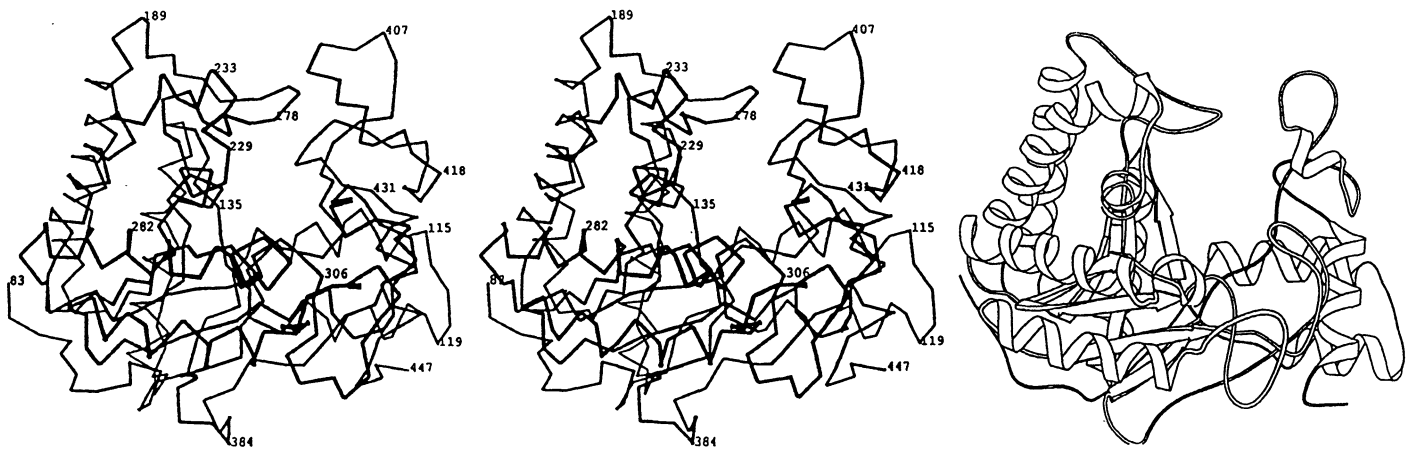

Fig. 3. The Cα skeleton and ribbon drawing (*52*) of the CBHII core. The view is chosen to clarify the active site tunnel and its position relative to the β barrel.

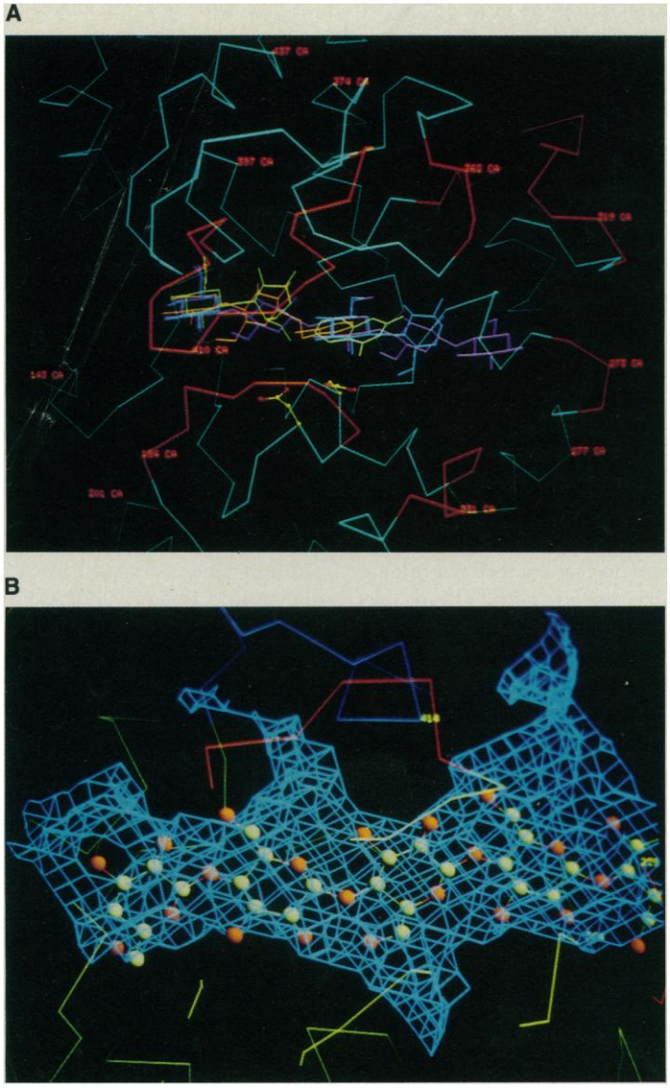

Fig. 4. (**A**) The Cα coloring code localizes the deletions within the CBHII family (Fig. 6). Green implies that there are no deletions, and red the likely deletions in the endoglucanases. The *o*-iodobenzy-1-thio-β-cellobioside ligand is colored brown, the 4-methylumbelliferyl β-D-cellobioside is colored yellow, the glucose-cellobiose is colored blue, and the modeled cellulose chain is purple. The aspartic acids that we suggest are involved in catalysis are also shown. Sites A, B, C, and D are located left to right. The nonreducing end of the cellulose chain is at the left. (**B**) The active site tunnel is illustrated by drawing the solvent accessible surface (*53*) as a density mesh.

ordered and similar in both molecules A and B. Many of the side chains in the tunnel are polar, forming a complex hydrogen bond and salt link network.

CBHII is glycosylated (*5, 14*) although it is not known if the carbohydrate interacts in any way with the cellulose substrate. As would be expected, both Asn^{289} and Asn^{310} are glycosylated. The linkage region between the core and the so-called cellulose binding domain is rich in serines and threonines that are O-glycosylated (*14*). However, we unexpectedly find that threonines 87 and 97 and serines 106, 109, 110, and 115 are also O-glycosylated (*39*). The electron density suggests that these sugars are α-linked mannose units.

CBHII-core contains six cysteine residues that we find result in two disulfide bridges (176–235 and 368–415). Both help to stabilize the loops involved in forming the tunnel. The two free cysteines are the principal binding sites for our mercurial reagents.

The active site. The natural function of CBHII is to hydrolyze cellulose into units of the disaccharide cellobiose. More detailed knowledge of the active site of CBHII has come from studies of its interaction with small substrates and ligands (*20, 21*). The results suggest that CBHII has an extended binding site with four subsites, one of which has a high affinity for glucose. Hydrolysis of 4-methylumbelliferyl-β-D-glycosides [MeUmb-$(Glc)_n$ where $n > 3$] results in cleavage of the glycosidic bond between the second and third residues from the nonreducing end when $n = 3$ or 4. In contrast, when $n = 5$, the substrate is cleaved both between the second and third residues and between the third and fourth (resulting in the creation of two different nonhydrolyzed products). The fluorescence of the ligands at $n = 1$, 2, or 3 is quenched on binding to the enzyme.

Although the reason is unclear, parallel β-sheet enzymes, whether nucleotide binding domains or TIM barrels, invariably have their active sites at the carboxyl-terminal end of the β-sheet structure (*40*). The unusual tunnel observed in the CBHII-core would therefore be an obvious suggestion for an active site that must bind a long, floppy polymer such as a single cellulose chain. This has been confirmed by determining the structure of three different enzyme-inhibitor complexes, *o*-iodobenzyl-1-thio-β-cellobioside, 4-methylumbelliferyl-β-cellobioside, and glucose-cellobiose complex. The tunnel contains four clear binding sites for glucosyl units that we refer to as A, B, C, and D. The sites for the ligands are shown in Table 2, and their structures overlaid in Fig. 4A. All ligands bind within the tunnel, with sites A and C occupied in each complex, and the same chain directionality is maintained for each oligosaccharide. This has allowed us to build a cellulose chain occupying all four sites (Figs. 4 and 5) with the extended conformation observed in crystalline cellulose. Only the B site glucosyl ring of 4-methylumbelliferyl-β-cellobioside deviates significantly from this model. Whether this ring conformation is of functional significance is unclear.

Changes in the protein structure on ligand binding are small, close to the expected experimental errors. At site A, protein side chain atoms made hydrogen bonds to the ring oxygens (Asp^{137} to O3 and O4, Glu^{399} to O6). The more hydrophobic β face of the ring packs onto Trp^{135}, while the other face packs against the main chain of Ala^{427}-Gly^{428}. Its edges are wedged between the side chains of Lys^{395} on one side and of Asp^{137} and Tyr^{169} on the other side. In site B, the ring is completely in the tunnel and is surrounded by the side chains of Asp^{401}, Lys^{395}, Tyr^{169}, and Ala^{304}.

Site C is also completely surrounded by protein atoms (Trp^{367}, Asp^{412}, His^{266}, Asn^{305}, Asp^{175}, Asp^{221}, and Ala^{178}). One face of the benzene ring of the inhibitor—the α face of a glucosyl ring in the glucose-cellobioside complex and in our cellulose model—stacks onto the ring of Trp^{367} while the side chain of Asp^{221} is positioned at the other face. Asp^{175} and Asp^{221} are closest to the acetal linkage between B and C in our modeling and are buried. The fourth site, D, is not so tightly packed, with the α face of the sugar packing onto the ring of Trp^{269}, and the rest of the site formed by His^{266}, Gly^{365}, His^{414}, Asn^{225}, and Asn^{229}. Tryptophans 272 and 364 may make another, external site.

The site of cellulose cleavage is likely to be between B and C, and involves aspartic acid residues 175 and 221. An oxygen atom from each carboxylate is close to the glycosidic oxygen between sites B and C in the modeled cellulose (in the range 4.5 to 4.9 Å). The other carboxylate oxygen of Asp^{175} is close to the glucosyl site B ring oxygen (4.3 Å). The enzyme is active at acidic pH values (between pH 4.0 and pH 7.0) where both charged and uncharged forms of the aspartyl side chains could be expected. In both the native and inhibitor structures, atoms 221 OD2 and 175 OD2 are close enough (2.8 Å) and have the correct geometry (angles 175 CG–175 OD2–221 OD2 and 175 OD2–221 OD2–221 CG are 119.6° and

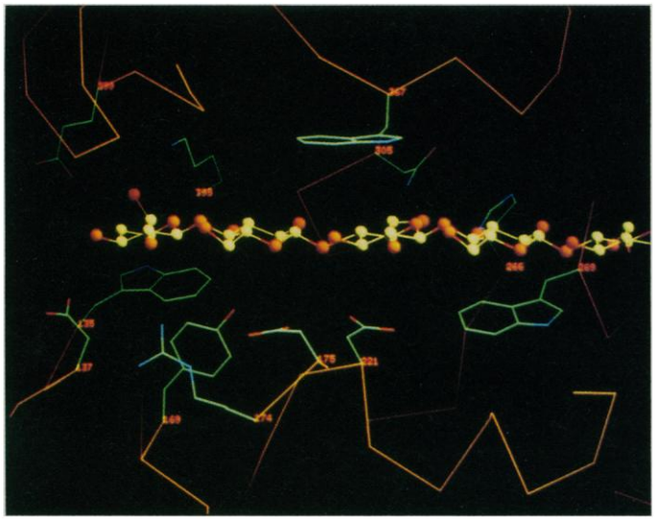

Fig. 5. Close-up of the proposed active site with some of the most important amino acid side chains. The protein coordinates are native model A25, and the cellulose chain is the result of model-building.

121.0°, respectively) to suggest the formation of a hydrogen bond. This in turn implies that one of the carboxylate groups is protonated. Asp^{175} forms a hydrogen bond with the guanido group of Arg^{174} (175 OD1–174 NE = 3.1 Å). The other guanido group atoms NH1 and NH2 hydrogen-bond to the carbonyl oxygens of 169 and 137 (2.8 and 3.3 Å, respectively) and possibly with the O6 hydroxyl in site B. These interactions lead us to suggest that Asp^{175} is more likely to be charged, and Asp^{221} to be protonated. We have demonstrated the importance of these residues by the use of site-directed mutagenesis. The Asp^{175}–Ala^{175} mutant shows only ~20 percent of wild-type activity, while the Asp^{221}–Ala^{221} mutant has no measurable activity (*41*). While Asp^{175} is conserved in all four members of the CBHII family, Asp^{221} is conserved in only three of the four proteins (Fig. 6).

The water molecule needed for catalysis could enter the active site either from the solvents interacting with the buried Asp^{263} or, more likely, through the narrow tube reaching to the external solvent referred to earlier (Fig. 4B). This tube, leading directly to the B-C binding site, is on the opposite side of the glucosyl rings relative to Asp^{175} and Asp^{221}, and could place a water molecule near the anomeric carbon atom in a position suitable for the inverting reaction. Although it may be premature to suggest a detailed mechanism for the cleavage reaction catalyzed by this enzyme, in a single displacement mechanism (*42*), Asp^{221} is in a position to act as a proton donor, thus assisting the departing aglycone by general acid catalysis. A candidate for a general base to assist the nucleophilic attack of the water molecule is more difficult to identify. Two candidates are the buried Asp^{263}, conserved in all members of the family, but placed below the anomeric carbon atom, and Asp^{401}, involved in salt links to Arg^{353} and Lys^{395}, but conserved in three members of the family. In this scheme, Asp^{175} acts to create the environment to force the protonation of Asp^{221} and, because of its proximity, to assist in the stabilization of any transitional positive charge that may build up at the oxygen ring atom. The residue conservation pattern of the *casA* gene product is different from the other members of the family where neither Asp^{221} nor Asp^{401} are conserved. Since both CBHII and the *cenA* gene product are inverting enzymes, it is possible that the *celA* and *casA* gene products may be inverting and retaining enzymes, respectively.

The structure of the enzyme explains why cellobiose is the product of cellulose degradation. The extended conformation of cellulose results in a zig-zag pattern of glycosidic linkages. Threading the chain into the tunnel allows the correct conformation to occur at the active site every two glucosyl units. The restricted volume of the tunnel prevents extensive conformational rearrangement when the chain is advanced by one unit. The degradation pattern observed for cello-oligosaccharides (*20, 21*) suggests that both the A and B sites should be occupied to position the substrate into position for the hydrolysis. The observed cleavage patterns for "$n = 5$" glycosides (cleavage at 2 to 3 or at 3 to 4) can be explained as follows: Once an extended cellulose chain enters the active site, half of the molecules, by chance, will have their 2-3–glycosidic linkage, and half will have their 3-4–glycosidic linkage in a conformation susceptible to catalytic hydrolysis.

It has been frequently suggested that the catalytic mechanism of cellulases resembles that of lysozymes (*43*). While this may be true, the difference in stereochemical course of the two classes of CBH enzymes (CBHI and lysozyme go by retention, but CBHII procedes by inversion) rules out a very close analogy. From the structural work of Phillips and co-workers (*44*), we know that in hen egg-white lysozyme, the binding site is a groove on the surface of the protein that comprises a number of subsites. The active-site carboxylate groups Glu^{35} (which acts as the proton donor), and Asp^{52} (which stabilizes the putative oxocarbonium ion intermediate) lie on either side of the bond to be cleaved with their Cα's separated by about 10 Å (compared with 5.9 Å in CBHII). However, T4 lysozyme, consists of two lobes with the active site in the interface (*45*). Two of the proposed subsites are completely enclosed by the protein and model building suggests that because of some close contacts, there is some opening up of the active site (*46*). The active site carboxylates (proton donor Glu^{11} and Asp^{20}) are positioned on either side of the polysaccharide (Cα separations ~10 Å). Both of these lysozymes are considerably smaller than CBHII and have different folding topologies.

We observe a greater similarity with a family of other enzymes, those that degrade starch and related oligosaccharides. Three-dimensional structures are known for two members of the family, Taka-amylase from the fungus *Aspergillus oryzae* (*47*) and porcine pancreatic α-amylase (*48*). Each contains a TIM barrel domain with the active site (*49*) at the carboxyl-terminal end of the barrel. In both molecules the active site is open and lacks the tunnel observed in CBHII. Some members of this family work by inversion, others by retention of configuration, while conserving the catalytic carboxylates (*49*). As we observe in CBHII, the enzyme structure must define the approach of water into the active site (*49*) to account for this stereospecificity.

The role of the noncatalytic domain. CBHII is built up of two domains separated by a flexible linker region (Fig. 2). This architecture is found in other cellulases, both from prokaryotes and eukaryotes and must be important for their function. At least two hypotheses can be made regarding the function of the noncatalytic, cellulose binding domain. One proposes that it functions merely as a cellulose binding domain, thereby increasing the interactions between the catalytic domain and cellulose. The second hypothesis supposes a more active role in cellulose degradation by displacing individual cellulose chains from the cellulose crystals. The core structure does not help us choose the correct hypothesis. However, the amino terminus of the core (residue 85) is about 45 Å from the proposed entrance to the active site. This separation should be compared with the length of the linker region estimated from low angle x-ray diffraction, about 140 Å (*19*). It is, therefore, possible to visualize the active site rim residues of the enzyme sitting on the surface of a cellulose crystal, while still tethered to the anchoring noncatalytic domain.

The difference between endoglucanases and exoglucanases.

The CBHII family (*23*) at present consists of an exoglucanase (CBHII) and three bacterial endoglucanases—the *casA* gene product from an alkalophilic *Strepomyces* strain (*50*), the *cenA* product from *Cellulomonas fimi* bacteria (*25*) and the *celA* product from *Microbispora bispora* (*51*). Their presumed enzymatic cores show an extensive overall sequence homology (Fig. 6) with some highly conserved regions and others having little similarity (particularly at the amino and carboxyl terminals of the chains). However, there are a number of clear deletions in the endoglucanases corresponding to regions at the carboxyl terminal of the β barrel. In particular, the two loops forming the lid of the active site tunnel in CBHII are missing in the endoglucanases (although the *casA* sequence is difficult to align for the loop around 410). These deletions result in structures with a much more open active site containing a groove for multiple subsites.

CBHI and EGI which share 45 percent sequence identity, show four clear deletions in EGI relative to CBHI (*9*). These observations suggest a general explanation for the difference between an exoglucanase and an endoglucanase. We propose that the tunnel-shaped active site is typical for exoglucanases, which are therefore restricted to hydrolysis occurring at the ends of the chain. Because hydrolysis continues processively, it may be stopped at the first substitution encountered. Endoglucanases, with a more traditional open active site cleft or groove, can hydrolyze internal glycosidic bonds. We believe that the structure of CBHII represents a general framework for cellulolytic enzymes and that the various cellulases differ in the number of surface loops forming the active site.

```
Cbh2  .......... .......... .......... .......... .....QACSS
Cena  APGCRVDYAV TNQWPGGFGA NVTITNLGDP VSSWKLDWTY TAGQRIQQLW
Casa  .......... .......... .......... .......... ..........
Cela  .......... .......... .......... .......... ..........

      6          16         26         36         46
Cbh2  VWGQCGGQNW SGPTCCASGS TCVYSNDYYS QCLPGAASSS SSTRAASTTS
Cena  NGTASTNGGQ VSVTSLPWNG SIPTGGTASF GFNGSWAGSN PTPASFSLNG
Casa  .......... .......... .......... .......... ..........
Cela  .......... .......... .......... .......... ..........

      56         66         76         86         96
Cbh2  RVSPTTSRSS SATPPPGSTT TRVPPVGSGT ATYSGNPFVG VTPWANAYYA
Cena  TTCTGTVPTT SPTPTPTPTT PTPTPTPTPT PTPTVTPQPT SGFYVDPTTQ
Casa  .......... .......... .......... .....GTTAL PSMELYRAEA
Cela  .......... .......... .......... .........D SPFYVDPQSN

      106        116        126        136        146
Cbh2  SEVSSLAIPS LTGAMATAAA AVAKVPSFMW LDTLDKTPLM EQTLADIRTA
Cena  GYRAWQAASG TDKA...LLE KIALTPQAYW VGNWADASHA QAEVADYTGR
Casa  GVHAWLDANP GDHRAPLIVE RIGSEPEAVW FAGAYNPGTI TQQVAEVTSR
Cela  .AAKWVAANP NDPRTPVIRD RIAAVPTGRW FAN.YNPSTV RAEVDAYVGR

      156        166        176        186        196
Cbh2  NKNGGNYAGQ FVVYDLPDRD CAALASNGEY SIADGGVAKY KNYIDTIRQI
Cena  AV.AAGKTPM LVVYAIPGRD C........G SHSGGGV.SE SEYARWVDTV
Casa  RQQPPGQLPV VVPYMIPFRD C........G NHSGGGAPSF AAYAEWSGLF
Cela  AA.AAGKIPI MVVYAMPNRD C........G GPSAGGAPNH TAYRAWIDEI

      206        216        226        236        246
Cbh2  VVEYSDIRTL LVIEPDSLAN LVTNLGTPKC ANAQSAYLEC INYAVTQLNL
Cena  AQGIKG.NPI VILEPDALAQ L......GDC SGQGDRVGFL KYAAKSLTLK
Casa  AAGLGSEPVV VVLSPMRFRW .......IDC LENQQRAERL AALQASPEAV
Cela  AAGLRNRPAV IILEPDALPI M......TNC MSPSEQAEVQ ASMAYAGKKF

      256        262        272        282        292
Cbh2  PN....VAMY LDAGHAGWLG WPANQDPAAQ LFANVYKNAS SPRALRGLAT
Cena  G.....ARVY IDAGHAKWLS ...VDTPVNR LNQVGFEYAV ......GFAL
Casa  TDANPEARVY YDVGHSAWHA ...PAAIAPT LVEAGILEHG .....AGIAT
Cela  KAASSQAKVY FDAGHDAWVP ...ADEMASR LRGADIANSA .....DGIAL

      302        312        322        332        342
Cbh2  NVANYNGWNI TSPPSYTQGN AVYNEKLYIH AIGPLLANHG WSNAFFITDQ
Cena  NTSNY..... .........Q TTADSKAYGQ QISQRLG... ..GKKFVIDT
Casa  NISNY..... .........R TTTDETAYAS AVIAELG... .GGLGAVVDT
Cela  NVSNY..... .........R YTSGLISYAK SVLSAIG... ASHLRAVIDT

      352        362        372        382        392
Cbh2  GRSGKQPTGQ QQWGDWCNVI GTGFGIRPSA NTGDSLLDSF VWVKPGGECD
Cena  SRNGNGSN.. ...GEWCNPR GRALGERPVA VNDGSGLDAL LWVKLPGESD
Casa  SRNGNGPTA. .......... ..ADLVNTRT VTRCPGVDAF LWITCPVTDG
Cela  SRNGNGPLG. ...SEWCDPP GRATGTWSTT DTGDPAIDAF LWIKPPGEAD

      402        412        422        432        442
Cbh2  GTSDSSAPRF DSHCALPDAL QPAPQAGAWF QAYFVQLLTN ANPSFL....
Cena  G......... ........AC NGGPAAGQWW QEIALEMARN ARW.......
Casa  G......... .......DGP VFSPPKLQLP RKPAAGRGCR DTIVRSARQQ
Cela  G......... .......... .CIATPGVFV PDRAYELAMN AAPPTY....

Cbh2  .......... .......... ...
Cena  .......... .......... ...
Casa  QTRPPGKPGL PAGRDSIRHG AAG
Cela  .......... .......... ...
```

Fig. 6. Sequence alignment of the CBHII family of proteins. The disulfide 176–235 is expected to be conserved throughout the family, while 368–415 is conserved except for *casA*. Abbreviations used for the amino acid residues are: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.

REFERENCES AND NOTES

1. B. S. Montenecourt, *Trends Biotechnol.* **1**, 156 (1983).
2. J. Knowles, P. Lehtovaara, T. Teeri, *ibid.* **5**, 255 (1987).
3. L. Fägerstam and L. G. Pettersson, *FEBS Lett.* **119**, 97 (1980).
4. B. Henrissat, H. Driguez, C. Viet, M. Schülein, *BioTechnology* **3**, 722 (1983).
5. M.-L. Niku-Paavola, A. Lappalainen, T.-M. Enari, M. Nummi, *Biotechnol. Appl. Biochem.* **8**, 449 (1986).
6. T. Teeri, I. Salovuori, J. Knowles, *BioTechnology* **1**, 696 (1983).
7. S. Shoemaker, V. L. Schweikart, M. B. Ladner, D. H. Gelfand, M. A. Innis, *ibid.*, p. 691.
8. J. N. Van Arsdell *et al.*, *ibid.* **4**, 60 (1987).
9. M. Penttilä, P. Lehtovaara, H. Nevalainen, R. Bhikhabhai, J. Knowles, *Gene* **45**, 253 (1986).
10. T. T. Teeri, P. Lehtovaara, S. Kauppinen, I. Salovuori, J. K. C. Knowles, *ibid.* **51**, 43 (1987).
11. C. M. Chen, M. Gritzali, D. W. Stafford, *BioTechnology* **5**, 274 (1987).
12. M. Saloheimo *et al.*, *Gene* **63**, 11 (1988).
13. H. Van Tilbeurgh, P. Tomme, M. Claeyssens, R. Bhikhabhai, G. Pettersson, *FEBS Lett.* **204**, 223 (1986).
14. P. Tomme *et al.*, *Eur. J. Biochem.* **170**, 575 (1988).
15. G. Johansson, J. Ståhlberg, G. Lindeberg, Å. Engström, G. Pettersson, *FEBS Lett.* **243**, 389 (1989).
16. P. J. Kraulis *et al.*, *Biochemistry* **28**, 7241 (1989).
17. I. Salovuori, M. Makarow, H. Rauvala, J. Knowles, L. Kääriäinen, *BioTechnology* **5**, 152 (1987).
18. M. Schmuck, I. Pilz, M. Hayn, H. Esterbauer, *Biotechnol. Lett.* **8**, 397 (1986).
19. P. M. Abuja, I. Pilz, M. Claeyssens, P. Tomme, *Biochem. Biophys. Res. Commun.* **156**, 180 (1988).
20. H. Van Tilbeurgh, G. Pettersson, R. Bhikhabhai, H. De Boeck, M. Claeyssens, *Eur. J. Biochem.* **148**, 329 (1985).
21. M. Claeyssens, H. Van Tilbeurgh, P. Tomme, P. M. Wood, S. I. McRae, *Biochem. J.* **261**, 819 (1989).
22. J. K. C. Knowles *et al.*, *J. Chem. Soc. Chem. Commun.* **1988**, 1401 (1988).
23. B. Henrissat *et al.*, *Gene* **81**, 83 (1989).
24. S. G. Withers *et al.*, *Biochem. Biophys. Res. Commun.* **139**, 487 (1986).
25. W. K. R. Wong *et al.*, *Gene* **44**, 315 (1986).
26. G. Joliff *et al.*, *BioTechnology* **4**, 896 (1986).
27. T. Bergfors *et al.*, *J. Mol. Biol.* **209**, 167 (1989).
28. M. Blum, P. Metcalf, S. C. Harrison, D. C. Wiley, *J. Appl. Cryst.* **20**, 235 (1987).
29. The crystallographic measurements were processed with the PROTEIN system of W. Steigemann. The heavy atom difference Patterson maps were solved with the aid of a vector search program RSPS written by S. Knight. The heavy atom sites were refined with the PHASEREF program of S. J. Remington and Tenn Eyck as implemented in PROTEIN.
30. T. A. Jones and L. Liljas, *Acta Crystallogr.* **A40**, 50 (1984).
31. T. A. Jones and S. Thirup, *EMBO J.* **5**, 819 (1986).
32. T. A. Jones, M. Bergdol, M. Kjeldgaard, in *Molecular Modelling*, S. Ealick and C. A. Bugg, Eds. (Springer, New York, 1990). The program O is now implemented on a DEC VMS Vax/Evans and Sutherland PS300, and on a Stardent GS1000 workstation. All computer graphics figures were photographed from the Stardent.
33. A. T. Brünger, J. Kuriyan, M. Karplus, *Science* **235**, 458 (1987); A. T. Brünger, *J. Mol. Biol.* **203**, 803 (1988).
34. T. A. Jones, J.-Y. Zou, S. W. Cowan, M. Kjeldgaard, *Acta Crystallogr. A*, in press. The standard crystallographic *R* factor gives an overall view of how good a model fits the experimental observations. This can result in incorrect structures (especially partially refined structures) having quite low *R* factors. The function plotted in Fig. 1 shows how good a fit each residue (main chain or side chain, or both) makes with the electron density. For each group of atoms being considered the function calculates $(\Sigma|\rho_{obs} - \rho_{calc}|)/(\Sigma|\rho_{obs} + \rho_{calc}|)$ on the grid of the observed density, for all grid points contained in the map calculated from the atoms. The plot for main chain atoms shows how continuous the traced chain actually is, while a plot for side chains could show out of register errors. Residues placed in zero density would have a value of 1.0 for this function.
35. D. W. Banner *et al.*, *Nature* **255**, 609 (1975).
36. Y. Lindqvist and C.-I. Brändén, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 6855 (1985).
37. A. D. McLachlan, *J. Mol. Biol.* **128**, 49 (1979); shear refers to the residue advance moving around the barrel to return the same starting residue.
38. A. M. Lesk *et al.*, *Proteins Struct. Function Genet.* **5**, 139 (1989).
39. The sequence of residues 103 to 117 (corresponding to most of the first helix of the core) can be aligned with two regions of the linker region (40 to 54 and 59 to 73) to overlap three and four serines, respectively. Since the linker region sequences contain serine repeats, it is unclear if this implies structural homology.
40. C.-I. Brändén, *Q. Rev. Biophys.* **13**, 317 (1980).

41. J. Rouvinen, L. Ruohonen, T. Teeri, J. K. C, Knowles, T. A. Jones, in preparation.
42. D. E. Koshland, Jr., *Biol. Rev.* **28**, 416 (1953).
43. M. Yaguchi, C. Roy, C. F. Rollin, M. G. Paice, L. Jurasek, *Biochem. Biophys. Res. Commun.* **116**, 408 (1983).
44. C. C. F. Blake *et al.*, *Nature* **206**, 757 (1965); D. C. Phillips, *Proc. Natl. Acad. Sci. U.S.A.* **57**, 484 (1967); an alternative mechanism involving ring breaking has been proposed by C. B. Post and M. Karplus [*J. Am. Chem. Soc.* **108**, 1317 (1986)]. The lysozyme mechanism has been reviewed by A. J. Kirby [*CRC Crit. Rev. Biochem.* **22**, 283 (1987)].
45. S. J. Remington *et al.*, *J. Mol. Biol.* **118**, 81 (1978).
46. W. F. Anderson, M. G. Grutter, S. J. Remington, L. H. Weaver, B. W. Matthews, *ibid.* **147**, 523 (1981).
47. Y. Matsuura, M. Kusunoki, W. Harada, M. Kakudo, *J. Biochem. (Tokyo)* **95**, 697 (1984).
48. G. Buisson, E. Duree, R. Haser, F. Payan, *EMBO J.* **6**, 3909 (1987).
49. B. Svensson, A. J. Clarke, I. Svendsen, H. Møller, *Eur. J. Biochem.* **188**, 29 (1990).
50. R. Nakai, S. Horinouchi, T. Beppu, *Gene* **65**, 229 (1988).
51. M. D. Yablonsky, K. O. Elliston, D. E. Eveleigh, in *Enzyme Systems for Lignocellulose Degradation*, M. P. Coughlan, Ed. (Elsevier Applied Science, New York, 1989).
52. Figure 3 is drawn with the program RIBBON; J. P. Priestle, *J. Appl. Crystallogr.* **21**, 572 (1988).
53. B. Lee and F. M. Richard, *J. Mol. Biol.* **55**, 379 (1971). The accessible surface refers to the closest approach that a water molecule can make to the protein. It is obtained by rolling a probe of radius 1.4 Å on the van der Waals surface of the protein. The contours shown in the figure were obtained from an "electron density" map file [R. Voointholt, M. T. Kosters, G. Vegter, G. Vriend, W. G. J. Hol, *J. Mol. Graph.* **7**, 243 (1989)].
54. Supported in part by the Swedish Natural Sciences Research Council and the Academy of Finland (J.R.). We thank M. Claeyssens for providing the iodine-labeled inhibitor compound, and J. Knowles, Cambridge, Massachusetts, for constructive comments. Coordinates have been deposited at the Protein Data Bank, Brookhaven.

5 February 1990; accepted 13 June 1990

Searching for Peptide Ligands with an Epitope Library

JAMIE K. SCOTT AND GEORGE P. SMITH*

Tens of millions of short peptides can be easily surveyed for tight binding to an antibody, receptor or other binding protein using an "epitope library." The library is a vast mixture of filamentous phage clones, each displaying one peptide sequence on the virion surface. The survey is accomplished by using the binding protein to affinity-purify phage that display tight-binding peptides and propagating the purified phage in *Escherichia coli*. The amino acid sequences of the peptides displayed on the phage are then determined by sequencing the corresponding coding region in the viral DNA's. Potential applications of the epitope library include investigation of the specificity of antibodies and discovery of mimetic drug candidates.

"Fusion phage" are filamentous bacteriophage vectors in which foreign antigenic determinants are cloned into phage gene III and displayed as part of the gene III protein (pIII) at one tip of the virion. Fusion phage whose displayed determinant binds an antibody (Ab) can be selected from a vast background of nonbinding phage by affinity purification (AP) as follows (*1*). First, phage are reacted with biotinylated Ab (bio-Ab), then diluted and placed on a streptavidin-coated petri dish, thereby specifically attaching Ab-reactive phage to the plastic surface through the Ab-biotin-streptavidin bridge. Free phage are washed away, and bound phage eluted in acid and used to infect *Escherichia coli* cells. A single round of AP can enrich Ab-binding phage by as much as a factor of 10^5 relative to unreactive phage; further enrichment is achieved by further rounds of AP after amplification on agar medium (*1*). Thus Ab serves as a powerful selective agent favoring the target clones, so that vast numbers of phage can be surveyed.

The idea of using fusion phage to develop an "epitope library" (*1, 2*) was inspired by the synthetic "mimotope" strategy of Geysen *et al.* (*3*). By synthesizing peptide mixtures on plastic pins and assessing their ability to bind an Ab against a protein antigen, these workers delineated a peptide that mimicks a discontinuous epitope—an Ab-binding determinant composed of residues distant in the primary sequence but adjacent in the folded structure. They called these peptide mimics mimotopes. In this way ligands can be discovered for an Ab whose specificity is not known in advance.

Fusion phage displaying short cloned peptides are infectious analogs of chemically synthesized mimotopes, with the key advantages of replicability and clonability. A large library of such phage—an "epitope library"—may display tens of millions of peptide epitopes. The peptides can in effect be individually surveyed for binding to an Ab or other binding protein by affinity purifying reactive phage from the library, progagating individual phage clones, and sequencing the relevant part of their DNA's to determine the amino acid sequences of their displayed peptides. A survey based on the epitope library undoubtedly would be imperfect because of bias introduced by the biology of the phage and other factors; still, it would represent a powerful new approach to the study of the specificity of Ab's and other binding proteins.

In this article we report on construction and characterization of epitope libraries. Our library is a mixture of fusion phage theoretically displaying approximately 4×10^7 different hexapeptide epitopes. Devlin *et al.* (*4*) and Cwirla *et al.* (*5*) have also described the construction and characterization of epitope libraries.

Construction of the library. We constructed phage fUSE5 as the vector for the epitope library (Fig. 1A); it has several advantages in common with other vectors in the fUSE series (*1*), including vector

The authors are in the Division of Biological Sciences, Tucker Hall, University of Missouri, Columbia, MO 65211; J. Scott is also in the Department of Medicine, Health Sciences Complex, University of Missouri, Columbia, MO 65212.

*To whom correspondence should be addressed.