**Biophysical Journal, Volume 99**

**Supporting Information**

**The *O*-glycosylated linker from the *Trichoderma reesei* Family 7 cellulase is a flexible, disordered protein**

Gregg T. Beckham, Yannick J. Bomble, James F. Matthews, Courtney B. Taylor, Michael G. Resch, John M. Yarbrough, Steve R. Decker, Lintao Bu, Xiongce Zhao, Clare McCabe, Jakob Wohlert, Malin Bergenstråhle, John W. Brady, William S. Adney, Michael E. Himmel, Michael F. Crowley

**Additional Methods**

Ramachandran dihedral angles of each residue were calculated, excluding end residues, and compared between the two linkers to determine if structural flexibility is impaired by the presence of glycosylation. The lifetimes of protein-protein, protein-sugar, and sugar-sugar hydrogen bonds were also characterized with a threshold value of 3.5 Å and an angle cutoff of 120°. All potential donors and acceptors (O-H---O or N-H---O) were included in the analysis and the hydrogen bonds presented have an occupancy greater than 1%.

Contacts were calculated for protein-protein, protein-sugar, and sugar-sugar interactions to measure the compactness of both structures and to quantify the interactions between the protein and glycosylation. Contacts are defined as non-sequential, heavy atom side chain (for protein) or heavy sugar atoms within 6.5 Å. For glycine, the alpha carbon is used to define a contact. For sugar-sugar contacts and protein-sugar contacts, sequential neighbors were excluded as contacts.

Convergence was measured for the REMD simulations using the transit number and the potential energy autocorrelation (57). The transit number is a metric to determine the efficiency of computation by calculating the number of RE attempts required for a 95% probability that one replica will visit the maximum temperature before visiting the minimum temperature. Transit numbers were calculated over a range of average exchange probability for various numbers of replicas. The average exchange probability can be estimated using the following:

$$P(exchange) = \min\left(1, \exp\left[-\frac{\varepsilon^2 N_{df}}{1+\varepsilon}\right]\right) \qquad (1)$$

where 1+ε represents the constant ratio between temperatures when replica temperatures are distributed exponentially, and $N_{df}$ represents the number of degrees of freedom, estimated as $N_{atoms}$ for this simulation. For the glycosylated linker, the average exchange probability equaled 0.088. Using this value in Figure 1a in Abraham and Gready with 16 replicas, the recommended transit number is approximately 1,100 (57). The authors recommend the simulation exceed the transit number by at least two orders of magnitude, which would be 110,000. The glycosylated linker simulation includes 40,000 swap attempts, and based on the error bars presented for Figures 3-5, our results seem to be converged.

The interval between exchanges is important because the transit number metric is a valid model of efficient exchange only when there is a sufficiently long enough interval between exchange attempts to result in independence of replica exchange events. There are various recommendations in literature regarding this number. Many simulations studied in (57) exceed the recommended transit number but exchange too frequently. This can be estimated by the autocorrelation time of the potential energy. The autocorrelation was calculated by

$$c(t) = \frac{1}{c_0(N-t)} \sum_{j=1}^{N-t} \left(U_j - \overline{U}\right)\left(U_{j+t-1} - \overline{U}\right) \qquad (2)$$

where $c_0$ and c(t) are the autocorrelation functions at time 0 and t, respectively, N represents the number of sample points, and U represents the potential energy. C(t) was calculated by analyzing 1.5 ns windows of a 15 ns NVT MD simulation of the glycosylated linker. The time step was 1.5 fs, and data were recorded every 0.15 ps. A suggested exchange interval for REMD was then calculated by the average of the cumulative sums of c(t)*Δt for each window. At long *t,* the noise of the simulation can begin to override *c(t)*, and thus the data were truncated at 400 ps, where *c(t)* shows increased instances of negative values, indicating noise dominance. The average indicates that the interval between swap attempts, or exchange interval, should be 3.5 ps which is on the order of our chosen interval of 3 ps. Based on the above calculations, our simulations with 3 ps/swap and 40,000 swaps per REMD run (for 120*12 ns = 1.4 μs and 120*16 = 1.92 μs) ensured reasonable mixing and thermodynamic efficiency as defined by Abraham and Gready (57).

Error analysis for the free energy curves shown in Figures 3-5 were calculated using a bootstrapping procedure available from Alan Grossfield. The bins were separated for each distance metric into 1 Å bins, and the number of bootstrapping Monte Carlo simulations were varied in each data set until convergence in the relative error was found.

Table S1. Free energy for the non-glycosylated linker REMD simulations.

| End-to-end distance [Å] | Free energy for the non-glycosylated linker [F/kT] |
|---|---|
| 4 | 8.09631 |
| 5 | 5.97605 |
| 6 | 5.01542 |
| 7 | 4.10117 |
| 8 | 3.54243 |
| 9 | 3.04574 |
| 10 | 2.66843 |
| 11 | 2.45954 |
| 12 | 2.24507 |
| 13 | 1.9878 |
| 14 | 1.73905 |
| 15 | 1.69994 |
| 16 | 1.62568 |
| 17 | 1.56673 |
| 18 | 1.42021 |
| 19 | 1.32044 |
| 20 | 1.20683 |
| 21 | 1.05189 |
| 22 | 0.890429 |
| 23 | 0.748084 |
| 24 | 0.725661 |
| 25 | 0.647524 |
| 26 | 0.605536 |
| 27 | 0.493353 |
| 28 | 0.409945 |
| 29 | 0.363356 |
| 30 | 0.317449 |
| 31 | 0.281104 |
| 32 | 0.241887 |
| 33 | 0.145732 |
| 34 | 0.130843 |
| 35 | 0.123079 |
| 36 | 0.0608884 |
| 37 | 0 |
| 38 | 0.0193816 |
| 39 | 0.0638428 |
| 40 | 0.113781 |
| 41 | 0.136765 |
| 42 | 0.184499 |
| 43 | 0.262757 |
| 44 | 0.302818 |
| 45 | 0.309069 |
| 46 | 0.303639 |
| 47 | 0.337695 |
| 48 | 0.391197 |
| 49 | 0.493801 |
| 50 | 0.554036 |
| 51 | 0.671679 |
| 52 | 0.723704 |
| 53 | 0.911093 |
| 54 | 1.10167 |
| 55 | 1.28165 |
| 56 | 1.32718 |
| 57 | 1.50525 |
| 58 | 1.85604 |
| 59 | 2.01587 |
| 60 | 2.15702 |
| 61 | 2.40934 |
| 62 | 2.57797 |
| 63 | 2.60325 |
| 64 | 2.84987 |
| 65 | 3.20429 |
| 66 | 3.46158 |
| 67 | 3.74389 |
| 68 | 4.16667 |
| 69 | 4.15581 |
| 70 | 4.36661 |
| 71 | 5.18155 |
| 72 | 6.10388 |
| 73 | 6.57996 |
| 74 | 6.92624 |
| 75 | 7.20249 |
| 76 | 7.99095 |

Table S2. Free energy for the glycosylated linkers REMD simulations.

| End-to-end distance [Å] | Free energy for the glycosylated linker [F/kT] |
|---|---|
| 5 | 10.3006 |
| 6 | 7.528 |
| 7 | 6.53939 |
| 8 | 6.01013 |
| 9 | 6.01013 |
| 10 | 5.01232 |
| 11 | 4.43129 |
| 12 | 4.68381 |
| 13 | 4.6481 |
| 14 | 4.49244 |
| 15 | 4.05836 |
| 16 | 3.97086 |
| 17 | 4.00348 |
| 18 | 3.7048 |
| 19 | 3.55065 |
| 20 | 3.43261 |
| 21 | 3.17852 |
| 22 | 2.71938 |
| 23 | 2.43963 |
| 24 | 2.20133 |
| 25 | 2.03108 |
| 26 | 1.83806 |
| 27 | 1.66312 |
| 28 | 1.5553 |
| 29 | 1.54123 |
| 30 | 1.54579 |
| 31 | 1.53077 |
| 32 | 1.39013 |
| 33 | 1.24866 |
| 34 | 1.10262 |
| 35 | 1.01699 |
| 36 | 0.864941 |
| 37 | 0.754915 |
| 38 | 0.658461 |
| 39 | 0.573417 |
| 40 | 0.459878 |
| 41 | 0.3796 |
| 42 | 0.355141 |
| 43 | 0.368656 |
| 44 | 0.265325 |
| 45 | 0.174754 |
| 46 | 0.126231 |
| 47 | 0.0861615 |
| 48 | 0.0632369 |
| 49 | 0.0249481 |
| 50 | 0.0165736 |
| 51 | 0.0381997 |
| 52 | 0.0282273 |
| 53 | 0 |
| 54 | 0.0511001 |
| 55 | 0.052514 |
| 56 | 0.0794445 |
| 57 | 0.194076 |
| 58 | 0.283145 |
| 59 | 0.292242 |
| 60 | 0.38147 |
| 61 | 0.472843 |
| 62 | 0.628715 |
| 63 | 0.754699 |
| 64 | 0.857071 |
| 65 | 0.969972 |
| 66 | 1.10658 |
| 67 | 1.28504 |
| 68 | 1.59625 |
| 69 | 1.88565 |
| 70 | 2.00279 |
| 71 | 2.11719 |
| 72 | 2.45321 |
| 73 | 2.90917 |
| 74 | 3.08461 |
| 75 | 3.20056 |
| 76 | 3.59128 |
| 77 | 4.25795 |
| 78 | 4.88005 |
| 79 | 5.64663 |
| 80 | 6.27523 |
| 81 | 6.33029 |
| 82 | 7.41021 |
| 83 | 8.50883 |

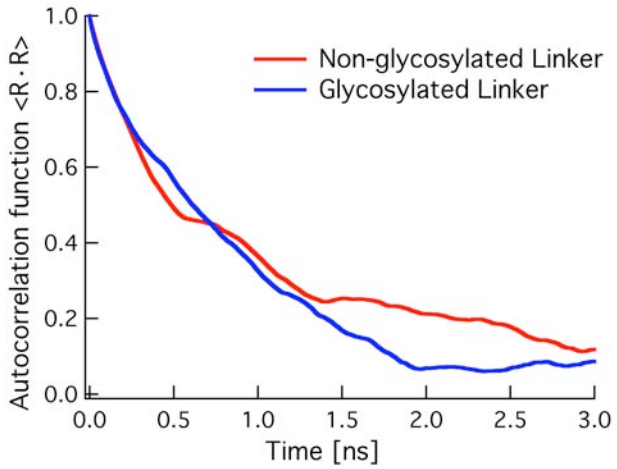



Figure S1. The autocorrelation function of the end-to-end distance (R) of the glycosylated and non-glycosylated Cel7A linker in implicit solvent from MD simulation. The autocorrelation time is 2-3 ns.

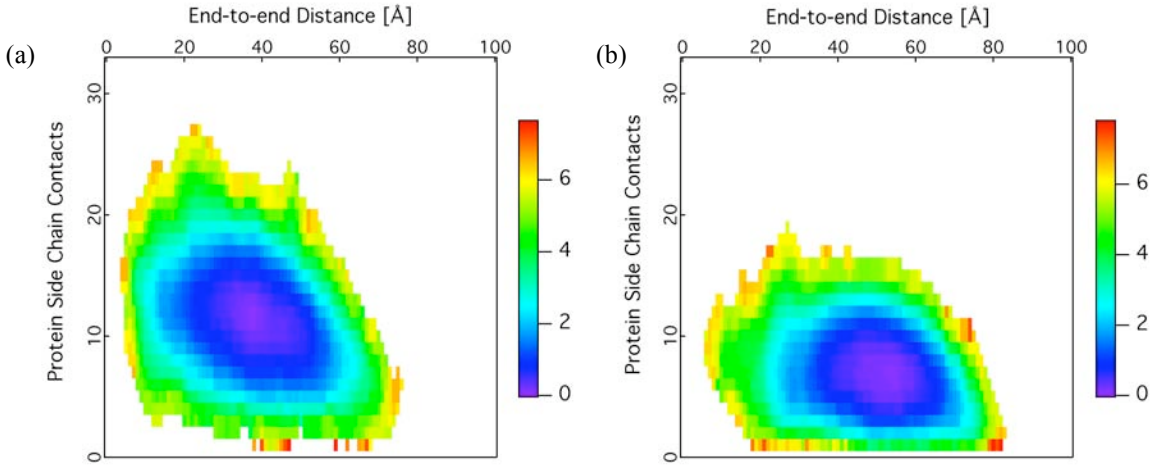
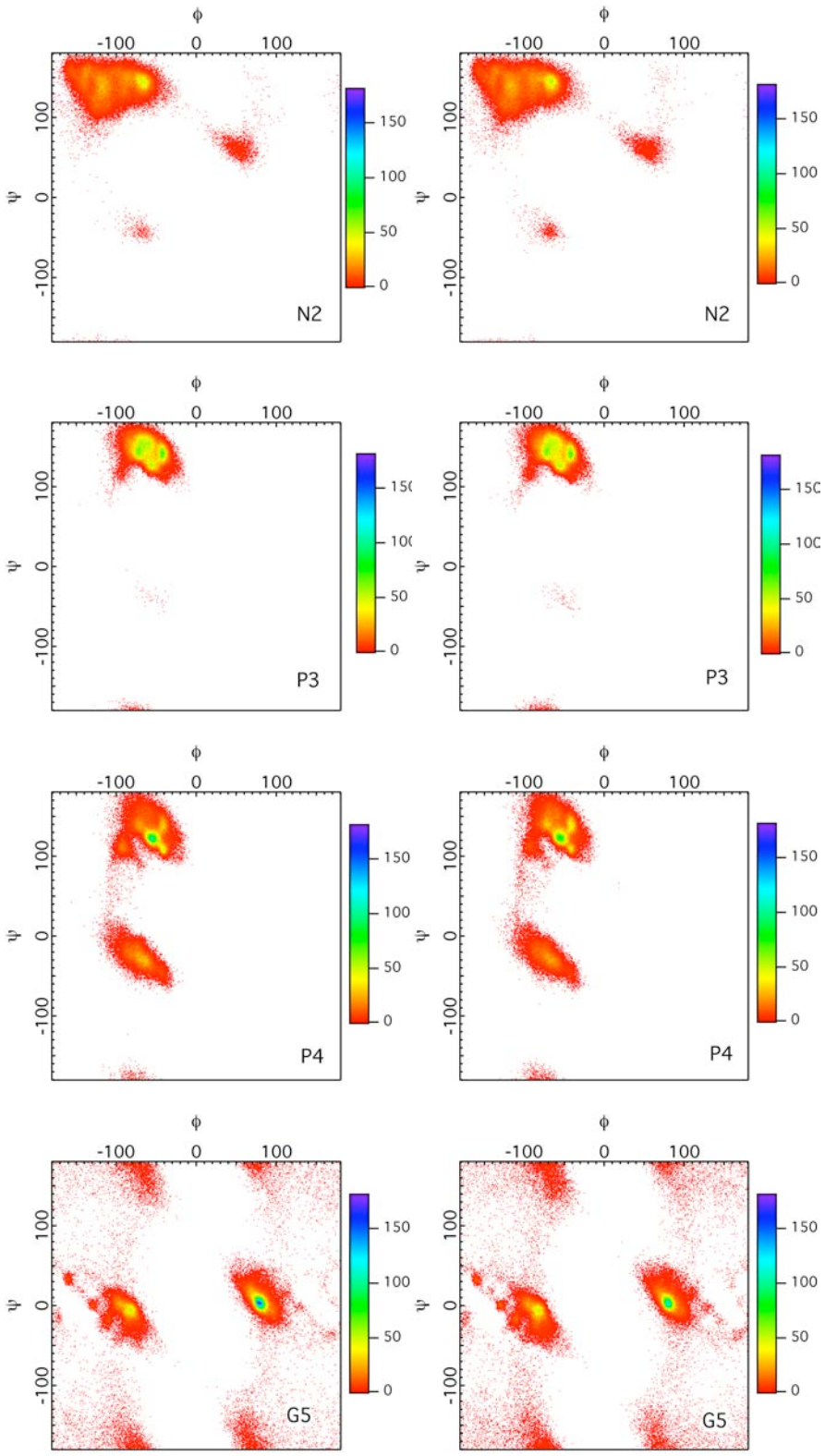
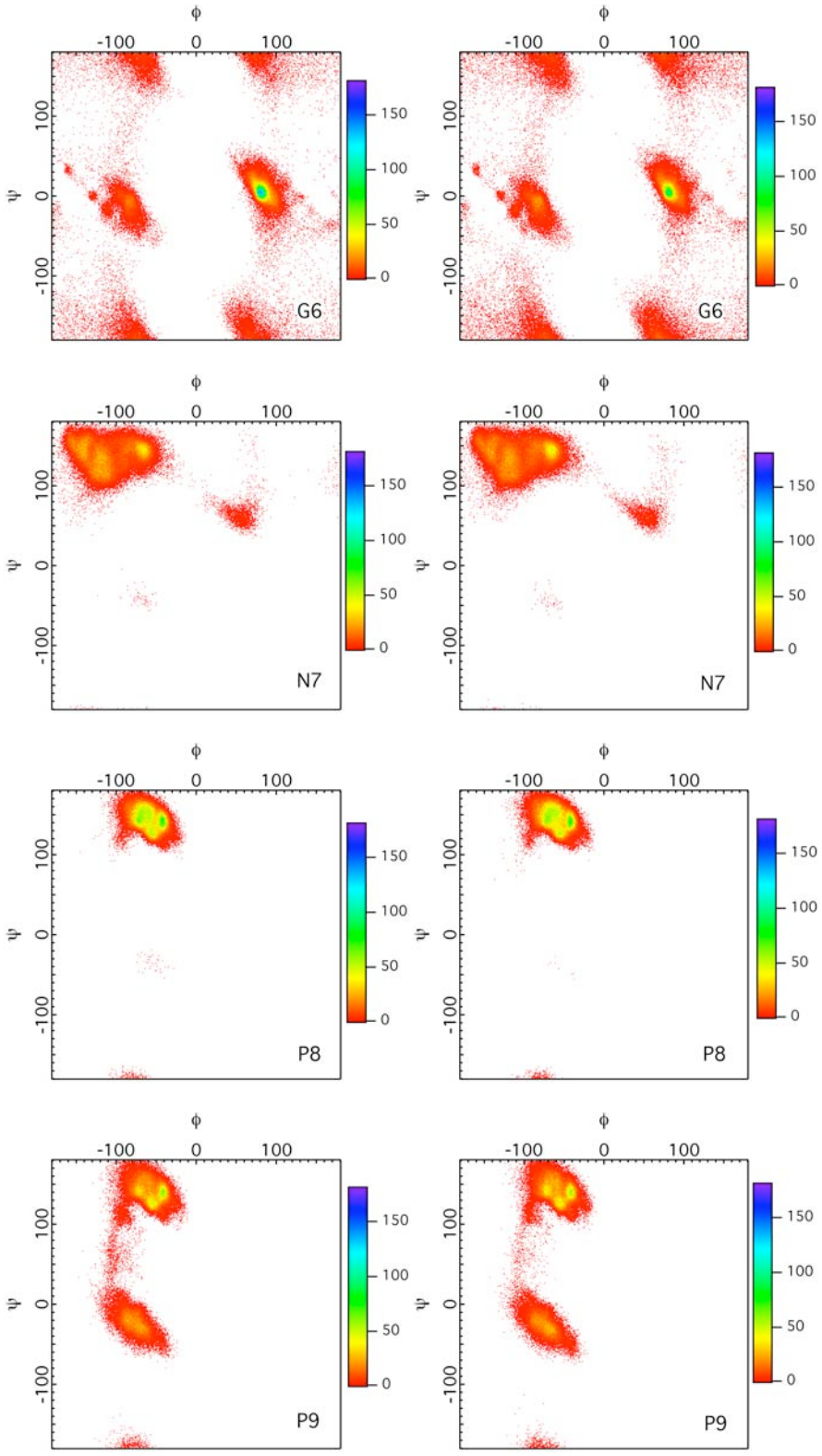
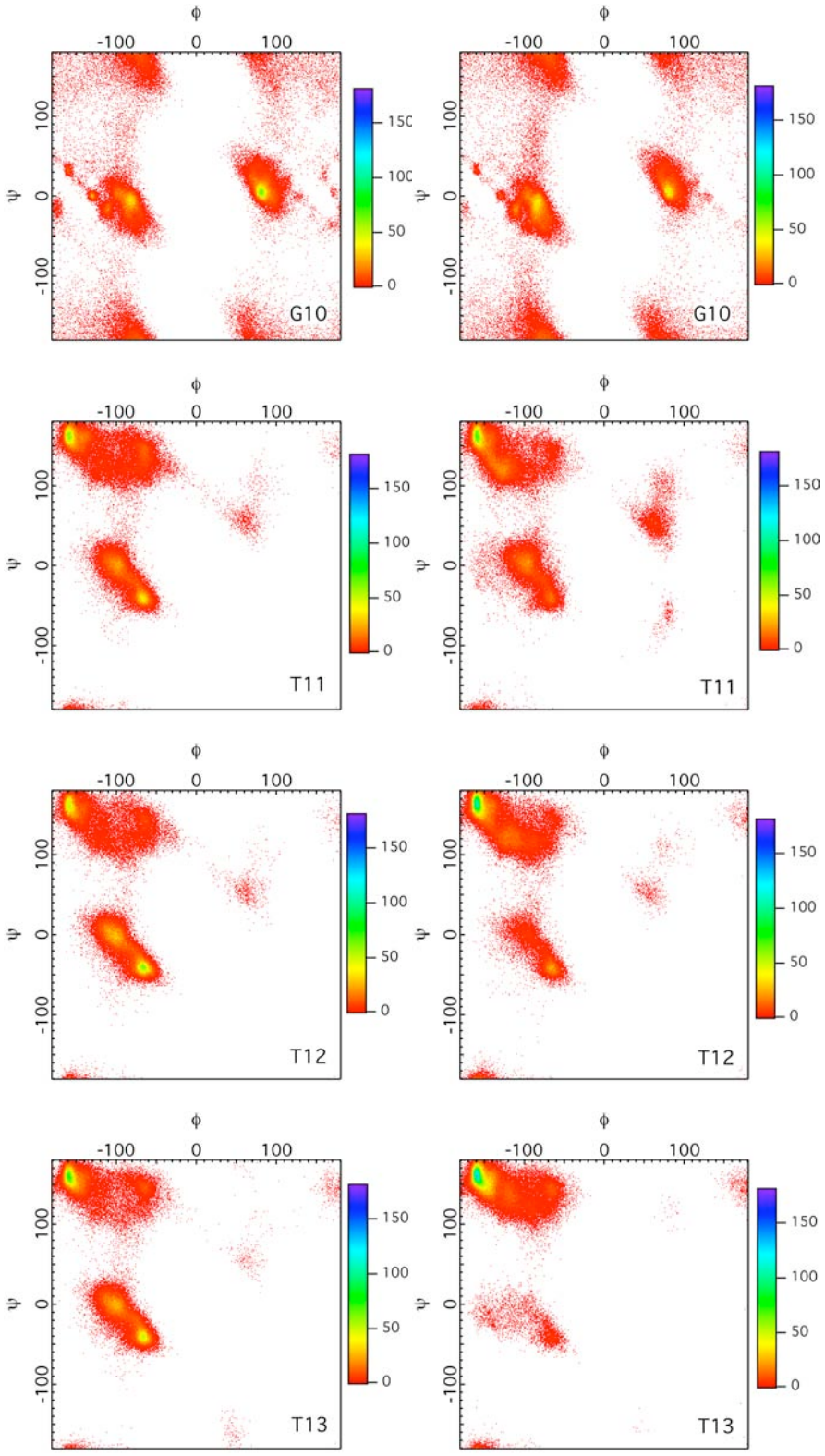
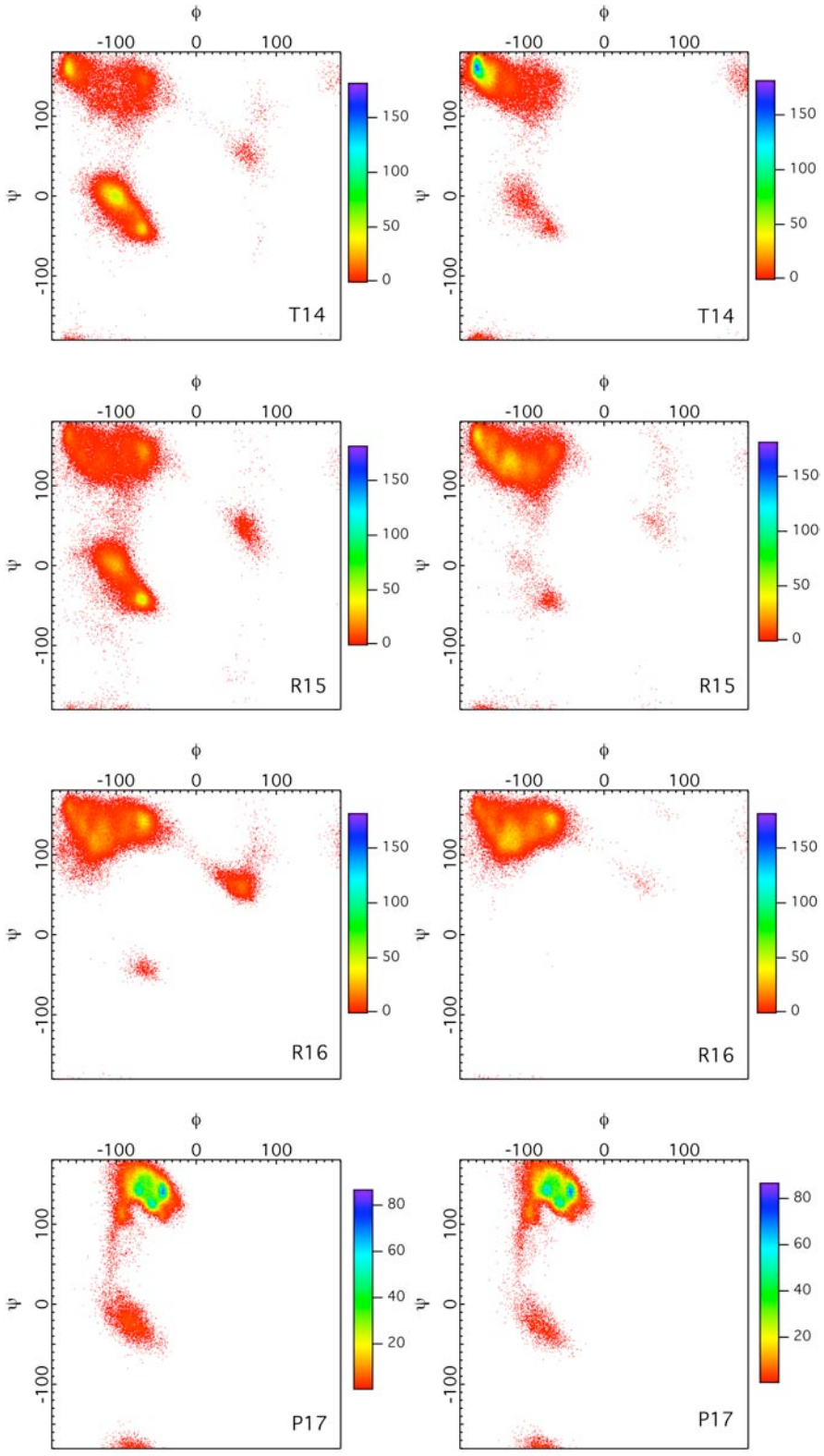


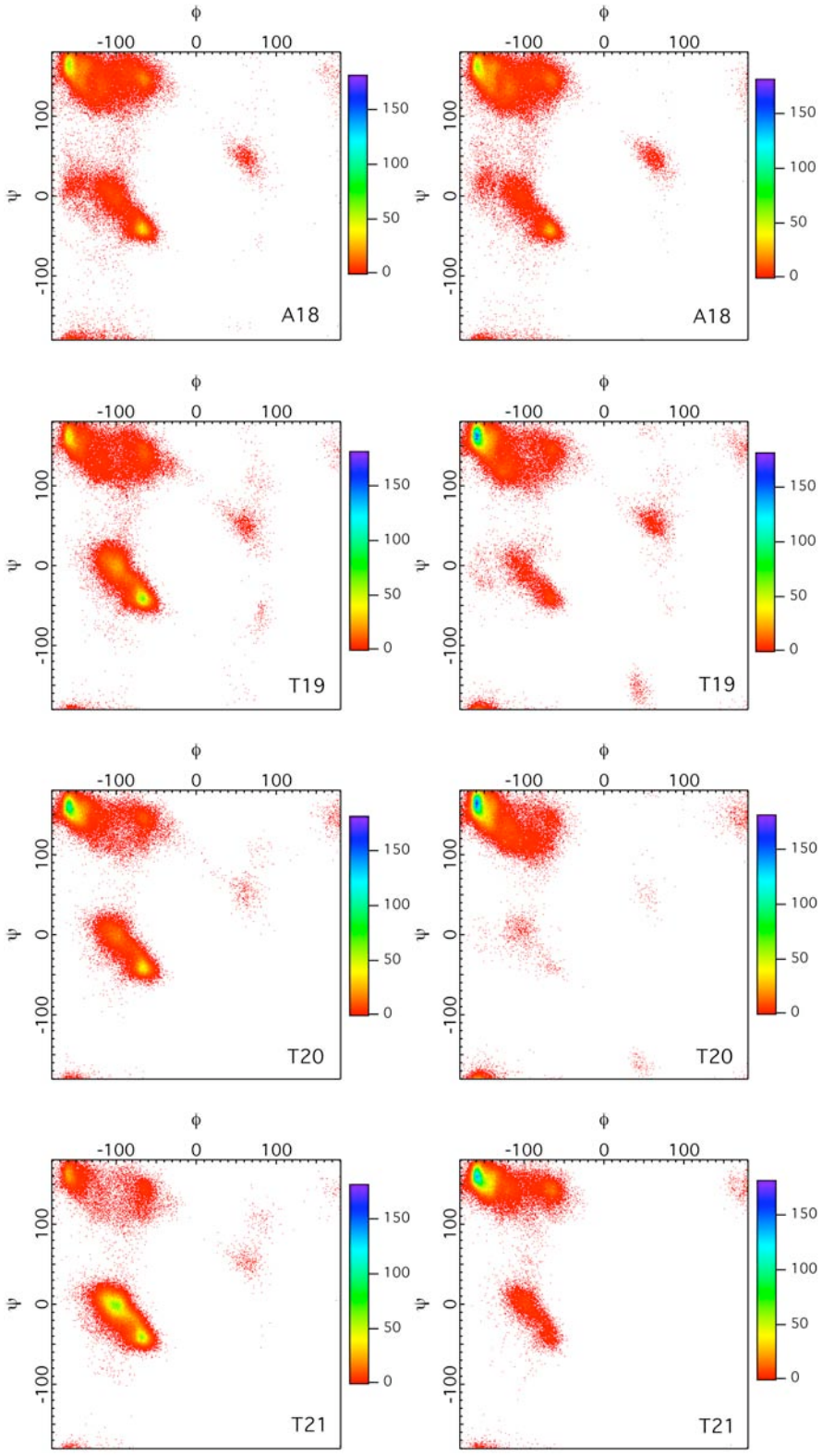

Figure S2. Relative free energy surfaces for the (a) non-glycosylated and (b) glycosylated Cel7A linker as a function of protein side chain contacts and end-to-end distance. The free energy units are dimensionless (F/kT). Based on the differences in the minima and shapes of the free energy contours, the non-glycosylated linker is able to maintain more protein side chain contacts by forming a slightly more compact structure, whereas the glycosylation provides extension and reduces the number of side chain contacts.

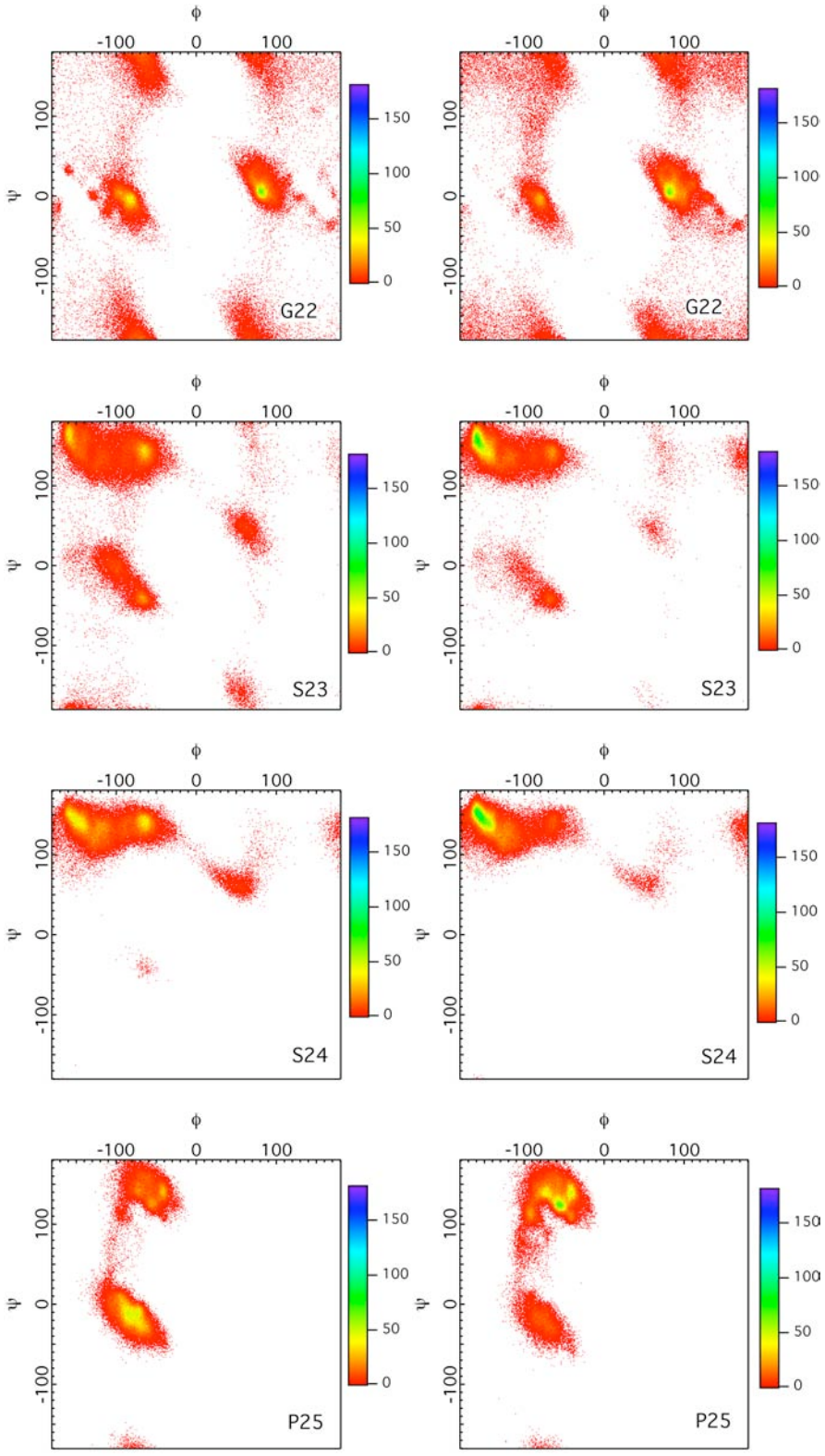

ϕ
ψ
-100
0
100
150
100
50
0
N2
N2
P3
P3
P4
P4
G5
G5


ϕ
ψ
-100
0
100
150
50
G6
N7
P8
P9


ϕ
ψ
-100
0
100
150
100
50
0
G10
G10
T11
T11
T12
T12
T13
T13


ϕ
ψ
-100
0
100
150
100
50
0
T14
T14
R15
R15
R16
R16
80
60
40
20
P17
P17


ϕ
ψ
-100
0
100
150
50
A18
A18
T19
T19
T20
T20
T21
T21


ϕ
ψ
-100
0
100
150
50
G22
S23
S24
P25

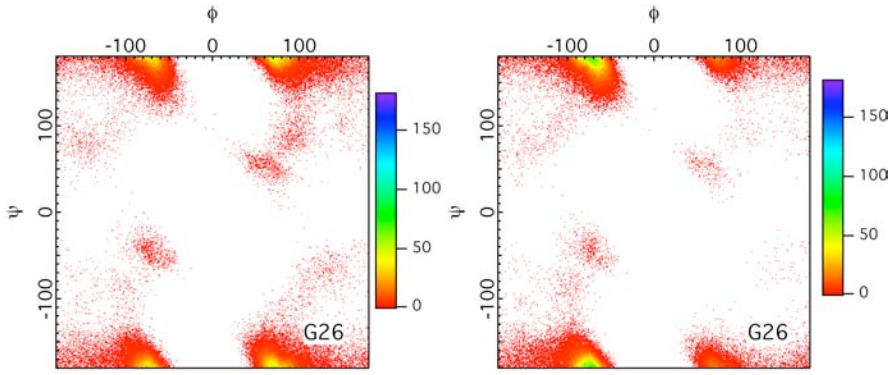



Figure S3. Ramachandran maps for the residues in the *T. reesei* Cel7A non-glycosylated and glycosylated linkers. The left map is from the non-glycosylated linker and the right map is from the glycosylated linker. Each map is on the same scale. The maps are labeled with the residue type and number. These data are taken from the REMD simulations.

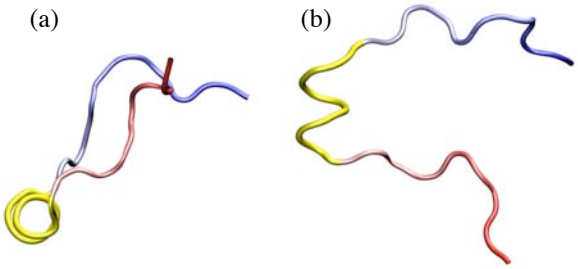



Figure S4. An alpha helix, highlighted in yellow, forms in the REMD simulations of the non-glycosylated Cel7A linker. The linker is shown in backbone shading from the N-terminal in red to the C-terminal in blue. (a) Front view down the helix. (b) Side view of the helix. From the REMD simulations, it is expected that this conformation is relatively thermodynamically equivalent to an extended conformation, and does not form in the glycosylated Cel7A linker REMD simulations.

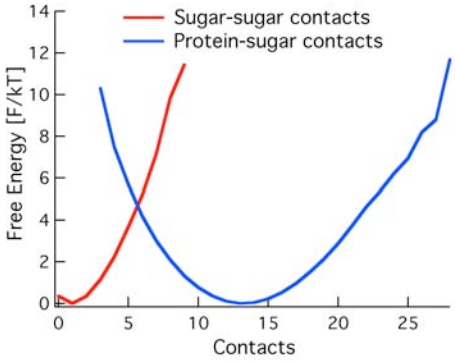



Figure S5. The free energy of the glycosylated Cel7A linker as a function of sugar-sugar contacts and protein-sugar contacts. As shown, there are few glycan-glycan interactions whereas there are a substantial number of non-sequential protein side chain interactions with the glycosylation. This figure implies that there are few sugar-sugar native contacts throughout the linker (non-covalent interactions), with more long-lived protein-sugar (again, non-covalent) interactions.

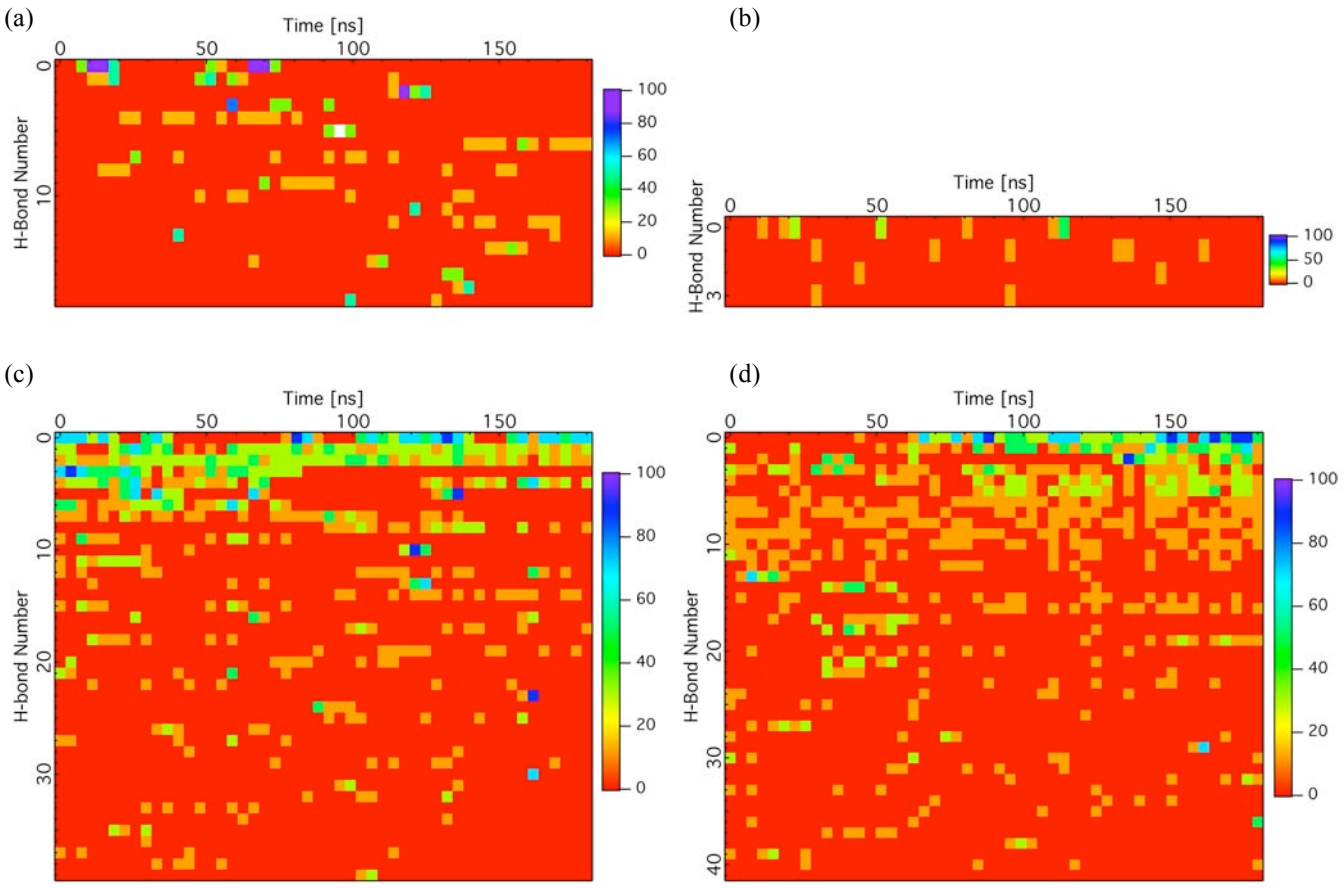



Figure S6. Hydrogen bonding lifetime diagrams as a function of time for a 180 ns of continuous MD simulation. The color contours represent the percentage lifetime of a given hydrogen bond within a window of 1.5 ns. (a) Protein-protein hydrogen bonds for the non-glycosylated linker. (b) Protein-protein hydrogen bonds for the glycosylated linker. (c) Protein-sugar hydrogen bonds for the glycosylated linker. (d) Sugar-sugar hydrogen bonds for the glycosylated linker.

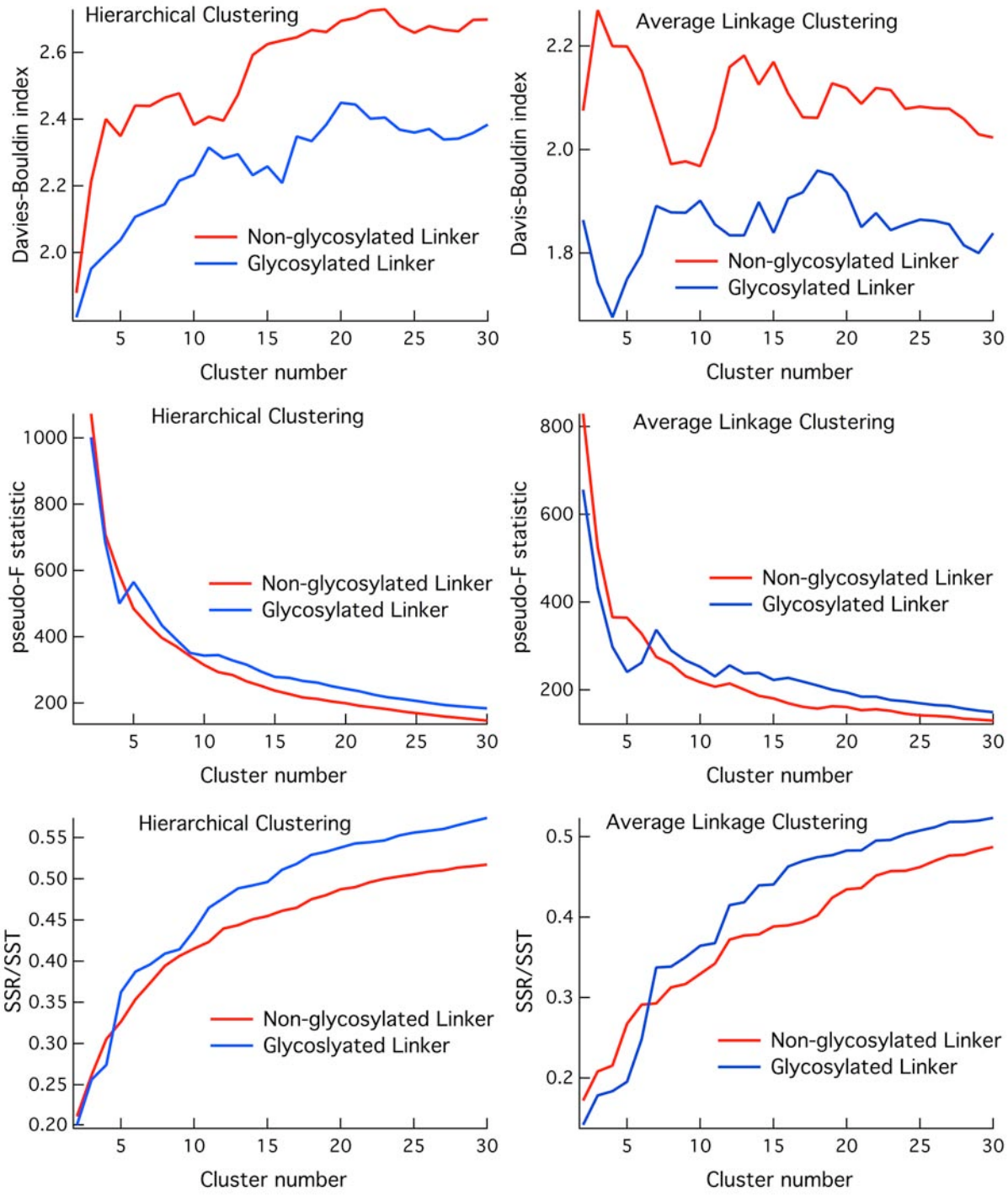



Figure S7. Clustering metric results for the REMD simulations with the hierarchical and average linking algorithms for the non-glycosylated and glycosylated linkers. These metrics are all useful for determining the optimal number of structural clusters found by REMD. For the results shown here, these diagrams illustrate that there are no significantly populated, distinct structural clusters of the *T. reesei* Cel7A linker.

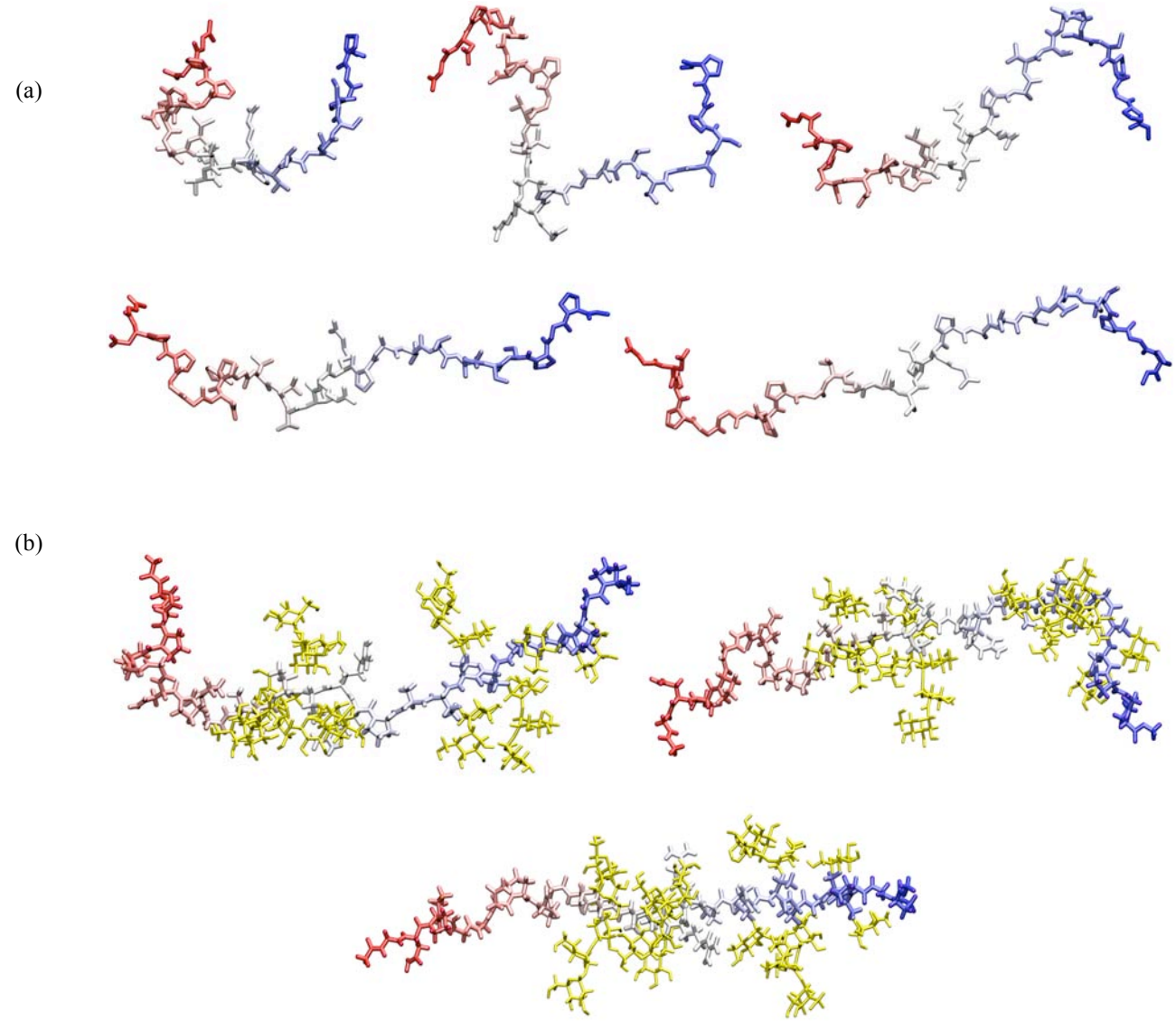



Figure S8. The *T. reesei* Cel7A linker displays significant conformational flexibility. (a) The non-glycosylated linker at different end-to-end distances. From left to right: R=18, 33, 44, 56, and 66 Å. (b) Various conformations of the glycosylated linker. From left to right: R=55, 62, and 69 Å. The linker peptide is colored from red at the N terminus to blue at the C terminus. Sugars are shown in yellow.

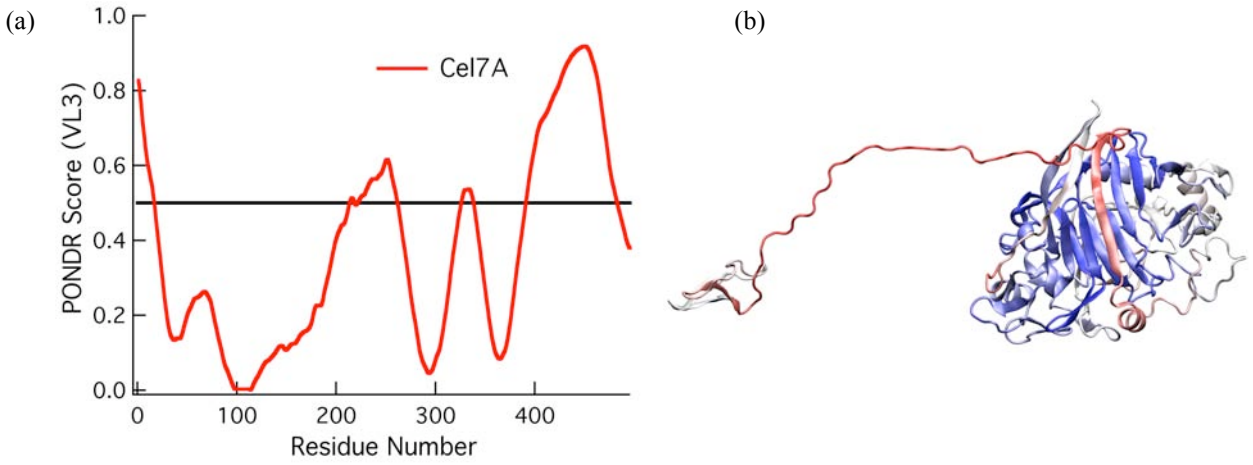



Figure S9. Sequence-based PONDR screen for protein disorder applied to the entire *T. reesei* Cel7A enzyme. The linker is approximately from residues 430 to 460. (a) The VL3 algorithm (58) predicts the Cel7A linker to be a disordered region. (b) The Cel7A enzyme colored by VL3 score from Figure 8(a) where the minimum score is 0 (blue) and the maximum VL3 score is 1.0 (red).

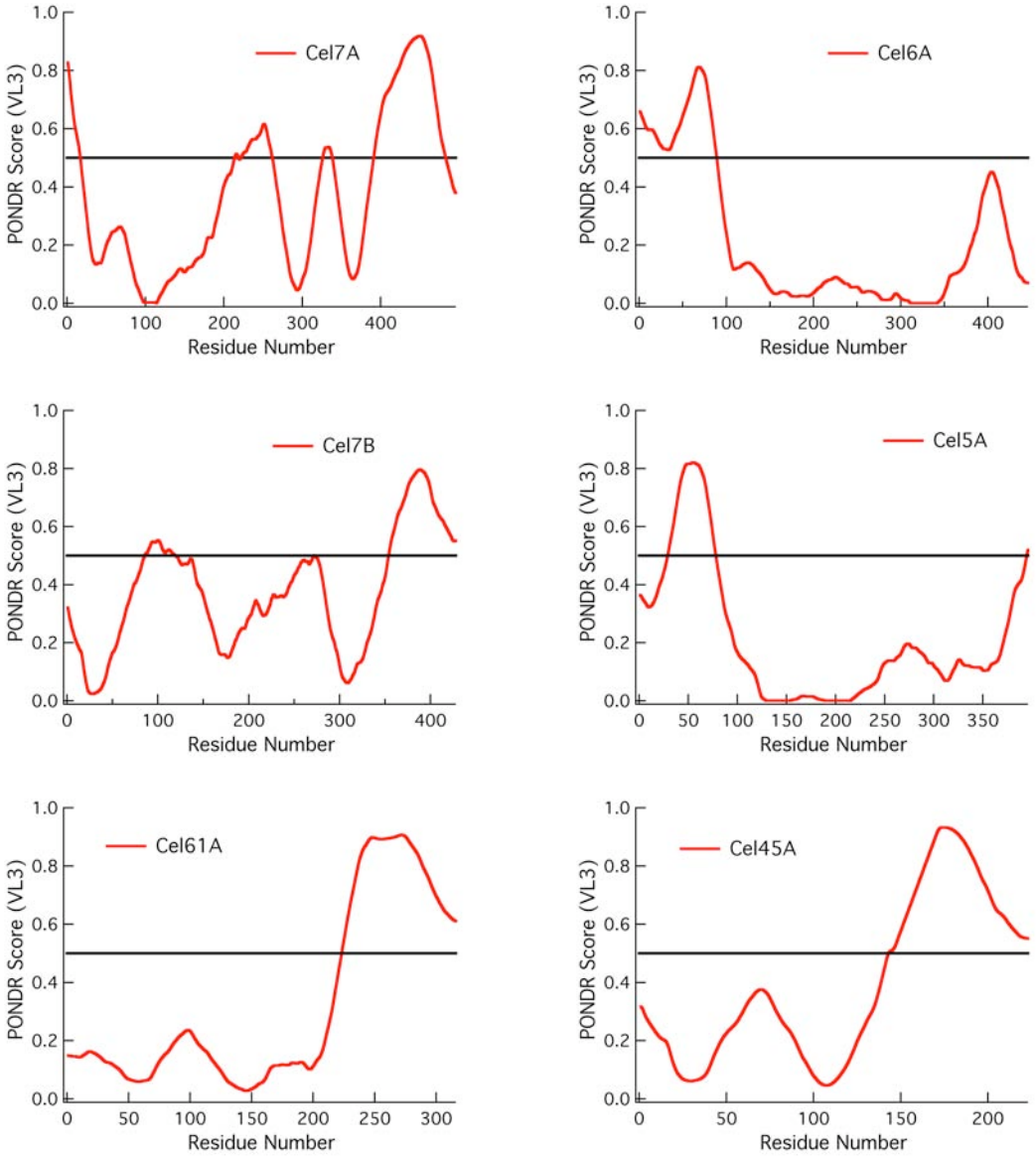



Figure S10. Predicted regions of disorder from the VL3 PONDR algorithm for the *T. reesei* Cel7A cellulase, *T. reesei* Cel6A cellulase, *T. reesei* endoglucanase I or Cel7B, *T. reesei* endoglucanase II or Cel5A, *T. reesei* endoglucanase IV or Cel61A, and *T. reesei* EGV endoglucanase or Cel45A. Signal sequences were removed in all cases. For all of the cellulases shown here, the VL3 PONDR algorithm predicts the linker regions both as the most disordered regions in the protein and as an intrinsically disordered protein as measured by the VL3 score significantly above 0.5.

***T. reesei* Cel6A**

qacssvwgqc ggqnwsgptc casgstcvys ndyysqclpg aasssstra asttsrvspt
tsrsssatpp pgstttrvpp vgsgtatysg npfvgvtpwa nayyasevss laipsltgam
ataaaavakv psfmwldtld ktplmeqtla dirtanknqg nyagqfvvyd lpdrdcaala
sngeysiadg gvakyknyid tirqivveys dirtllviep dslanlvtnl gtpkcanaqs
aylecinyav tqlnlpnvam yldaghagwl gwpanqdpaa qlfanvykna sspralrgla
tnvanyngwn itsppsytqg navyneklyi haigpllanh gwsnaffitd qgrsgkqptg
qqqwgdwcnv igtgfgirps antgdsllds fvwvkpggec dgtsdssapr fdshcalpda
lqpapqagaw fqayfvqllt nanpsfl

***T. reesei* endoglucanase I, Cel7B**

rwmhdanyns ctvnggvntt lcpdeatcgk ncfiegvdya asgvttsgss ltmnqympss
sggyssvspr lylldsdgey vmlklngqel sfdvdlsalp cgengslyls qmdengganq
yntaganygs gycdaqcpvq twrngtlnts hqgfccnemd ilegnsrana ltphsctata
cdsagcgfnp ygsgyksyyg pgdtvdtskt ftiitqfntd ngspsgnlvs itrkyqqngv
dipsaqpggd tisscpsasa ygglatmgka lssgmvlvfs iwndnsqymn wldsgnagpc
sstegnpsni lannpnthvv fsnirwgdig sttnstappp ppassttfst trrssttsss
psctqthwgq cggigysgck tctsgttcqy sndyysqcl

***T. reesei* endoglucanase II, Cel5A**

qqtvwgqcg gigwsgptnc apgsacstln pyyaqcipgat tittstrpps gpttttrats
tssstppts sgvrfagvni agfdfgcttd gtcvtskvypp lknftgsnny pdgigqmqhf
vnedgmtifr lpvgwqylv nnnlggnlds tsiskydqlvq gclslgayci vdihnyarwn
ggiigqggpt naqftslws qlaskyasqs rvwfgimneph dvnintwaat vqevvtairn
agatsqfisl pgndwqsag afisdgsaaa lsqvtnpdgst tnlifdvhky ldsdnsgtha
ecttnnidga fsplatwlr qnnrqailte tgggnvqsciq dmcqqiqyln qnsdvylgyv
gwgagsfdst yvltetpts sgnswtdtsl vssclark

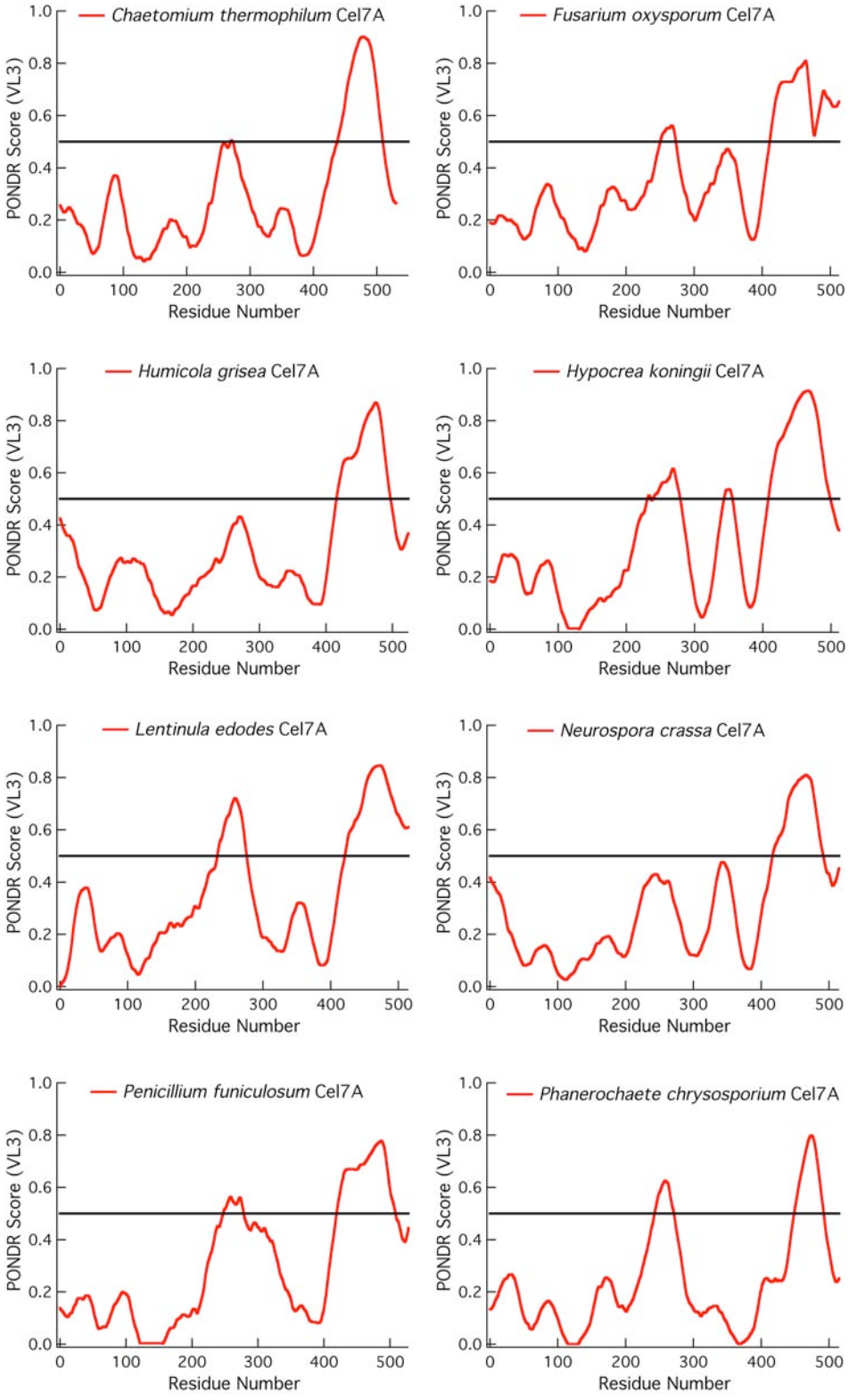
***T. reesei* endoglucanase IV, Cel61A**

dngfvspday qnpdiichkn atnakghasv kagdtilfqw vpvpwphpgp ivdylancng
dcetvdkttl effkidgvgl lsggdpgtwa sdvlisnnnt wvvkipdnla pgnyvlrhei
ialhsagqan gaqnypqcfn iavsgsgslq psgvlgtdly hatdpgvlin iytsplnyii
pgptvvsglp tsvaqgssaa tatasatvpg ggsgptsrtt ttarttqass rpsstppatt
sapaggptqt lygqcggsgy sgptrcappa tcstlnpyya qcln

***T. reesei* endoglucanase V, Cel45A**

aykatttryy dgqegacgcg sssgafpwql gignqvytaa gsqalfdtag aswcgagcgk
cyqltstgqa pcsscgtgga agqsiivmvt nlcpnngnaq wcpvvggtnq ygysyhfdim
aqneifgdnv vvdfepiacp gqaasdwgtc lcvgqqetdp tpvlgndtgs tppgssppat
sssppsgggq qtlygqcgga gwtgpttcqa pgtckvqnqw ysqclp

Figure S11. Sequences of the *T. reesei* cellulases examined in the charge-hydropathy scale. The linkers screened in this algorithm are highlighted in yellow. Signal peptides are not shown for all sequences. The boundaries between the linker and the CBM as well as the linker and the catalytic domain are taken from the gene annotations. The sequences of the Family 1 CBMs are highlighted in light blue.


Chaetomium thermophilum Cel7A
Fusarium oxysporum Cel7A
Humicola grisea Cel7A
Hypocrea koningii Cel7A
Lentinula edodes Cel7A
Neurospora crassa Cel7A
Penicillium funiculosum Cel7A
Phanerochaete chrysosporium Cel7A
PONDR Score (VL3)
Residue Number
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
400
500

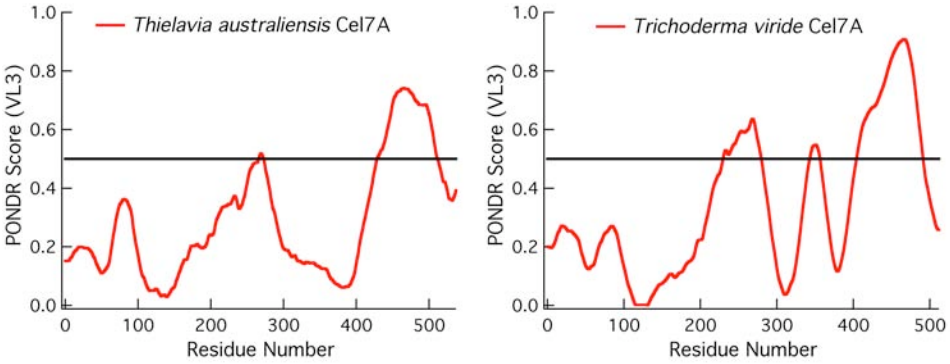



Figure S12. Predicted regions of disorder from the VL3 PONDR algorithm for multiple Family 7 cellobiohydrolases from different fungal species. In all cases, the linker regions are predicted to be disordered.

***Chaetomium thermophilum*** **Cel7A**

mmykkfaala alvagasaqq acsltaenhp sltwkrctsg gscstvngav tidanwrwth
tvsgstncyt gnqwdtslct dgkscaqtcc vdgadyssty gittsgdsln lkfvtkhqyg
tnvgsrvylm endtkyqmfe llgneftfdv dvsnlgcgln galyfvsmda dggmskysgn
kagakygtgy cdaqcprdlk fingeanvgn wtpstndana gfgrygsccs emdvweannm
ataftphpct tvgqsrcead tcggtyssdr yagvcdpdgc dfnayrqgdk tfygkgmtvd
tnkkmtvvtq fhknsagvls eikrfyvqdg kiianaeski pgnpgnsitq eycdaqkvaf
sntddfnrkg gmaqmskala gpmvlvmsvw ddhyanmlwl dstypidqag apgaergacp
ttsgvpaeie aqvpnsnvif snirfgpigs tvpgldgsnp gnptttvvpp aststsrpts
stsspvstpt gqpggcttqk wgqcggigyt gctncvagtt ctqlnpwysq cl

***Chrysosporium lucknowense*** **Cel7A**

myakfatlaa lvagaaaqna ctltaenhps ltwskctsgg sctsvqgsit idanwrwthr
tdsatncyeg nkwdtsycsd gpscaskcci dgadysstyg ittsgnslnl kfvtkgqyst
nigsrtylme sdtkyqmfql lgneftfdvd vsnlgcglng alyfvsmdad ggmskysgnk
agakygtgyc dsqcprdlkf ingeanvenw qsstndanag tgkygsccse mdvweannma
aaftphpcxv igqsrcegds cggtystdry agicdpdgcd fnsyrqgnkt fygkgmtvdt
tkkitvvtqf lknsagelse ikrfyvqngk vipnsestip gvegnsitqd wcdrqkaafg
dvtdxqdkgg mvqmgkalag pmvlvmsiwd dhavnmlwld stwpidgagk pgaergacpt
tsgvpaevea eapnsnvifs nirfgpigst vsglpdggsg npnppvssst pvpsssttss
gssgptggtg vakhyeqcgg igftgptqce spytctklnd wysqcl

***Fusarium oxysporum*** **Cel7A**

myrivatasa liaaaraqqv cslntetkpa ltwskctssg csdvkgsvvi danwrwthqt
sgstncytgn kwdtsictdg ktcaekccld gadysgtygi tssgnqlslg fvtngpyskn
igsrtylmen entyqmfqll gneftfdvdv sgigcglnga phfvsmdedg gkakysgnka
gakygtgycd aqcprdvkfi ngvansegwk psdsdvnagv gnlgtccpem diweansist
aftphpctkl tqhsctgdsc ggtyssdryg gtcdadgcdf nayrqgnktf ygpgsnfnid
ttkkmtvvtq fhkgsngrls eitrlyvqng kvianseski agnpgsslts dfcskqksvf
gdiddfskkg gwngmsdals apmvlvmslw hdhhsnmlwl dstyptdstk vgsqrgscat
tsgkpsdler dvpnskvsfs nikfgpigst yksdgttpnp passsttgss tptnppagsv
dqwgqcggqn ysgpttcksp ftckkindfy sqcq

***Humicola grisea*** **Cel7A**

mrtakfatla alvasaaaqq acsltterhp slswnkctag gqcqtvqasi tldsnwrwth
qvsgstncyt gnkwdtsict dakscaqncc vdgadytsty gittngdsls lkfvtkgqhs
tnvgsrtylm dgedkyqtfe llgneftfdv dvsnigcgln galyfvsmda dgglsrypgn
kagakygtgy cdaqcprdik fingeanieg wtgstndpna gagrygtccs emdiweannm
ataftphpct iigqsrcegd scggtysner yagvcdpdgc dfnsyrqgnk tfygkgmtvd
ttkkitvvtq flkdangdlg eikrfyvqdg kiipnsesti pgvegnsitq dwcdrqkvaf
gdiddfnrkg gmkqmgkala gpmvlvmsiw ddhasnmlwl dstfpvdaag kpgaergacp
ttsgvpaeve aeapnsnvvf snirfgpigs tvaglpgagn ggnnggnppp pttttssapa
ttttasagpk agrwqqcggi gftgptqcee pyictklndw ysqcl

***Hypocrea koningii*** **Cel7A**

myrklavisa flataraqsa ctlqsethpp ltwqkcssgg tctqqtgsvv idanwrwtha
tnsstncydg ntwsstlcpd netcaknccl dgaayastyg vttsgnslsi gfvtqsaqkn
vgarlylmas dttyqeftll gnefsfdvdv sqlpcglnga lyfvsmdadg gvskyptnta
gakygtgycd sqcprdlkfi ngqanvegwe pssnnantgi gghgsccsem diweansise
altphpcttv gqeicegdgc ggtysdnryg gtcdpdgcdw npyrlgntsf ygpgssftld
ttkkltvvtq fetsgainry yvqngvtfqq pnaelgsysg nelnddycta eeaefggssf
sdkggltqfk katsggmvlv mslwddyyan mlwldstypt netsstpgav rgscstssgv
paqvesqspn akvtfsnikf gpigstgnps ggnppggnrg ttttrrpatt tgsspgptqs
hygqcggigy sgptvcasgt tcqvlnpyys qcl

***Lentinula edodes*** **Cel7A**

mfrtaallsf aylavvygqq agtstaethp pltweqctsg gscttqsssv vldsnwrwth
vvggytncyt gnewnttvcp dgttcaanca ldgadyegty gistsgnalt lkfvtasaqt
nvgsrvylma pgseteyqmf nplnqeftfd vdvsalpcgl ngalyfsemd adgglseypt
nkagakygtg ycdsqcprdi kfiegkanve gwtpsstspn agtggtgicc nemdiweans
isealtphpc taqggtactg dscsspnsta gicdqagcdf nsfrmgdtsf ygpgltvdtt
skitvvtqfi tsdntttgdl tairriyvqn gqviqnsmsn iagvtptnei ttdfcdqqkt
afgdtntfse kggltgmgaa fsrgmvlvls iwdddaaeml wldstypvgk tgpgaargtc
attsgqpdqv etqspnaqvv fsnikfgaig stfsstgtgt gtgtgtgtgt gtttssapaa
tqtkygqcgg qgwtgatvca sgstctssgp yysqcl

***Neurospora crassa* Cel7A**

mrasllafsl aaavaggqqa gtltakrhps ltwqkctrgg cptlnttmvl danwrwthat
sgstkcytgn kwqatlcpdg kscaancald gadytgtygi tgsgwsltlq fvtdnvgara
ylmaddtqyq mlellnqelw fdvdmsnipc glngalylsa mdadggmrky ptnkagakya
tgycdaqcpr dlkyingian vegwtpstnd angigdhgsc csemdiwean kvstaftphp
cttieqhmce gdscggtysd drygvlcdad gcdfnsyrmg nttfygegkt vdtsskftvv
tqfikdsagd laeikafyvq ngkviensqs nvdgvsgnsi tqsfcksqkt afgdiddfnk
kgglkqmgka laqamvlvms iwddhaanml wldstypvpk vpgayrgsgp ttsgvpaevd
anapnskvaf snikfghlgi spfsggssgt ppsnpsssas ptsstakpss tstasnpsgt
gaahwaqcgg igfsgpttcp epytcakdhd iysqcv

***Penicillium funiculosum* Cel7A**

msalnsfnmy ksalilgsll atagaqqigt ytaethpsls wstcksggsc ttnsgaitld
anwrwvhgvn tstncytgnt wntaicdtda scaqdcaldg adysgtygit tsgnslrlnf
vtgsnvgsrt ylmadnthyq ifdllnqeft ftvdvsnlpc glngalyfvt mdadggvsky
pnnkagaqyg vgycdsqcpr dlkfiagqan vegwtpstnn sntgignhgs ccaeldiwea
nsisealtph pcdtpgltvc taddcggtys snryagtcdp dgcdfnpyrl gvtdfygsgk
tvdttkpftv vtqfvtddgt ssgslseirr yyvqngvvip qpsskisgis gnvinsdfca
aelsafgeta sftnhgglkn mgsaleagmv lvmslwddys vnmlwldsty panetgtpga
argscpttsg npktvesqsg ssyvvfsdik vgpfnstfsg gtstggsttt tasgttstka
sttstsstst gtgvaahwgq cggqgwtgpt tcasgttctv vnpyysqcl

***Phanerochaete chrysosporium* Cel7A**

mfrtatllaf tmaamvfgqq vgtntaenhr tltsqkctks ggcsnlntki vldanwrwlh
stsgytncyt gnqwdatlcp dgktcaanca ldgadytgty gitasgsslk lqfvtgsnvg
srvylmaddt hyqmfqllnq eftfdvdmsn lpcglngaly lsamdadggm akyptnkaga
kygtgycdsq cprdikfing eanvegwnat sanagtgnyg tcctemdiwe anndaaaytp
hpcttnaqtr csgsdctrdt glcdadgcdf nsfrmgdqtf lgkgltvdts kpftvvtqfi
tndgtsagtl teirrlyvqn gkviqnssvk ipgidpvnsi tdnfcsqqkt afgdtnyfaq
hgglkqvgea lrtgmvlals iwddyaanml wldsnyptnk dpstpgvarg tcattsgvpa
qieaqspnay vvfsnikfgd lnttytgtvs sssvssshss tstssshsss stpptqptgv
tvpqwgqcgg igytgsttca spytchvlnp yysqcy

***Thielavia australiensis* Cel7A**

myakfatlaa lvagasaqav csltaethps ltwqkctapg sctnvagsit idanwrwthq
tssatncysg skwdssictt gtdcaskcci dgaeysstyg ittsgnalnl kfvtkgqyst
nigsrtylme sdtkyqmfkl lgneftfdvd vsnlgcglng alyfvsmdad ggmskysgnk
agakygtgyc daqcprdlkf ingeanvegw esstndanag sgkygsccte mdvweannma
taftphpctt igqtrcegdt cggtyssdry agvcdpdgcd fnsyrqgnkt fygkgmtvdt
tkkitvvtqf lknsagelse ikrfyaqdgk vipnsestia gipgnsitka ycdaqktvfq
ntddftakgg lvqmgkalag dmvlvmsvwd dhavnmlwld styptdqvgv agaergacpt
tsgvpsdvea napnsnvifs nirfgpigst vqglpssggt sssssaapqs tstkasttts
avrttstatt kttssapaqg tntakhwqqc ggngwtgptv cespykctkq ndwysqcl

***Trichoderma viride* Cel7A**

myqklalisa flataraqsa ctlqaethpp ltwqkcssgg tctqqtgsvv idanwrwtha
tnsstncydg ntwsstlcpd netcaknccl dgaayastyg vttsadslsi gfvtqsaqkn

vgarlylmas dttyqeftll gnefsfdvdv sqlpcglnga lyfvsmdadg gvtkyptnta
gakygtgycd sqcprdlkfi ngqanvegwe pssnnantgi gghgsccsem diweansise
altphpcttv gqeicegdsc ggtysgdryg gtcdpdgcdw npyrlgntsf ygpgssftld
ttkkltvvtq fetsgainry yvqngvtfqq pnaelgdysg nsldddycaa eeaefggssf
sdkggltqfk katsggmvlv mslwddyyan mlwldstypt detsstpgav rgssstssgv
paqlesnspn akvvysnikf gpigstgnps ggnppggnpp gtttprpats tgsspgptqt
hygqcggigy igptvcasgs tcqvlnpyys qcl

Figure S13. Sequences of other processive Family 7 cellobiohydrolases examined in the charge-hydropathy scale. The linkers screened in this algorithm are highlighted in yellow. The sequences of the Family 1 CBMs are highlighted in light blue. The boundary between the linker and the CBM are taken from the homology of the Family 1 CBMs in these organisms and *T. reesei* Cel7A. The boundary between the linker and the catalytic domain is less well defined.

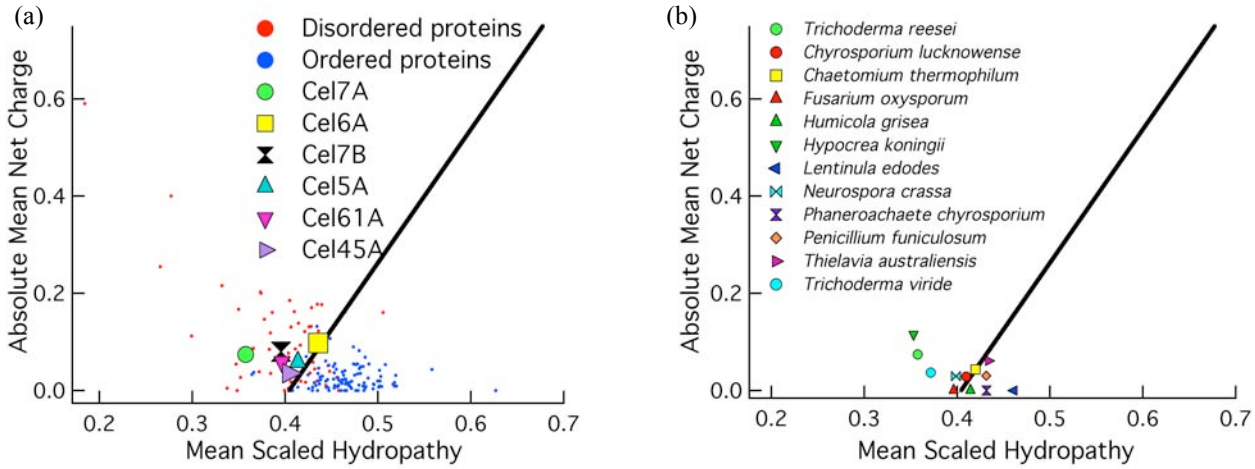



Figure S14. (a) Charge as a function of hydropathy for the Cel7A, the Cel6A and the four *T. reesei* endoglucanases linkers. The training sets for disordered and ordered proteins are shown in red and blue, respectively from (58). (b) Charge as a function of hydropathy for a library of Cel7A enzymes from other organisms. The black lines show the approximate delineation between ordered and disordered proteins.