Publication IV

Binding properties of single and double cellulose binding modules reveal differences between cellulose substrate. Suvi Arola and Markus B. Linder. Submitted to *Nature Chemical Biology.*

Binding properties of single and double cellulose binding modules reveal differences between cellulose substrate

Suvi Arola[1], and Markus B. Linder[2*]

[1]School of Science, Aalto University, P. O. Box 11100, FI-00076, Aalto, [2] School of Chemical Technology, Aalto University, P .O. Box 16100, FI-00076, Aalto

[*]To whom correspondence should be addressed: Markus B. Linder (Aalto University)
markus.linder@aalto.fi, aalto.hyber.fi

Abstract

Here we report for the first time that the binding of *Trichoderma reesei* major cellulase Cel6A and Cel7A cellulose binding modules (CBM) have a very different way of binding to the substrates originating from wood compared to microbial cellulose. Our findings indicate that the CBM proteins see the two substrates, nanofibrillated cellulose and bacterial microcrystalline cellulose, as very different from one another. We show that the substrate has a large impact on the exchange rate of the two CBM, and moreover, that CBM-Cel7A seems to have an extra mode of binding on NFC but not on BMCC. This mode is not translated to the double CBM comprising of CBM-Cel7A and CBM-Cel6A. These results have impact on the cellulase research mostly done with non-wood cellulose substrates and offer new understanding on how these industrially relevant enzymes act.

Introduction

Carbohydrate binding modules (CBMs) are found in a large variety of proteins that are active in one way or another on plant cell walls. These are mainly cellulases but also for example xylanases and mannanases[1–4]. There are several examples of convergent evolution leading to

several different protein families with similar function[5]. Especially in bacterial these families are very diverse[6]. Analogously the CBMs can show affinity for chitin and are also found as parts of chitin active proteins[7]. In this work we focus on the type of CBM that is found in fungi and are classified as belonging to Family 1[2,8,9]. These have been the subject of much investigation because of the efficiency and importance of fungal enzyme systems for degrading cellulosic material both as a part of the ecosystem and also increasingly for technical applications[10,11]. These fungal CBMs have a compact structure comprising of only about 35 amino acids[12]. The structure has a stable fold that is referred to as a cysteine knot and which is also found in other small adhesion proteins such as conotoxins[13]. As the name indicates, cysteine knots are stabilized by disulfide bonds and in the case of family 1 CBMs, two or three of them are found. On one face of the folded Family 1 CBM proteins there is a distinctive arrangement of three aromatic residues that has been shown to play a key role in cellulose adhesion. The interaction between aromatics and pyranose rings is widely observed also elsewhere, and it has also been shown that by changing the character of the aromatic residues also the binding characteristics can be changes[14]. In addition to these pi-electron interactions, also hydrogen bonding is involved in forming affinity and specificity between protein and cellulose[12]. The binding is highly specific and shows even selectivity for the different crystalline faces of cellulose[15,16].

The CBMs have proved to play an essential role for how enzymes function. There are effects on both substrate recognition and catalytic activities.[17] The most straightforward explanation of their function is to guide the enzyme towards the substrate and enhance local concentrations on surfaces, but in many cases the role have been found to be more subtle[2]. Another reason for an interest in these CBMs is that they provide a way to modify cellulose through a self-assembly mechanism[18,19]. This can be utilized for protein immobilization[20]. However more recently there

has been a large interest in cellulose as a nanoscale material[21]. By either acid hydrolysis or by mainly mechanical disintegration cellulose fibrils and crystallites are released [22–27]. Because cellulose as a polymer has very attractive mechanical and physical properties many types of advanced materials have been made, including photonic materials and high performance composites[28,29]. Interestingly for example the strength of cellulose surpasses steel on a weight-to-weight comparison.

In this work we set out to gain understanding of both cellulose as a structure and how macromolecular architectures can be designed for adhesion to it by a protein engineering approach using family 1 CBMs. It was already known that linking two CBMs together does affect the overall interaction to a very large extent and that this two-domain interaction can to some extent explain the architecture of cellulose degrading enzymes in general[7]. The thermodynamic principles behind such cooperativity are well known [30] and depend on geometric constraints and the general architecture of the complexes that are formed. We hypothesized that making a series of different linkers between two CBMs could reveal how binding sites are located on the cellulose surface and how the linking of binding modules changes the dynamics of interactions. Using a set of differently designed molecules, different sources of cellulose, and accurate measurement techniques enabling the study of binding kinetics allowed a new set of tools to investigate cellulose for use as a material or for the enzymatic breakdown of it.

Results

Analysis of proteins

DCBM-12, -24, and 48 were produced as HFBI-DCBM fusions to aid the purification. Different DCBM proteins were gained by trypsin cleavage of the corresponding HFBI-DCBM.

The CDB-Cel7A and CBM-Cel6A were gained by papain cleavage of HFBI-DCBM-12. Amino acid analysis (AAA) of the fractions from RP-HPLC run after papain cleavages showed the identity of the two fractions; i.e. which of them CBM-Cel7A was and which CBM-Cel6A.

Matrix-assisted laser desorption/ionization mass spectrometry (MALDI-TOF MS) was used to verify the amount of linker region present in the proteins and the extent of glycosylation of the proteins. According to MALDI-TOF DCBM-24 contained14-25 glycan groups and DCBM-48 34-50. Both of the proteins contained multiple O-glycosylation sites on the linker (T, S, see figure S1 for sequences), DCBM-48 double the amount of DCBM-24. The glycosylation patterns were heterogenic and showed a broad peak at m/z ~12730-14500 (DCBM-24) and m/z ~18400-20730 (DCBM-48). DCBM-12 and the single CBMs were not glycosylated. There were no broad peaks on their MALDI-TOF spectra nor do the peaks on the spectra contain differences by m/z 162 (hexose) or m/z 203 (GlcNAc). The multiple peaks on their spectra come from a varying lengths of linker attached on them. Especially the CBM-Cel6A protein that was cleaved by papain showed variation in the linker length on both sides, the main peak was at m/z 5414. CBM-Cel7A main peak was at m/z 3850 (CBM-Cel7A + serine from linker) and minor peaks referring to CBM-Cel7A only (m/z 3763) and CBM-Cel7A + one to four threonine residues. DCBM-12 has a major peak at m/z 9039 and minor peaks showing that trypsin has cleaved on both sides of the arginine in the linker sequence between HFBI and DCBM-12.

Binding isotherms

Figure 1 shows the binding isotherms of CBMs and DCBMs on A: BMCC, and B: on NFC. On BMCC an increase in affinity was seen for DCBM-12 and DCBM-24 constructs compared to single CBMs. Linker length affects the binding of the double constructs; DCBM-48 construct

shows lower affinity compared to DCBM-12 and DCBM-24 and CBM-Cel7A. DCBM-24 shows the highest affinity.

On NFC for the DCBMs the linker length affected the binding in the same way as on BMCC; DCBM-24 and DCBM-12 having a higher affinity than DCBM-48 and CBM-Cel6A having the lowest affinity. On NFC there was no advantage of linking CBM-Cel7A and CBM-Cel6A together because CBM-Cel7A had the highest affinity on its own. Table 1 summarizes the partitioning coefficients, K_r, obtained from the data in the beginning of the isotherm. The corresponding affinities, k_d, and maximum binding capacities, B_{max}, are presented in supplementary table 1. Table 2 shows that the correlation of K_r-values of the five proteins between two substrates. This correlation is consistent for all other proteins being 1.5 except for CBM-Cel7A being much higher 3.6. This indicates clearly that CBM-Cel7A behaves differently on the two substrates compared to the other proteins.

The Gibbs free energy of binding, ΔG

For the calculation of Gibb's free energy for binding the binding capacity, B_{max}, is needs. Binding isotherms with larger protein concentrations (up to 50 µM free protein concentration) are presented for all proteins on both substrates in semi-logarithmic plots in figure 2A-D. The maximum binding capacities obtained from this data are presented in table 3. The linear forms of the binding isotherm graphs are presented in supplementary figure 2. The data with higher protein concentrations shows that CBM-Cel7A binds with a greater capacity than any of the other proteins on both substrates. The data also shows that the single CBMs are far from their actual maximum binding capacities, although substantial amount of protein was used in the experiments. On the other hand, it can be seen that the maximum binding capacity of DCBM-48

has been reached on both substrates, and that DCBM-12 and -24 are not far from their maximum capacities. These values for DCBM-48 can be used to calculate accurate Gibb's free energies, ΔG, for the binding event using equation (4) presented in Online methods. To calculate an estimated range of the ΔG-value for the other proteins on both substrates, we used the K_r-values obtained from the initial slope of the data and varied the B_{max}-value in equation (4). We used three B_{max}-values; B_{max}, low, which was obtained from data in figure 2 and which is lower than the actual B_{max} for all other proteins than DCBM-48, a double that of the low value (B_{max}, intermediate), and a triple that of the low value (B_{max}, high). A schematic illustration of the three B_{max} –value levels are presented in supplementary figure 3 with the binding isotherm for CBM-Cel7A on NFC to illustrate the range of B_{max}-values.

Table 4 summarizes ΔG-values for the different binding events. From table 4 it is evident that the free energies of binding are lower for DCBM-proteins than for CBM-proteins. Thus the binding of DCBMs is more favorable than CBMs. The differences for the ΔG-values are not very large between different DCBMs. We can also see that ΔG-values for CBM-CeI7A on NFC and BMCC are very close to each other and also to the ones obtained for CBM-Cel6A. The ΔG-values for DCBMs are also very close to each other for both substrates indicating that the binding is energetically similar for the double constructs on both substrates. Since the peculiar binding of CDB-Cel7A cannot be explained by energetic differences of the binding events, it seems to a matter of geometry or the substrate or both.

Xylanase assay for CBM-Cel7A

There is roughly 27 % xylan in the NFC material that we use. Of this, 30 % can be specifically hydrolyzed by pI9 xylanase from *T. reesei*, which corresponds to about 10 % loss of

the total mass. To test how the xylan on NFC affects the CBM-Cel7A binding we used xylanase pI9 enzyme to hydrolyze xylan from NFC and tested the binding of CBM-Cel7A on this xylanase treated NFC compared to non-treated NFC. The binding of CBM-Cel7A increased about 20% on the xylanase treated NFC compared to the control. These results suggest that the xylan hydrolysis opens up substantial amount of binding sites on NFC for CBM-Cel7A rather than the CBM-Cel7A binding to xylan.

Nanoscale effect of the NFC substrate

Since NFC has not been widely used as a substrate for CBMs in binding studies it could be that the nanoscale structure somehow induces a boost in CBM-Cel7A affinity. To test this we performed binding studies on the pulp material that the NFC was prepared from. The binding isotherm is shown in supporting information figure S4. The order of binding is the same for the proteins CBM-Cel7A having greater affinity and capacity and thus partitioning coefficient than the DCBMs and CBM-Cel6A. The K_r-values are very close to those obtained for NFC and thus the binding is the same regardless of the nanoscale structure.

Exchange rates of CBM and DCBM on NFC and BMCC

Next, to investigate the kinetics of the bindings, we performed a series of experiments to measure the extent of exchange of proteins from the substrate surface. At equilibrium exactly the same amount of non-labelled protein initially used for the equilibrium reaction at concentration equal to free protein concentration was added to reaction mixture. The amount of ^{3}H-protein in solution was monitored during time and compared to the original amount of ^{3}H-protein at equilibrium state. The results presented in figure 5 shows that the only combination of protein

and substrate that fully exchange to 50% from the original amount of ^{3}H-protein bound are CBM-Cel7A and CDB-Cel6A on BMCC. The exchange of these two proteins on BMCC is fast and fully reversible. On the other hand, none of the DCBMs on BMCC exchanged fully during the time course of the experiment (3600 s). DCBM-12 and DCBM-24 exchange roughly to 10% and DCBM-48 exchanges roughly to 30%. The coupling of the CBMs together clearly hinders the exchange rate. On NFC none of the proteins fully exchange to 50% during the time course of the experiment. The CBMs exchange roughly to 20% showing that the exchange rate is a lot slower for these proteins on NFC compared to BMCC. The DCBMs exchange even slower on NFC than the CBMs showing again that the coupling of the individual pieces together affects the exchange rate by slowing it down. The DCBM-24 seems not to exchange at all from NFC during the time course of the experiment and DCBM-12 and -48 exchanges roughly to 5-10%.

Discussion

By studying the binding initially on BMCC we observed that there was a significant effect of the linker length to the binding properties of the DCBM proteins. The data at low protein concentrations was highly reproducible between experiments due to the excellent accuracy of analytical technique based on scintillation counting (Figure 1). As a measure of affinity we calculated partitioning coefficients ($K_r = B_{max}/k_d$). This coefficient was obtained by fitting the equation for a one-site Langmuir isotherm (equation (1)) to the data and using the values for k_d and B_{max} that were obtained. We note that the uncertainty of the B_{max} (and hence the k_d) is large when only using initial parts of the isotherms for fitting, as will be discussed below. However, the ratio K_r can be reliably estimated since it is mathematically equal to the limit of the derivative of the one-site Langmuir equation as the concentration of free protein approaches zero.

Since this part of the curve is accurate we obtain a reliable value for this ratio. The ratio can also be thought of as the partitioning of free and bound protein at low concentrations when competition for binding sites is negligible. As presented in Table 1 the DCBM-24 shows the highest K_r, closely followed by DCBM-12. Interestingly the CBM-Cel7A showed an affinity that was higher than the DCBM-48, and the CBM-Cel6A showed the lowest K_r. When the experiments were repeated with birch wood NFC a surprising difference was noticed. For NFC the clearly highest partitioning was found for CBM-Cel7A. The rest of the constructs showed the same relative order of affinity although the affinities were higher overall. To investigate the difference between BMCC and NFC more closely we calculated the ratio of K_r-values for each protein at the two different substrates (Table 2). The ratios for the partitioning coefficients were about 1.5 for all proteins except CBM-Cel7A. This indicates that the substrates are seen by these proteins in a very similar way, except that the NFC as expected has 1.5 times more surface area per weight than the BMCC. The CBM-Cel7A on the other hand has a very high partitioning coefficient on NFC, but also the ratio as calculated in Table 2 is over twice that of the other constructs. We can deduce that the CBM-Cel7A can find some binding mode or site on NFC that is not found on BMCC, and which is not found by CBM-Cel6A. Some feature in NFC is available for CBM-Cel7A that does not exist in BMCC, but all other constructs behave identically in relation to each other on the two substrates. It is especially notable that the synergies between modules seen for the DCBM-12 and DCBM-24 constructs are not able to incorporate this extra mode of binding for CBM-Cel7A.

Next we evaluated if full isotherms, allowing the calculation of binding energies would be feasible. Data points were collected at very high protein concentrations. These data points are showed in semi-log plots in Figure 2. The highest protein concentrations were around 50 µM. At

these concentrations only about 1-10 % (DCBM) or 20 % (CBM) of the total protein are bound to cellulose. Due to the high overall protein concentrations and low percentage of total protein bound to cellulose it was not feasible to acquire data points at even higher concentrations. We note that for the DCBM constructs the capacity levels off to a value that can be reliably determined but in the case of the free modules the estimated capacities may be too low. Because these values are likely to be too low it can lead to an error in the estimation of change in ΔG.

The change in free energy was calculated using standard equations (equation (4) in methods). We then used upper and lower estimates of B_{max} (see Supplementary Fig. 3) to evaluate the value of ΔG (Table 4). This allowed us to estimate the degree of synergy in the DCBM constructs. The ΔG values show clearly that in all DCBM constructs there is a synergy that comes from the linkage. It is thus energetically favorable for the proteins to be linked together.

We next considered how differences between the two substrates could explain the observations. The most obvious difference between NFC and BMCC is their polysaccharide composition: BMCC is pure cellulose whereas NFC has, in addition to cellulose, roughly 27% xylan[31]. To try and understand the difference in binding sites available for CBM-Cel7A on NFC but not on BMCC, we thus investigated how the hydrolysis of xylan affected the binding. This was done by using xylanase to hydrolyse xylan from NFC and to study the binding of CBM-Cel7A on this substrate. The results showed that the binding of CBM-Cel7A was increased about 20% by the xylanase treatment. This indicates that the xylan hydrolysis opens up substantial amounts of binding sites on NFC for CBM-Cel7A. From this it seems not plausible that the greater binding affinity of CBM-Cel7A on NFC would be caused by an affinity towards xylan. However, it could be that CBM-Cel7A binds to xylan but when xylan is released it binds to cellulose with a much larger affinity. In this case we could expect to see small differences in the

binding behavior of CBM-Cel7A to hydrolyzed and non-hydrolyzed NFCs since surface area of the fibrils is not changed drastically by hydrolysis of xylan from the fibril surface[31], and it is not expected that these proteins bind elsewhere in the fibril than on the surface[16,17]. On the other hand, it seems still more logical that xylan inhibits the binding of CBM by masking the cellulose surface. This has been shown with cellulases, that xylanase are needed in order for the cellulases to fully hydrolyze NFC[32]. We show that the binding of CBM-Cel7A is greatly affected by xylan hydrolysis and this also suggests that at least partly the xylan available for the xylanase is located on the surface of the fibril at the CBM binding sites.

As the peculiar binding of CBM-Cel7A could not be explained by the xylan content, we sought answers from another difference between the substrates: the nanoscale architecture. NFC from wood source has not been widely used for CBM binding studies; usually substrates include BMCC, Avicel, tunicin and Valonia cellulose. The mechanical processing of pulp to NFC and the nanoscale structure could potentially have affected the substrate in adding features, which are not present in BMCC. To investigate this, we performed similar binding studies on the pulp material from which the NFC is prepared. The results showed that pulp has much less surface are for the proteins to bind but it is seen as the same material as NFC by the proteins since the partitioning coefficients of the proteins on pulp were very close to those obtained for NFC and the order of the binding was the same; namely CBM-Cel7A having the greatest affinity. The results show that cellulose substrate from wood source, no matter how fine or coarse the structures, is seen in a different way by CBM-Cel7A than BMCC. The source of the material seems to count for the protein.

A third difference between the two substrates is their crystalline structure; wood cellulose crystalline regions composing mainly of cellulose I_β and BMCC of cellulose I_α. Also wood

cellulose amorphous to crystalline ratio is different than that of BMCC; BMCC being more crystalline.[26,33,34] This could be of great significance for the two proteins having a difference of one amino acid residue in their binding faces[7,12,35]. We next studied the exchange rate of the different constructs on the two substrates in order to evaluate if they show differences. This was done by equilibrium exchange experiments where the extent of exchange of ^{3}H-labelled protein is determined by competition with unlabeled protein. From the results presented in Figure 3 it was evident that the coupling of the individual CBM together greatly reduced the extent of exchange. On BMCC both CBM alone fully exchange to 50% of the original within the time scale of the experiment (<600 s) and thus the binding can be considered truly reversible and fast (Figure 3 B). The linking slowed the exchange rate and the extent of exchange to only 10-30% for the DCBM constructs (Figure 3 D). On NFC, the exchange of the individual CBM did not reach 50% from the original amount of bound ^{3}H-protein (Figure 3 A) and the exchange was thus markedly slower than for BMCC. This more slow exchange from NFC for the CBM-Cel7A and CBM-Cel6A both could be due to the different structure of NFC compared to BMCC. The exchange of DCBM constructs was even slower from NFC than those of CBM (Figure 3 C) and thus consistent with results on BMCC: linking the two CBM together slows down the exchange rate. These results suggest that CBM-Cel7A and CBM-Cel6A both experience NFC as a different substrate compared to BMCC, yet is doesn't provide a good explanation on why the CBM-Cel7A has a boost in affinity on NFC and why the DCBM proteins do not exhibit this property.

One way would be to experiment whether the two proteins compete on binding sites on the two substrates fully, partially or not at all and if there would be differences between the competitions on different substrates, i.e. to see if the binding of the one is affected by the

presence of the other. This was examined by competing labelled CBM with unlabeled CBM of the same kind and of different and comparing the results (experimental and data in supplementary materials, supplementary Fig 6.). The results showed that both proteins are affected by the presence of the other in a same way as they are affected by the presence of the unlabeled protein of the same kind. This suggests full competition of the CBM-Cel7A and CBM-Cel6A on both substrates and it does not provide straight forward answers to the deviant binding behavior of CBM-Cel7A on NFC.

In conclusion, we were able to show that the linker length plays a critical role in protein constructs with two binding units, in our case the CBM proteins. We showed that the linking provides an increase in affinity at certain lengths for DCBM constructs and it increases the affinity when the proteins binding to BMCC. On the other hand, we were able to show that CBM-Cel7A binding to NFC substrate behaved in a very different way than was expected. The CBM-Cel7A had an increased affinity compared to the DCBM and to the CBM-Cel6A, and the linking of the individual CBM proteins was not favorable affinity or capacity vise. However, the linking of the CBM proteins together was shown to be energetically more favorable, and is beneficial in this regard.

The increased affinity of CBM-Cel7A was shown not to be induced by the nanoscale structure of NFC nor the xylan content. However, we were able to show that the exchange rate of the proteins is greatly dependent on the substrate, since individual CBD proteins exchange reversibly from BMCC but not from NFC. Moreover, the extent of exchange was dependent on the linking; the DCBM constructs exchanged much slower than the CBM proteins on both substrates.

Our findings indicate that the CBM-Cel7A sees the two substrates, NFC and BMCC, as very different from one another. It seems to have another binding mode on NFC rather than another binding site, since we saw no evidence on non-competitive binding of the two CBM on NFC. The other mode of binding of Cel7A-CBM is not translated to the DCBM constructs, since the affinity of DCBM constructs on NFC is lower than that of CBM-Cel7A. It seems that the linking of CBM-Cel7A together with CBM-Cel6A inhibits the other binding mode of CBM-Cel7A, and causes the DCBM construct to behave more like a double CBM-Cel6A. We show for the first time that for the binding of CBM onto substrate the substrate origin matters greatly and results gained from experiments on one substrate are not comparable to results gained with another substrate, or at least the comparison should be done by caution.

Our results can offer new consideration for cellulase studies, especially to the energetics of the cellulase binding. It is clearly more favorable for the enzyme to bind to a substrate together with the CBM than without it[36,37]. The gain seems to be also in the energetics of the binding because the DCBM binding, regardless of the linker length, is more favorable than the binding of an individual CBM. Our results also raise a question whether it is somehow evolutionary that the CBM-Cel7A recognizes wood cellulose with a higher affinity than other cellulose materials because wood is the primary source of energy for the organism is originates from?

Acknowledgements

This work was funded by The Academy of Finland and The National Graduate School for Biomass Refining. This work was carried out under the Academy of Finland's Centre of Excellence Program (2014–2019) and supported by project # 259034 from the Academy of Finland.

Authors Contribution

S.A. planned and carried out all experimental work, except for protein production that was conducted by Michael Bayley, and protein cleavage and subsequent protein purificaton which was done by Riitta Suihkonen. S.A. prepared all figures. S.A. and M.B.L. wrote the manuscript together. M.B.L. supervised the research.

Figure Legends

Figure 1. **(A)** The beginning of the binding isotherms of Ce-7A, Cel6A, DCBM-12, DCBM-24, and DCBM-48 on BMCC. **(b)** The beginning of the binding isotherms of Ce-7A, Cel6A, DCBM-12, DCBM-24, and DCBM-48 on NFC. (+) CBM-CeI7A, (x) CBM-Cel6A, (□) DCBM-12, (○) DCBM-24, (Δ) DCBM-48.

Figure 2. **(A)** Semi-logarithmic plots of the binding isotherms of CBM-Cel7A and CBM-Cel-6A on NFC. **(B)** Semi-logarithmic plots of the binding isotherms of CBM-Cel7A and CBM-Cel-6A on BMCC. **(C)** Semi-logarithmic plots of the binding isotherms of DCBM-12, DCBM-24, and DCBM-48 on NFC. **(D)** Semi-logarithmic plots of the binding isotherms of DCBM-12, DCBM-24, and DCBM-48 on BMCC. (+) CBM-Cel7A, (x) CBM-Cel6A, (□) DCBM-12, (○) DCBM-24, (Δ) DCBM-48.

Figure 3. (A) Exchange of ^{3}H-labelled CBM-Cel7A and CBM-Cel6A from NFC. **(B)** Exchange of ^{3}H-labelled CBM-Cel7A and CBM-Cel6A from BMCC. **(C)** Exchange of ^{3}H-labelled DCBM-12, DCBM-24, and DCBM-48 from NFC. **(D)** Exchange of ^{3}H-labelled DCBM-12, DCBM-24, and DCBM-48 from BMCC. (+) CBM-CBHI, (x) CBM-CBHII, (□) DCBM-12, (○) DCBM-24, (Δ) DCBM-48

Tables

Table 1. The partitioning coefficients ($K_r = B_{max}/K_d$) for single and double CBM obtained from the Langmuir isotherm fitted to the data on Figure 1 A and B

	CBM-Cel7A	CBM-Cel6A	DCBM-12	DCBM-24	DCBM-48
BMCC	1,4	0,61	2,47	2,82	1,07
NFC	4,98	1,05	3,87	4,18	1,67

Table 2. The correlations of partitioning coefficients, K_r, for single and double CBM between NFC and BMCC calculated from K_r-values in table 1

	CBM-Cel7A	CBM-Cel6A	DCBM-12	DCBM-24	DCBM-48
K_r^{NFC}/K_r^{BMCC}	3,57	1,72	1,56	1,49	1,56

Table 3. Binding capacities, B_{max}, for single and double CBM on NFC and BMCC obtained from the data shown in figure 2*

	CBM-Cel7A	CBM-Cel6A	DCBM-12	DCBM-24	DCBM-48
BMCC	8,20	5,78	4,01	4,09	1,07
NFC	18,24	10,74	9,21	8,61	4,05

*The unit of B_{max} is µmol/g

Table 4. Gibb's free energy difference for the binding, ΔG, for CBM-Cel7A, CBM-Cel6A, DCBM-12, DCBM-24, and DCBM-48 on NFC and BMCC.

substrate	B_{max}*	CBM-Cel7A	CBM-Cel6A	DCBM-12	DCBM-24	DCBM-48
NFC	low	-30,7	-28,2	-31,8	-32,1	-31,7
	inter	-29,0	-26,5	-30,1	-30,4	
	high	-28,0	-25,5	-29,1	-29,4	
BMCC	low	-29,6	-28,4	-32,7	-33,0	-33,9
	inter	-27,9	-26,7	-31,0	-31,3	
	high	-26,9	-25,7	-30,0	-30,3	

* B_{max}-value for DCBM-48 from data in figure 2 is an accurate value, thus it can be used to calculate the exact binding energy associated with the binding event. The ΔG-values for CBM-Cel7A, CBM-Cel6A, DCBM-12 and DCBM-24 are estimates (see supplementary information for the estimation, supplementary Fig. 3). B_{max}, low is the value gained from data on figure 2, inter is 2x (B_{max}, low), and high is 3x(B_{max}, low). The unit of ΔG is kJ/mol.

Methods

Protein production and purification

The DCBMs were produced as HFBI-DCBM fusions in *T. reesei* (sequences shown in supplementary information, figure S1) and the following transformants were used: VTT-D-133335 (HFBI-DCBM-12), VTT-D-133336 (HFBI-DCBM-24), and VTT-D-133337 (HFBI-DCBM-48). The strains were then cultivated in 7 L bioreactors on media containing 50 vol-% spent grain extract, 60 g/L lactose, 1 g/L yeast extract, 4 g/L KH2PO4, 2.8 g/L (NH4)2SO4, 0.6 g/L MgSO4 · 7H2O, 0.8 g/L CaCl2 · 2H2O, 2 ml/L trace solution. The pH was let to drop from 5

to about 3 during cultivation. At 24 h intervals 48 mg pepstatin A and 28 mg soy bean trypsin inhibitors (both from Sigma-Aldrich) were added to the cultures to minimize protein degradation. The culture supernatants were separated from the biomass by filtration through GF/A filters (Whatman). Protein expression levels were analyzed by RP-UPLC and were 0.2 g/L, 0.4 g/L, and 3.0 g/L for HFBI-DCBM-12, -24, and -48, respectively. The proteins were purified using aqueous two phase extraction and reverse-phase high-performance liquid chromatography (RP-HPLC) as described earlier29 followed by lyophilization.

Protein preparation

The fusion proteins were cleaved with sequencing grade modified trypsin (Promega) in 25mM Tris-HCl – 150mM NaCl buffer for 2 hours in room temperature to yield DCBMs. The trypsin digestion reaction was followed by RP-UPLC using a 2.1 x 100 mm, 1.7 µm, C4 Acquity BEH300 prST column and an Acquity I-Class system with a photodiode array detector (Waters, MA, USA). The proteins were eluted in a linear mobile phase gradient from 20 – 60% B using water (A) and acetonitrile (B), both containing 0.1% trifluoroacetic acid. Concentrations of the analysed proteins were determined using standard samples with protein concentrations determined by amino acid analysis (Amino Acid Analysis Lab, Uppsala University, Sweden).

In a similar fashion, the HFBI-DCBM-12 was cleaved using papain (Sigma), in 0.1 mM Sodium Phosphate buffer pH 7 for o/n in room temperature to yield CBM-Cel7A and CBM-Cel6A. The papain cleavage was followed and the proteins were purified as described above for the trypsin cleavage. The identity of the individual CBMs were verified by amino acid analysis (Amino Acid Analysis Lab, Uppsala University, Sweden) and matrix-assisted laser desorption/ionization – time of flight mass spectrometry (MALDI-TOF MS, Brucer Autoflex II

MALDI-TOF). MALDI-TOF was also used to analyze the extent of glycosylation of the DCBMs and CBMs.

Protein labelling

All five proteins were labelled with tritium for interaction studies by reductive methylation as reported previously[7,18,38]. 1.9 mg of lyophilized protein was dissolved in 1.9 mL of 0.2 M borate buffer, pH 8.96 and cooled on an ice bath. 13.2 µL of 0.37 % formaldehyde and 100 mCi of 3H enriched NaBH4 (10 Cimmol-1, NET023H100MC, PerkinElmer) in 150 µL of 0.01 M NaOH were added and reacted for 30 minutes, gently mixing every now and then. The reaction was terminated by RP-HPLC. The specific activities were 0.52 Cimmol-1 for DCBM-12, 0.69 Cimmol-1 for DCBM-24 and 1.25 Cimmol-1 for DCBM-48.

Binding isotherms

Binding studies were done essentially in the same way as reported previously[18,38–40]. Shortly, a 100 µM stock solution of DCBM and CBM containing 10 % 3H–labelled proteins in 100 mM sodium acetate buffer (pH 5.0) with 100 mM NaCl and 0.1% BSA was prepared. Solutions with different fusion protein concentrations were made from the stock solution by diluting with the same buffer. The BSA was used to minimize nonspecific binding of the proteins to test tube walls and filters. 100-200 µL of each protein solution was mixed with an equal volume of 1-2 g L-1 NFC in MilliQ–water (or 1.28 g L-1 BMCC) and stirred in ambient temperature with 250 rpm for 1 h. After equilibration, the suspensions were filtered through disposable filters (Millipore, Millex®–GV filter unit, PVDF, hydrophilic, 0.22 µm, 13 mm, non–sterile) and the

amount of free protein was determined by liquid scintillation counter (PerkinElmer, Tri-Carb 2810TR) from the filtrate.

Xylanase assay

The pI9 xylanase of *T. reesei* was used to hydrolyze xylan from NFC[41]. The enzyme specifically hydrolyses xylan from NFC as described by previous reports[31]. The hydrolysis of NFC was carried out for 24h in 45 °C and a control NFC was treated exactly the same but without xylanase. The hydrolysed and non-hydrolysed NFC was used to study the effect of xylan hydrolysis with CBM-Cel7A using a 2.5 µM stock solution of the CBM containing 10% 3H-labelled protein. 100 µL of either the NFC were let to reach equilibrium with 100 µL of the protein solution. The reactions were ended by filtrating and the amount of free protein was measured by liquid scintillation counter.

Exchange rate assay

The exchange rate experiments were done according to previous reports[35,42]. Shortly, to determine the equilibrium exchange rates for CBM and DCBM on both BMCC and NFC, 3H-labelled proteins were allowed to react with the same unbound protein. A 100 µL aliquot of BMCC (1.28 g L-1) or NFC (2 g L-1) was allowed to react with 100 µL of the 3H-labelled protein. When equilibrium was reached the same amount of unlabeled protein was added. The reaction was stopped by filtering the sample after a certain time (11-3600s) from dilution. The concentration of the added protein was equal to the concentration of the free protein in the equilibrium reaction. By this way the equilibrium was not affected, but only the amount of labelled protein in the solution was diluted. The experiments were carried out by taking multiple

time points from a set of parallel samples and comparing to the original equilibrium concentration of the protein.

Calculations used for Gibb's free energy difference

Langmuir one-site binding model is written in equation (1).

$$B = \frac{[B_{max}] \times [Free]}{k_d \times [Free]} \quad (1),$$

where B_{max} is the maximum binding capacity, k_d is the binding affinity, and [Free] is the concentration of the free protein. When the free protein concentration approaches zero the initial slope for Langmuir isotherm is written as follows in equation (2):

$$B'_{[Free] \to 0} = \frac{[B_{max}]}{k_d} = K_r \quad (2),$$

where K_r is the partitioning coefficient and describes the relation of adsorption and desorption at very low protein concentrations. Gibb's free energy is written as follows,

$$\Delta G = RTln(K_d) \quad (3),$$

where K_d is the dissociation constant, which is comparable to the binding affinity k_d, and thus the equation can be written in terms of the B_{max} and K_r, using equation 2, as shown in equation 4,

$$\Delta G = RT(ln(B_{max}) - ln(K_r)) \quad (4),$$

where R is the universal gas constant and T is temperature.

References

1. Henrissat, B. Cellulases and their interaction with cellulose. *Cellulose* **1,** 169–196 (1994).

2. Várnai, A. *et al.* Carbohydrate-Binding Modules of Fungal Cellulases. Occurrence in Nature, Function, and Relevance in Industrial Biomass Conversion. *Adv. Appl. Microbiol.* **88,** 103–165 (2014).

3. Hägglund, P. *et al.* A cellulose-binding module of the Trichoderma reesei β-mannanase Man5A increases the mannan-hydrolysis of complex substrates. *J. Biotechnol.* **101,** 37–48 (2003).

4. Raghothama, S. *et al.* Solution structure of the CBM10 cellulose binding module from pseudomonas xylanase A. *Biochemistry* **39,** 978–984 (2000).

5. Stern, D. L. The genetic causes of convergent evolution. *Nat. Rev. Genet.* **14,** 751–64 (2013).

6. Conant, G. C. & Wagner, A. Convergent evolution of gene circuits. *Nat. Genet.* **34,** 264–266 (2003).

7. Linder, M., Salovuori, I., Ruohonen, L. & Teeri, T. T. Characterization of a Double Cellulose-binding Domain. *J. Biol. Chem.* **271,** 21268–21272 (1996).

8. Linder, M. & Teeri, T. T. The roles and function of cellulose-binding domains. *Journal of Biotechnology* **57,** 15–28 (1997).

9. Tomme, P. & Warren, R. *Cellulose-binding domains: classification and properties*. *ACS Symposium Series, American Chemical Society* (1995). doi:10.1021/bk-1995-0618

10. Viikari, L., Vehmaanperä, J. & Koivula, A. Lignocellulosic ethanol: From science to industry. *Biomass and Bioenergy* **46,** 13–24 (2012).

11. Wilson, D. B. Cellulases and biofuels. *Current Opinion in Biotechnology* **20,** 295–299 (2009).

12. Linder, M. *et al.* Identification of functionally important amino acids in the cellulose-binding domain of Trichoderma reesei cellobiohydrolase I. *Protein Sci.* **4,** 1056–1064 (1995).

13. Daly, N. L. & Craik, D. J. Structural studies of conotoxins. *IUBMB Life* **61,** 144–150 (2009).

14. Linder, M., Lindeberg, G., Reinikainen, T., Teeri, T. T. & Pettersson, G. The difference in affinity between two fungal cellulose-binding domains is dominated by a single amino acid substitution. *FEBS Lett.* **372,** 96–98 (1995).

15. Lehtiö, J. *et al.* The binding specificity and affinity determinants of family 1 and family 3 cellulose binding modules. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 484–489 (2003).

16. Igarashi, K. *et al.* Traffic jams reduce hydrolytic efficiency of cellulase on cellulose surface. *Science* **333,** 1279–1282 (2011).

17. Igarashi, K. *et al.* High speed atomic force microscopy visualizes processive movement of Trichoderma reesei cellobiohydrolase I on crystalline cellulose. *J. Biol. Chem.* **284,** 36186–36190 (2009).

18. Varjonen, S. *et al.* Self-assembly of cellulose nanofibrils by genetically engineered fusion proteins. *Soft Matter* **7,** 2402 (2011).

19. Laaksonen, P. *et al.* Genetic engineering of biomimetic nanocomposites: Diblock proteins, graphene, and nanofibrillated cellulose. *Angew. Chemie - Int. Ed.* **50,** 8688–8691 (2011).

20. Levy, I. & Shoseyov, O. Expression, refolding and indirect immobilization of horseradish peroxidase (HRP) to cellulose via a phage-selected peptide and cellulose-binding domain (CBD). *J. Pept. Sci.* **7,** 50–57 (2001).

21. Klemm, D. *et al.* Nanocelluloses: A new family of nature-based materials. *Angewandte Chemie - International Edition* **50,** 5438–5466 (2011).

22. Fleming, K., Gray, D. G. & Matthews, S. Cellulose crystallites. *Chem. - A Eur. J.* **7,** 1831–1835 (2001).

23. De Souza Lima, M. M. & Borsali, R. Rodlike cellulose microcrystals: Structure, properties, and applications. *Macromol. Rapid Commun.* **25,** 771–787 (2004).

24. Beck-Candanedo, S., Roman, M. & Gray, D. G. Effect of reaction conditions on the properties and behavior of wood cellulose nanocrystal suspensions. *Biomacromolecules* **6,** 1048–1054 (2005).

25. Henriksson, M., Henriksson, G., Berglund, L. a. & Lindström, T. An environmentally friendly method for enzyme-assisted preparation of microfibrillated cellulose (MFC) nanofibers. *Eur. Polym. J.* **43,** 3434–3441 (2007).

26. Pääkko, M. *et al.* Enzymatic hydrolysis combined with mechanical shearing and high-pressure homogenization for nanoscale cellulose fibrils and strong gels. *Biomacromolecules* **8,** 1934–1941 (2007).

27. Nakagaito, A. N. & Yano, H. The effect of morphological changes from pulp fiber towards nano-scale fibrillated cellulose on the mechanical properties of high-strength plant fiber based composites. *Appl. Phys. A Mater. Sci. Process.* **78,** 547–552 (2004).

28. Aulin, C., Salazar-Alvarez, G. & Lindström, T. High strength, flexible and transparent nanofibrillated cellulose–nanoclay biohybrid films with tunable oxygen and water vapor permeability. *Nanoscale* **4,** 6622 (2012).

29. Purandare, S., Gomez, E. F. & Steckl, A. J. High brightness phosphorescent organic light emitting diodes on transparent and flexible cellulose films. *Nanotechnology* **25,** 094012 (2014).

30. Mammen, M., Choi, S.-K. & Whitesides, G. M. Polyvalent Interactions in Biological Systems: Implications for Design and Use of Multivalent Ligands and Inhibitors. *Angew. Chemie Int. Ed.* **37,** 2754–2794 (1998).

31. Arola, S., Malho, J., Laaksonen, P., Lille, M. & Linder, M. B. The role of hemicellulose in nanofibrillated cellulose networks. *Soft Matter* **9,** 1319 (2013).

32. Várnai, A., Huikko, L., Pere, J., Siika-aho, M. & Viikari, L. Synergistic action of xylanase and mannanase improves the total hydrolysis of softwood. *Bioresour. Technol.* **102,** 9096–9104 (2011).

33. Wada, M., Sugiyama, J. & Okano, T. Native celluloses on the basis of two crystalline phase (Iα/Iβ) system. *J. Appl. Polym. Sci.* **49,** 1491–1496 (1993).

34. Czaja, W., Romanovicz, D. & Brown, R. malcol. Structural investigations of microbial cellulose produced in stationary and agitated culture. *Cellulose* **11,** 403–411 (2004).

35. Carrard, G. & Linder, M. Widely different off rates of two closely related cellulose-binding domains from Trichoderma reesei. *Eur. J. Biochem.* **262,** 637–643 (1999).

36. Kim, D. W., Jang, Y. H. & Jeong, Y. K. Adsorption behaviors of two cellobiohydrolases and their core proteins from Trichoderma reesei on Avicel PH 101. *Biotechnol. Lett.* **19,** 893–897 (1997).

37. Ståhlberg, J., Johansson, G. & Pettersson, G. A New Model For Enzymatic Hydrolysis of Cellulose Based on the Two-Domain Structure of Cellobiohydrolase I. *Bio/Technology* **9,** 286–290 (1991).

38. Carrard, G. & Linder, M. Widely different off rates of two closely related cellulose-binding domains from Trichoderma reesei. *Eur. J. Biochem.* **262,** 637–643 (1999).

39. Linder, M. & Teeri, T. T. The cellulose-binding domain of the major cellobiohydrolase of Trichoderma reesei exhibits true reversibility and a high exchange rate on crystalline cellulose. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 12251–12255 (1996).

40. Palonen, H., Tenkanen, M. & Linder, M. Dynamic interaction of Trichoderma reesei cellobiohydrolases Cel6A and Cel7A and cellulose at equilibrium and during hydrolysis. *Appl. Environ. Microbiol.* **65,** 5229–5233 (1999).

41. Tenkanen, M., Puls, J. & Poutanen, K. Two major xylanases of Trichoderma reesei. *Enzyme Microb. Technol.* **14,** 566–574 (1992).

42. Linder, M. & Teeri, T. T. The cellulose-binding domain of the major cellobiohydrolase of Trichoderma reesei exhibits true reversibility and a high exchange rate on crystalline cellulose. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 12251–12255 (1996).

Supplementary Information

Sequences of the HFBI-DCBM proteins

HFBI-DCBM-12
SNGNGNVCPPGLFSNPQCCATQVLGLIGLDCKVPSQNVYDGTDFRNVCAKTGAQPLCCVAPVAGQALLCQTAVGA
ENLYFQGPGASTSTGRGPGGQACSSVWGQCGGQNWSGPTCCASGSTCVYSNDYYSQCLPGANPPGTTTTSTQSH
YGQCGGIGYSGPTVCASGTTCQVLNPYYSQCL

HFBI-DCBM-24
SNGNGNVCPPGLFSNPQCCATQVLGLIGLDCKVPSQNVYDGTDFRNVCAKTGAQPLCCVAPVAGQALLCQTAVGA
ENLYFQGPGASTSTGRGPGGQACSSVWGQCGGQNWSGPTCCASGSTCVYSNDYYSQCLPGANPPGTTTTSQPAT
TTGSSPGPTQSHYGQCGGIGYSGPTVCASGTTCQVLNPYYSQCL

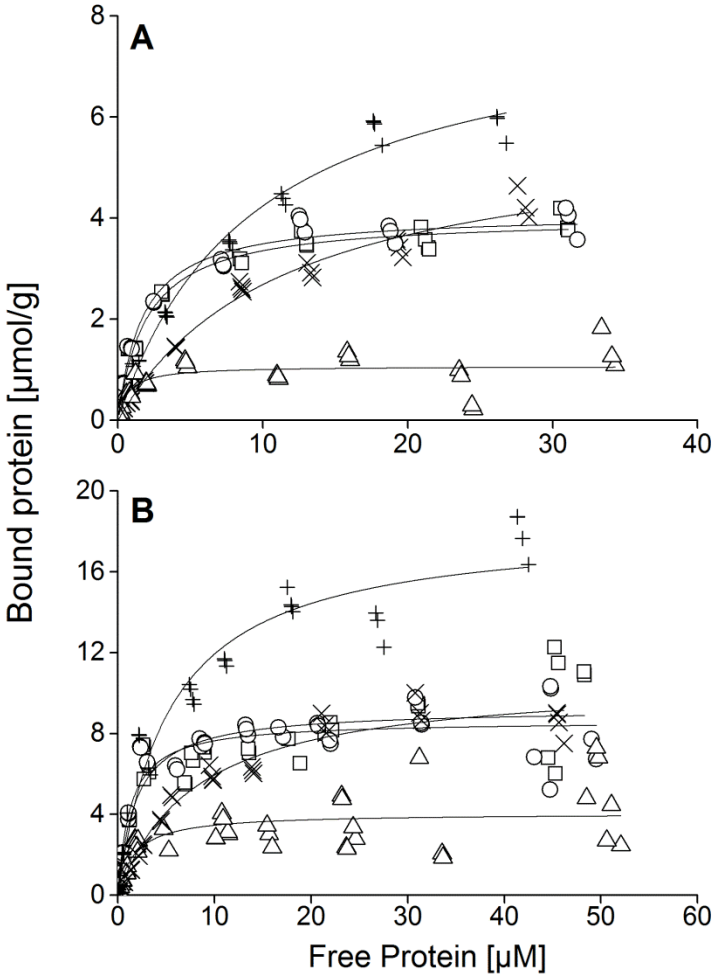
HFBI-DCBM-48
SNGNGNVCPPGLFSNPQCCATQVLGLIGLDCKVPSQNVYDGTDFRNVCAKTGAQPLCCVAPVAGQALLCQTAVGA
ENLYFQGPGASTSTGRGPGGQACSSVWGQCGGQNWSGPTCCASGSTCVYSNDYYSQCLPGANPPGTTTTSQPAT
TTGSSPGPPGANPPGTTTTSQPATTTGSSPGPTQSHYGQCGGIGYSGPTVCASGTTCQVLNPYYSQCL

Supplementary figure 1. Amino acid sequences of the three different HFBI-fusion proteins that were used to gain the DCBM-12, -24, and -48 by trypsin cleavage. DCBM-12 was used to gain CBM-Cel7A and CBM-Cel6A by papain cleavage. Linker regions are in black (the arginine where trypsin cleaves is shown in pink), HFBI sequence is shown in blue, CBM-CEL7A sequence is shown in red, and CBM-CEL6A sequence is shown in green.

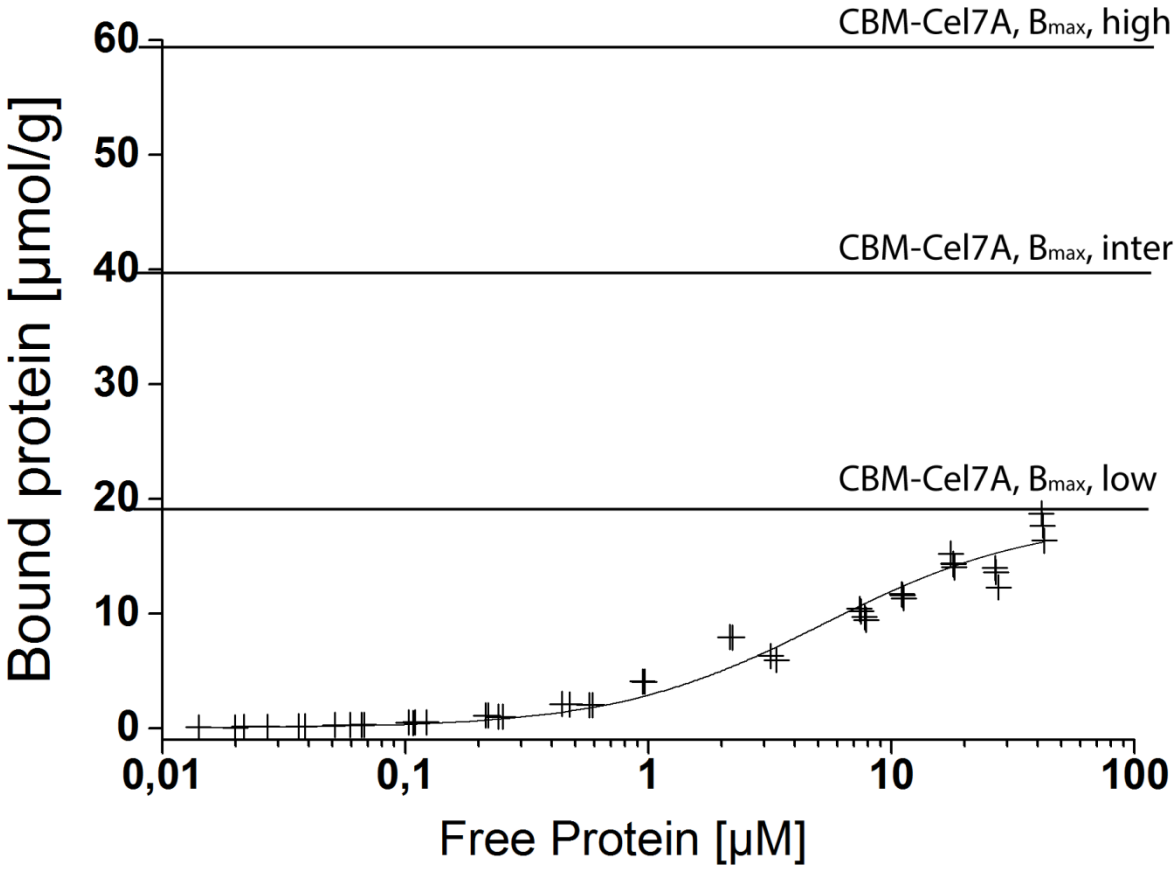
Binding affinities and capacities of CBM and DCBM on BMCC and NFC

Table 1. Binding affinities (k_d) and capacities (B_{max}) for single and double CBM gained from Langmuir one-site binding fit on data in figure 1 of the main text

		CBM-CeI7A	CBM-Cel6A	DCBM-12	DCBM-24	DCBM-48
BMCC	B_{max} ($\mu molg^{-1}$)	4,26	5,15	7,48	7,07	3,19
	k_d (μM)	3,03	8,38	3,03	2,51	2,97
NFC	B_{max} ($\mu molg^{-1}$)	28,96	18,45	26,22	29,36	13,8
	k_d (μM)	5,81	17,58	6,77	7,02	8,31

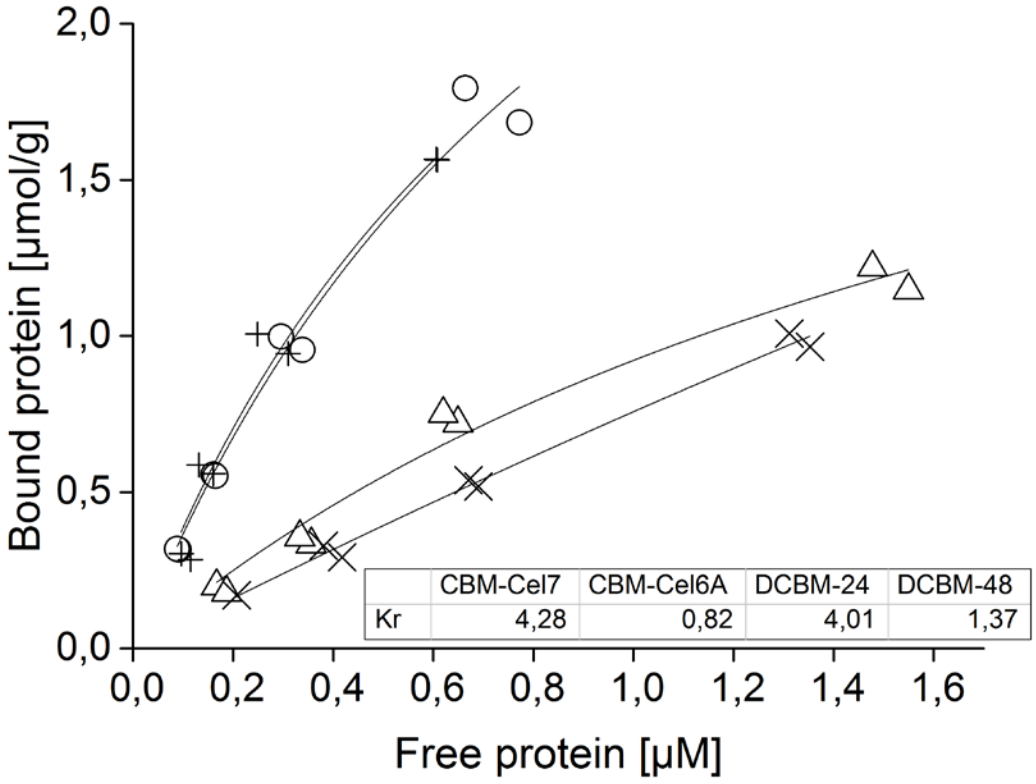


Supplementary figure 2. (A) Linear representation of the binding isotherms of CBM-Cel7A, CBM-Cel6A, DCBM-12, DCBM-24, and DCBM-48 on BMCC. **(B)** Linear representation of the binding isotherms of CBM-Cel7A, CBM-Cel6A, DCBM-12, DCBM-24, and DCBM-48 on NFC. (+) CBM-CBHI, (x) CBM-CBHII, (□) DCBM-12, (○) DCBM-24, (Δ) DCBM-48.



Supplementary figure 3. Schematic illustration of the levels of variation of B_{max}-values shown on semi-log plot of CBM-Cel7A binding isotherm on NFC. The values were used to calculate the variation of free energies of binding for CBM-Cel7A, CBM-Cel6A, DCBM-12, and DCBM-24. B_{max}, low presents the highest value of the maximum capacity that was obtained experimentaly. This value is not an actual maximum and thus represents a minimum which is always exeeded when the cellulose surface is fully covered. B_{max}, intermediate (B_{max}, inter) is double that of B_{max}, low and B_{max}, high is thriple that of B_{max}, low.

Binding isotherms of CBM and DCBM on pulp used to prepare NFC

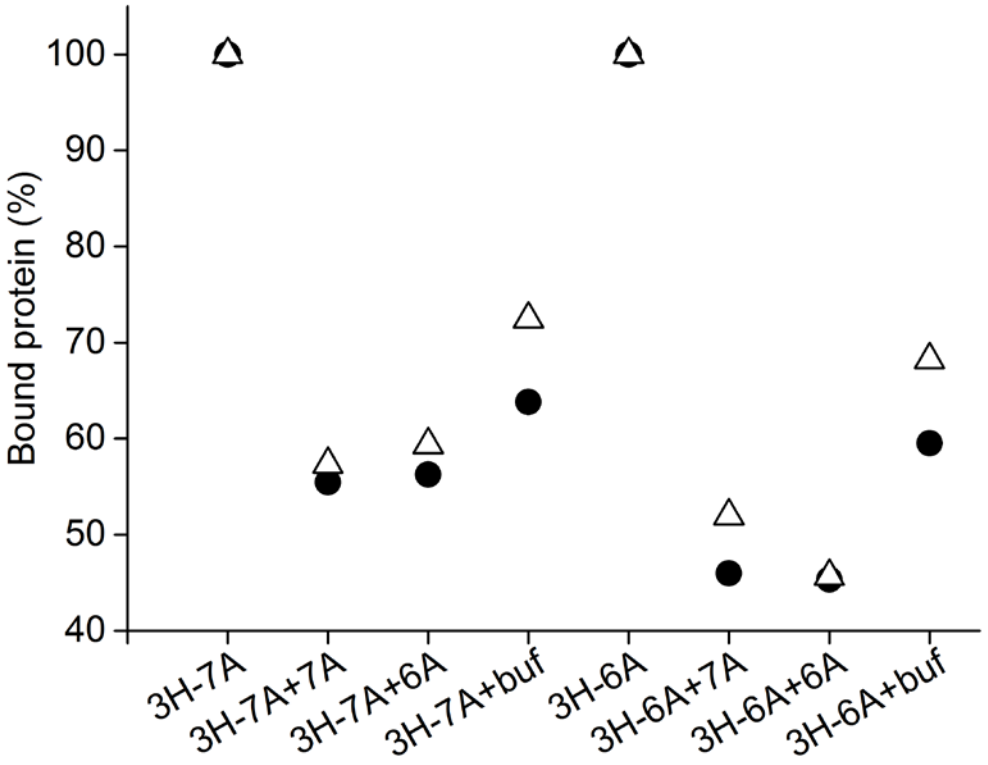


Supplementary figure 4. Binding isotherms of CBM-Cel7A, CBM-Cel6A, DCBM-24, and DCBM-48 on pulp. The partitioning coefficients, K_r, for the proteins are shown in the table within the figure. (+) CBM-CBHI, (x) CBM-CBHII, (○) DCBM-24, (Δ) DCBM-48

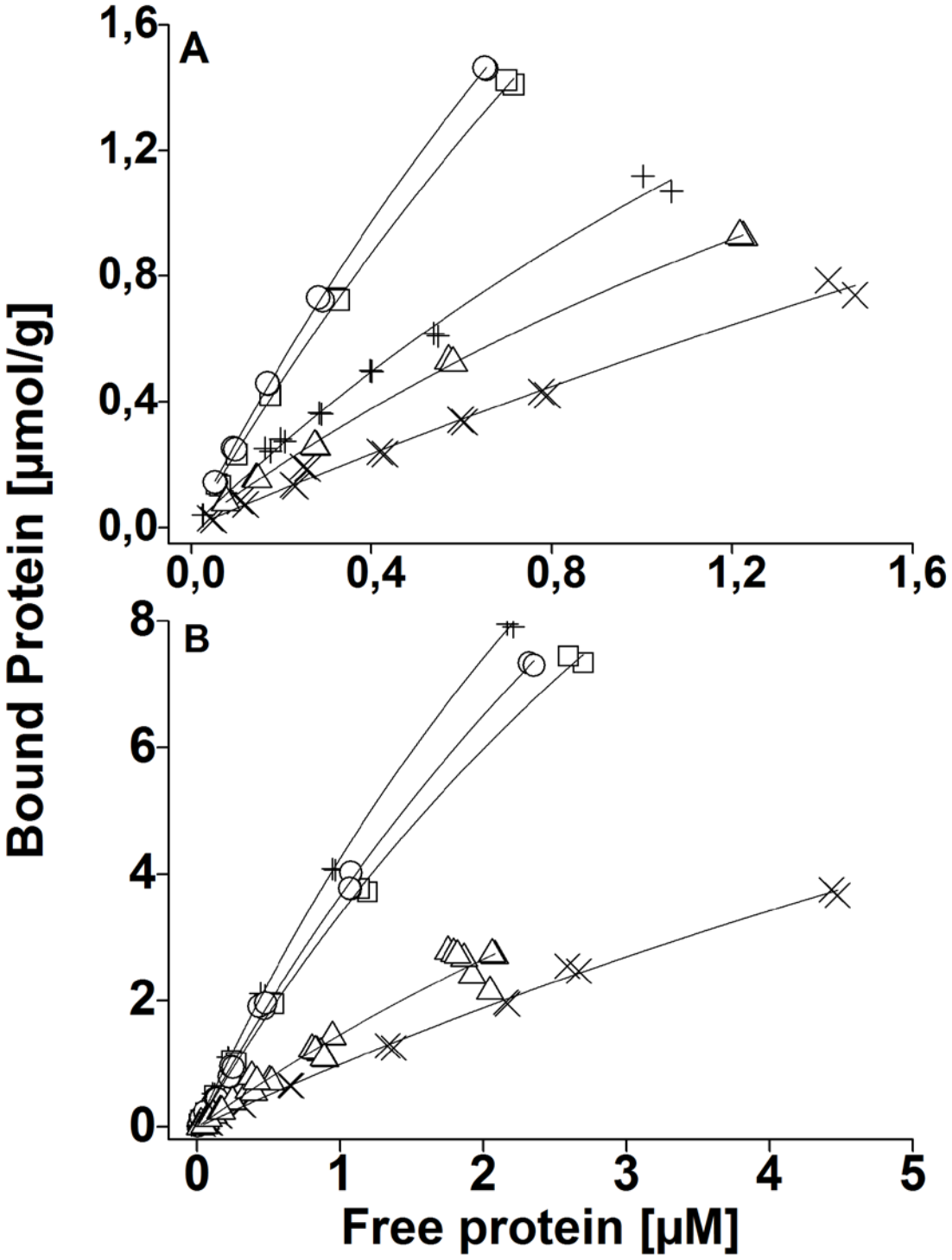
Competition of CBM-Cel7A and CBM-Cel6A on NFC and BMCC

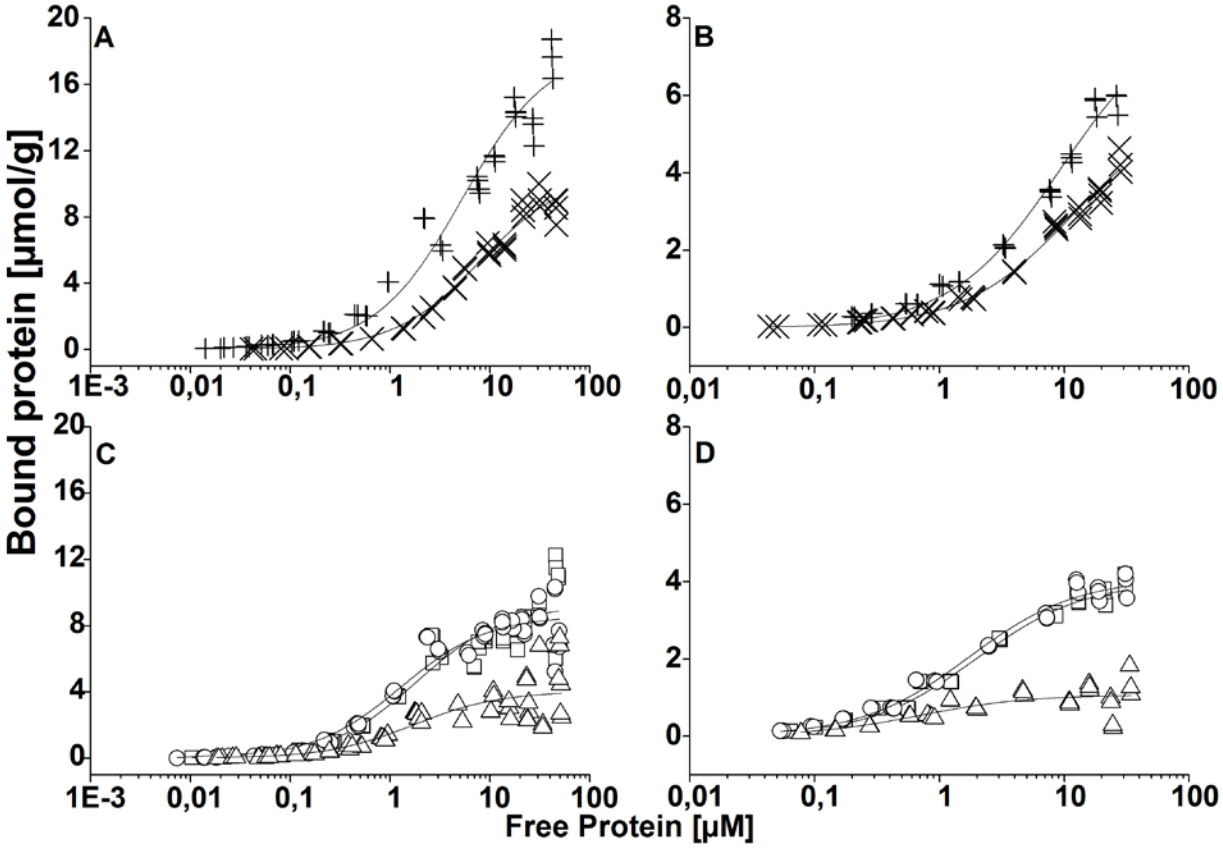
To examine the possible competition of the CBMs on binding sites on BMCC and NFC, we compared how the binding of ^{3}H-labelled CBM-Cel7A was affected by non-labelled CBM-Cel6A and vice versa. A control experiment with buffer only showed the behavior of the protein when the free protein concentration is diluted with no CBM in the solution. This was done in order to see if the added non-labelled CBM affects the binding at all. From the results it was evident that the added non-labelled CBMs affect the binding and they compete with the labelled counterpart in the solution because the results are different form the buffer dilution control. There seems not to be a difference on either substrate whether the competing counterpart is the same or the other CBM because the experiments give very similar results regardless of the components of the experiments. These results suggest that the CBM-Cel7A and CBM-Cel6A fully compete on binding sites on both substrates. This is in agreement with the Gibb's free energies associated with the bindings, also very similar for both proteins on both substrates.

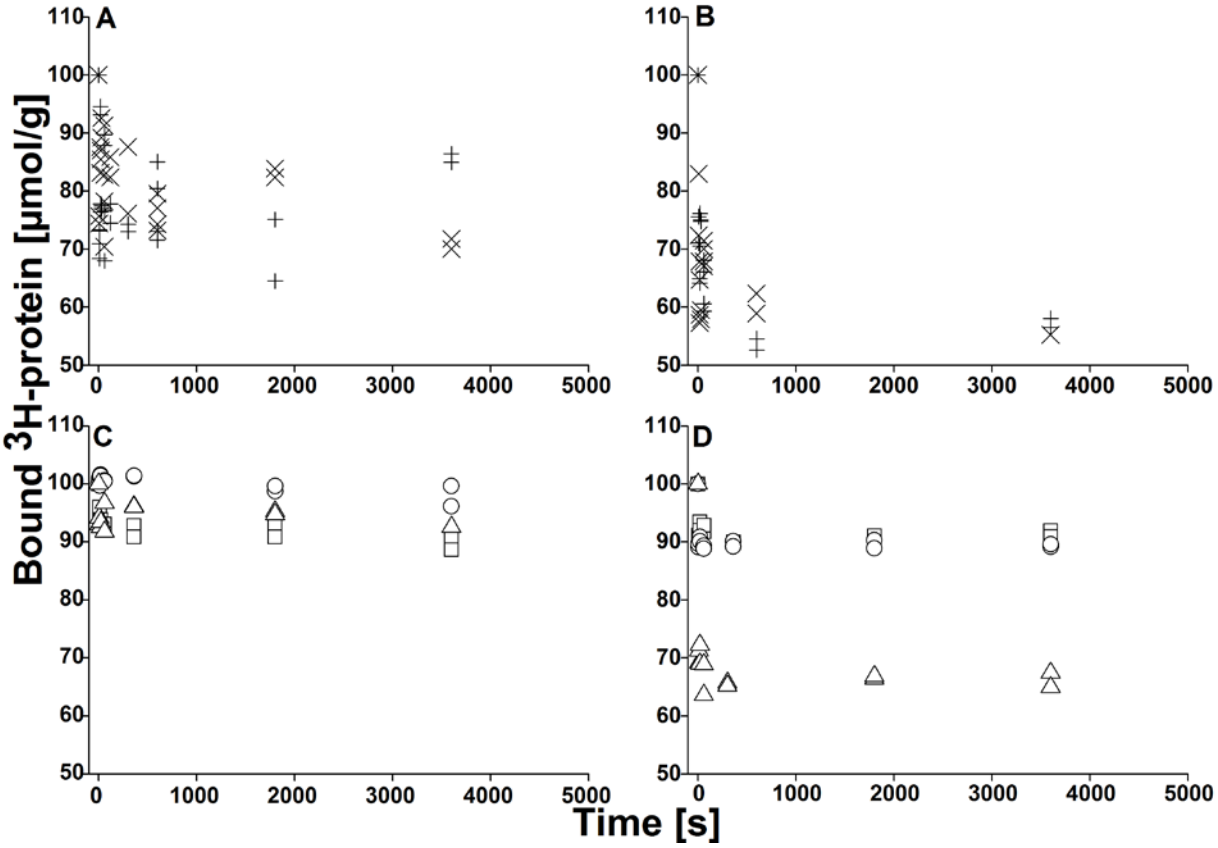
A 25µM solution of CBM containing 10% ^{3}H-protein was diluted 1:1 with a 25 µM solution of the same or the other unlabeled CBM. 100µL of these mixture solutions were let to react with 100 µL of NFC (2 g L-1) and BMCC (1.28 gL-1). As controls the original 25µM solutions with 10% labelled protein and 12.5µM solution with 10% labelled protein (prepared by dilution of 1:1 of the 25µM solution) were used.



Supplementary figure 6. Competition of ^{3}H-labelled CBM-Cel7A and CBM-Cel6A with non-labelled CBM- Cel7A (3H-7A+7A and 3H6A+7A) and CBM-Cel6A (3H7A+6A and 3H7A+6A) on both substrates, NFC and BMCC. Control experiment with ^{3}H- Cel7A and CBM-Cel6A without dilution (3H-7A and 3H-6A), and with buffer dilution (3H-7A+buf and 3H-6A+buf) were made. Δ BMCC, ● NFC.


A
B
1,6
1,2
0,8
0,4
0,0
0,0
0,4
0,8
1,2
1,6
8
6
4
2
0
0
1
2
3
4
5
Bound Protein [µmol/g]
Free protein [µM]


A
B
C
D
Bound protein [µmol/g]
Free Protein [µM]
1E-3
0,01
0,1
1
10
100


A
B
C
D
Bound ^{3}H-protein [µmol/g]
Time [s]
110
100
90
80
70
60
50
0
1000
2000
3000
4000
5000