Supporting Information

Sethaphong et al. 10.1073/pnas.1301027110

SI Materials and Methods

Approaches to computational structure prediction fall under the spectrum of knowledge-based algorithms spanning template-based modeling to use of physical force-fields (or de novo modeling) when no highly similar structures are available. Almost the entire region was modeled (506 amino acids; Q220–R725) beginning just after transmembrane helix 2 (TMH2). Only a small loosely conserved linker between the C-terminal region including QVLRW and TMH3 was excluded to reduce computational complexity. Knowledge based 3D structure prediction from a linear protein sequence was accomplished with the SAM-T08 prediction server of the K. Karplus aboratory (1) A FASTA file of the putative cytosolic domain amino acid sequence was submitted to the prediction server: http://compbio.soe.ucsc.edu/SAM_T08/T08-query.html. This method has exhibited good performances across diverse proteins, and high quality structures result when there is a good match between the target and available templates (1, 2). Two of the top selected structures were from the bacterial protein templates of spore coat polysaccharide biosynthesis protein (SpsA) and *Escherichia coli* K4 (K4CP) that have been extensively used to examine the molecular basis for catalysis and substrate recognition of glycosyltransferases (3–5). SpsA is a glycosyltransferase involved in producing the *Bacillus subtilis* spore coat that cocrystallized with Mg^{2+}- or Mn^{++}-UDP. K4CP catalyzes alternative transfers of glucuronic acid and *N*-acetylgalactosamine to form chondroitin (glycosaminoglycan) in *Escherichia coli* (3).

Because the resulting homology model, Fig. S2*D*, is fragmentary in form, it was initially manually refined with DS Visualizer from Accelerys to correct for steric clashes and breakages. An Amber molecular dynamics package with the force field FF99SB and TIP3P water model was used for relaxing this structure (6, 7). Atom types were converted into Amber-acceptable format via an in-house script before equilibration and subsequent MD production run.

All structures were subjected to conjugate gradient energy minimization for 5,000 steps. Minimized protein structures were then neutralized with Na^+ ions and immersed in a water box with at least 10 Å-deep solvation shell using the TIP3P water model (7). Additional Na^+ and Cl^- ions were added to represent a 0.3-M effective salt concentration. The equilibration of each system was carried out in 11 stages starting from the solvent minimization for 10,000 steps and keeping the protein restrained for 200 kcal/mol. The system was heated to 300 K in 100 ps while imposing a 200 kcal/mol constraint on the structure. A brief constant pressure (NPT) MD run was performed for 40 ps with the protein restraint maintained at 200 kcal/mol. Another constrained minimization step follows with the restraint of 25 kcal/mol for 10,000 steps. A second NPT MD run was performed at 25 kcal/mol restraint for 20 ps. Subsequently, four additional 1,000-cycle minimization steps were performed while relaxing the positional constraint from 20 kal/mol to 5 kcal/mol in 5 kcal/mol increments. A final unconstrained minimization stage of 1,000 cycles was performed before reheating the system to 300 K at constant volume within 40 ps. Subsequently, NPT equilibrations were performed to ensure uniformity in solvent density. Long-range electrostatic interactions were calculated by Particle Mesh Ewald summation (PME) (8), and the nonbonded interactions were truncated at 9 Å cutoff along with a 0.00001 tolerance of Ewald convergence. A Berendsen thermostat maintained temperature at 300 K (9). The SHAKE algorithm was used to constrain the position of hydrogen atoms (10). The production simulations were performed for a constant volume (NVT) ensemble. Each production simulation was performed for 10 ns with a 2-fs time step.

Intermediate structures were evaluated for quality; gross misfold errors were unfolded using a protocol starting directed MD with a harmonic force followed by free Langevin self-guided dynamics. Several series of such MD simulations were performed (for more than 150 ns simulations time) until a reasonable *z*-score was reached. The final structure from the MD simulations was energy minimized for 10,000 cycles with a convergence criterion of less than 1.0E-4 kcal/mole Å.

Initial evaluation of the final predicted structure of the native GhCESA1 cytosolic region was performed using Pro-SA (https://prosa.services.came.sbg.ac.at/prosa.php) (11). Two characteristics of the structure were derived: the *z*-score and a graphic of the residue energies. The *z*-score measures the deviation of the total energy from an energy distribution of random conformations, and an acceptable *z*-score of the computed structure must fall within the distribution of those derived from experimentally determined structures. High energy residues contributing to poor *z*-scores are likely areas that need further refinement or may have intrinsically high conformational entropy. The stereochemical quality of the intermediate and final structures was analyzed by PROCHECK (www.ebi.ac.uk/thornton-srv/software/PROCHECK/) (12). WhatCheck, another protein verification tool, was also used (http://swift.cmbi.ru.nl/gv/whatcheck/). The final structure was analyzed comprehensively using the protein structure validation software suite (PSVS; http://psvs-1_4-dev.nesg.org/), which integrates the analyses performed by PROCHECK, MolProbity, Verify3D, Prosa II, and the PDB validation software (13). Additional validation of our protein model was performed using ERRAT (14) (http://nihserver.mbi.ucla.edu/ERRATv2/), which is a protein structure verification algorithm mainly used to assess crystallographic models where a nine-residue sliding window is used to generate the value of the error function: ERRAT2 (Quality Factor). Earlier approaches that coupled a de novo prediction with further refinement under molecular dynamics simulations have not shown additive improvements (15). In this work, we achieved appreciable gains in structure quality over time (Fig. S2*E*).

The symmetric docking protocol of Rosetta 3.4 was used to generate homooligomeric assemblies (16); the algorithm allows translation occurring on the plane connecting the center of mass for the monomers. A slide degree of freedom is randomly chosen, and subunits are translated into contact. An optimization of the rigid body orientation proceeds with a Monte Carlo search under a low-energy resolution function followed by a high-resolution optimization of side-chain and rigid body conformation via Monte Carlo Minimization.

Related to docking UDP-Glc into the catalytic site, Density Functional Theory (DFT) calculations were carried out on the Mn^{2+}- and Mg^{2+}-UDP-Glc + DxD models using the B3LYP (17, 18) exchange and correlation functionals and the 6–311+G(d,p) basis set (19, 20) using the Gaussian 03 program (21). All atoms were allowed to relax without constraint or symmetry. After energy minimization, frequency analyses were performed to ensure an energy minimum had been found.

For the native predicted structure, as well as three mutant structures, the flexibility of each residue was assessed using molecular dynamic simulations (22). Each residue position was used as a variable in four simulations to generate four observations for each residue, allowing cross correlation analysis for coupled motions to be derived from the fluctuation data. The total atomic fluctuation data were calculated using the PTRAJ tool of Amber 11 (23) and

then imported into MATLAB (R2011a, MathWorks) with an in-house script to generate the correlation matrix. The input 4 × 506 matrix was constructed such that the rows corresponded to each simulation, with columns corresponding to individual residues.

1. Karplus K (2009) SAM-T08, HMM-based protein structure prediction. *Nucleic Acids Res* 37(Web Server issue):W492-7.
2. Das R, Baker D (2008) Macromolecular modeling with rosetta. *Annu Rev Biochem* 77:363–382.
3. Sobhany M, Kakuta Y, Sugiura N, Kimata K, Negishi M (2008) The chondroitin polymerase K4CP and the molecular mechanism of selective bindings of donor substrates to two active sites. *J Biol Chem* 283(47):32328–32333.
4. Keenleyside WJ, Clarke AJ, Whitfield C (2001) Identification of residues involved in catalytic activity of the inverting glycosyl transferase WbbE from Salmonella enterica serovar borreze. *J Bacteriol* 183(1):77–85.
5. Urresti S, et al. (2012) Mechanistic insights into the retaining glucosyl-3-phosphoglycerate synthase from mycobacteria. *J Biol Chem* 287(29):24649–24661.
6. Duan Y, et al. (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J Comput Chem* 24(16):1999–2012.
7. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 79:926–935.
8. Essmann U, et al. (1995) A smooth particle mesh Ewald method. *J Chem Phys* 103:8577–8593.
9. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR (1984) Molecular dynamics with coupling to an external bath. *J Chem Phys* 81:3684–3690.
10. Ryckaert JP, Ciccotti G, Berendsen HJC (1977) Numerical integration of the Cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J Comput Phys* 23:327–341.
11. Wiederstein M, Sippl MJ (2007) ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res* 35(Web Server issue):W407-10.
12. Laskowski RA, Macarthur MW, Moss DS, Thornton JM (1993) Procheck: A program to check the stereochemical quality of protein structures. *J Appl Cryst* 26:283–291.
13. Bhattacharya A, Tejero R, Montelione GT (2007) Evaluating protein structures determined by structural genomics consortia. *Proteins* 66(4):778–795.
14. Colovos C, Yeates TO (1993) Verification of protein structures: Patterns of nonbonded atomic interactions. *Protein Sci* 2(9):1511–1519.
15. Lee J, Wu S, Zhang Y (2009) Ab initio protein structure prediction. *From Protein Structure to Function with Bioinformatics*, ed Rigden DJ (Springer, London), Chap 1, pp 1-26.
16. André I, Bradley P, Wang C, Baker D (2007) Prediction of the structure of symmetrical protein assemblies. *Proc Natl Acad Sci USA* 104(45):17656–17661.
17. Becke AD (1993) A new mixing of Hartree-Fock and local density-functional theories. *J Chem Phys* 98:1372–1377.
18. Lee CT, Yang WT, Parr RG (1988) Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys Rev B Condens Matter* 37(2):785–789.
19. Krishnan R, Binkley JS, Seeger R, Pople JA (1980) Self-consistent molecular-orbital methods. 20. Basis set for correlated wave-functions. *J Chem Phys* 72:650–654.
20. Mclean AD, Chandler GS (1980) Contracted Gaussian-basis sets for molecular calculations. 1. 2nd row atoms, Z =11–18. *J Chem Phys* 72:5639–5648.
21. Frisch MJ, et al. (2004) *Gaussian 03* (Gaussian, Inc., Wallingford, CT).
22. Kormos BL, Baranger AM, Beveridge DL (2007) A study of collective atomic fluctuations and cooperativity in the U1A-RNA complex based on molecular dynamics simulations. *J Struct Biol* 157(3):500–513.
23. Case DA, et al. (2010) *Amber 11* (Univ of California, San Francisco).

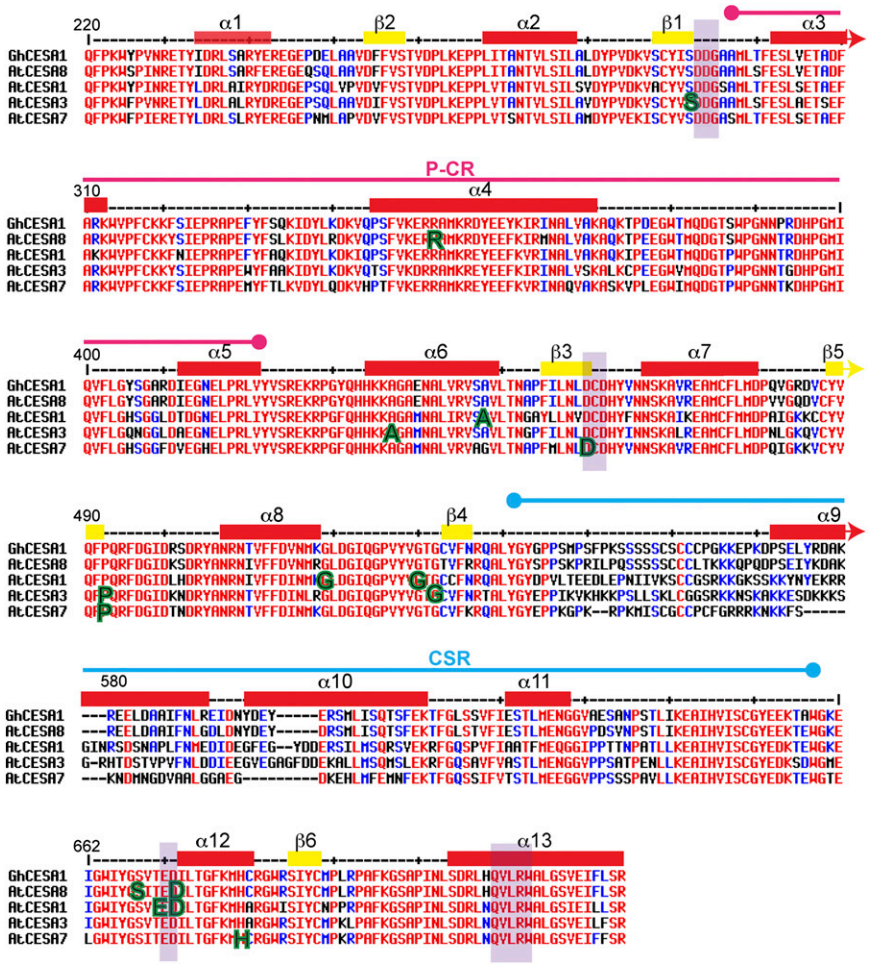


Fig. S1. Residues from GhCESA1 that were included in the Gh506 structure are aligned with the same regions of *Arabidopsis* CESAs with missense mutations. Numbering is relative to residue position in full-length GhCESA1. Plant-specific regions in CESA are highlighted by pink and blue lines, which indicate the positions of the plant-conserved region (P-CR) and class-specific region (CSR), respectively. Red and yellow rectangles indicate α–helices and β–sheets, respectively. By comparison with the structure of RsBcsA (see the main text), α2, -6, -7, -8, and -13 and β1–6 are predicted to be in the core GT domain. Light purple vertical highlights show the position of selected conserved domains. Large green letters indicate sites of missense mutation in the AtCESA indicated.

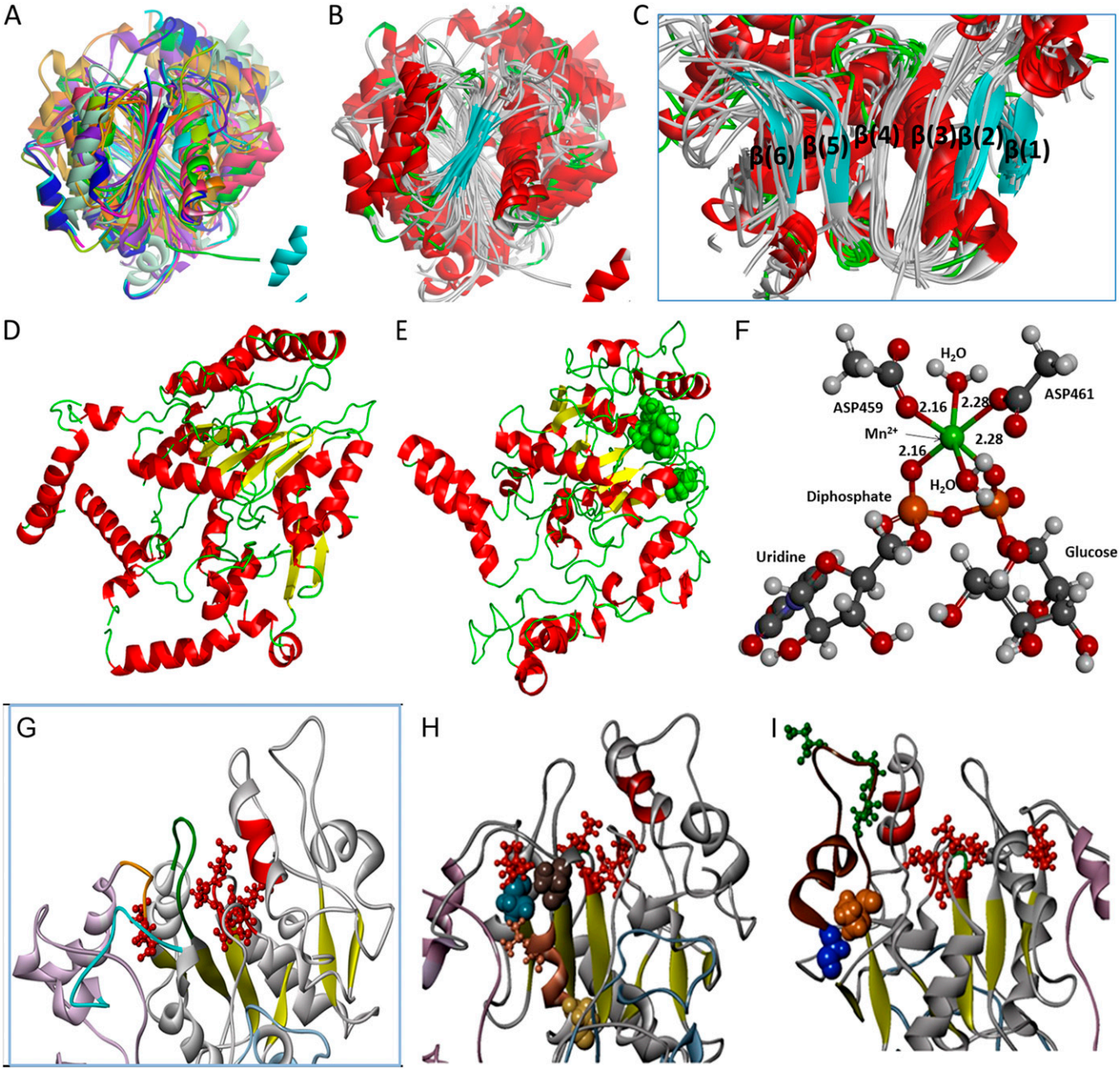
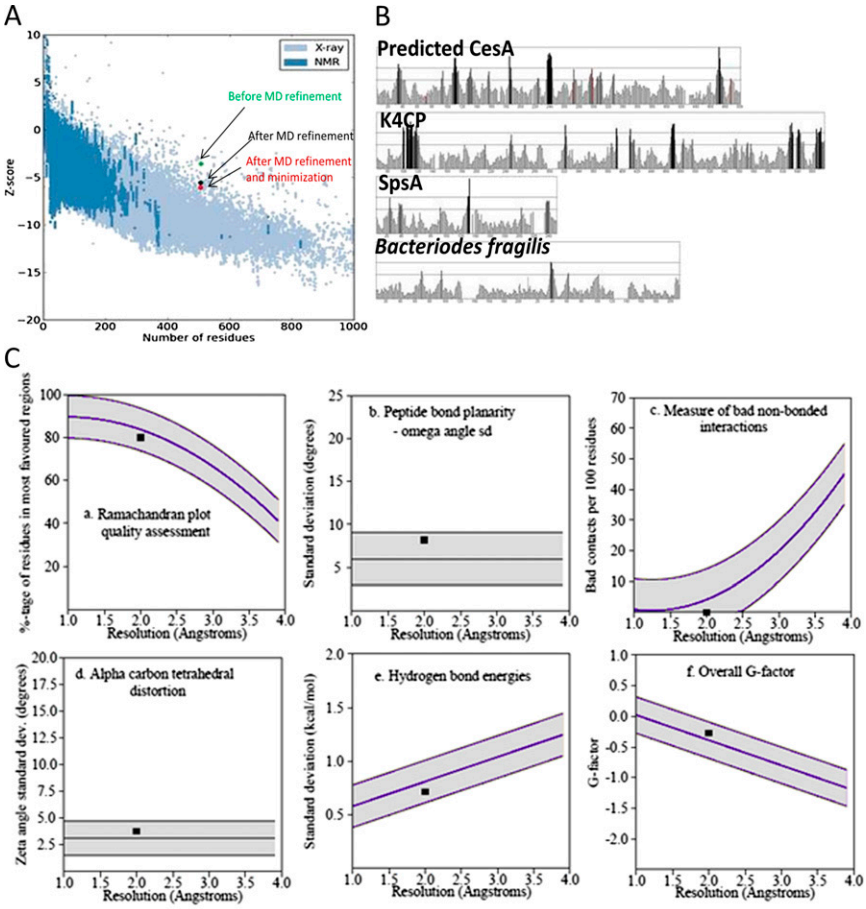


Fig. S2. (*A*–*C*) Aligned structures used in model prediction as listed in Table S1. Side view of the β-strands (*A*) colored by individual structure and (*B*) colored by secondary structure. (*C*) View of the slice to expose the β-sheet region, with individual β-strands numbered β1–β6. (*D*) The snapshot of the starting structure from the SAM-T08 HMM structure prediction server. (*E*) The predicted structure after molecular dynamics refinement with six β-strands in yellow and DD, DCD, and ED in green. The α-helices dispersed throughout the structure are red. (*F*) Interaction of manganese uridine diphosphate glucose (MnUDP-G) complex with residues of the modeled CESA. The positions of the "D" residues were taken from the CESA structure generated in this study, and all atomic positions were allowed to relax to minimum energy positions determined by our DFT methodology. Mn-O distances to carboxylate group of the D residues and to the diphosphate moiety of UDP are given in Angstroms. H, white; C, gray; O, red; N, blue; P, orange; Mn, green. This geometry was used to dock the UDP-Glc into the Gh506 structure in Fig.1. (*G*) Three loops in the vicinity of the UDP-Glc binding site of the Gh506 structure that may help to control catalysis through modulation of local accessibility to key residues: (*i*) T258–L267 at the end of β-2 (green); (*ii*) A294-F300, just after DDG and leading into α3 of the PCR (orange); and (*iii*) Y421-H432, leading from α5 into core α6 (aqua). The conserved motifs DD, DCD, ED, and QLVRW are highlighted red, the β sheet is yellow, and the P-CR is pink. (*H* and *I*) The locations of previously undescribed missense mutations in the predicted structure helped to support the existence of previously undescribed functionally important regions within CESA. Conserved residues are shown in red. (*H*) S291 (teal) just below DD is the analog of the previously undescribed *Atcesa3*^{S377F}, *ixr1-6*, mutation. In the predicted structure, it contacts L442 (rust ball and stick residue) within α-6 (rust), which has the analogs of *Atcesa3*^{A522V} (*eli1-2*; brown) and *Atcesa1*^{A549V} (*rsw1-1;* tan) at either end. (*I*) The P492-G518 loop (brown) contains native aspartates (green ball and stick residues) near QVLRW. At its base are G518 (blue; the analog of the previously undescribed *Atcesa1*^{G620E}, *lycos*, mutation) and P492 (rust; the analog of *Atcesa7*^{P557T}, *fra5*), where they may putatively act as hinge points.



Stereochemical parameter	No. of data pts	Parameter value	Comparison values Typical value	Band width	No. of band widths from mean	
a. %-tage residues in A, B, L	440	80.0	83.8	10.0	-0.4	Inside
b. Omega angle st dev	494	8.2	6.0	3.0	0.7	Inside
c. Bad contacts / 100 residues	0	0.0	4.2	10.0	-0.4	Inside
d. Zeta angle st dev	474	3.8	3.1	1.6	0.4	Inside
e. H-bond energy st dev	250	0.7	0.8	0.2	-0.5	Inside
f. Overall G-factor	506	-0.3	-0.4	0.3	0.4	Inside

Fig. S3. Comparison of the quality of the Gh506 structure to experimentally solved structures. (*A*) Pro-SA Z scores for various stages of GhCESA1 structure prediction (labeled green, black, and red dots) compared with scores of solved structures from the PDB databank (dense blue dots). The initial *Z* score of the predicted GhCESA1 cytosolic structure (−3.4, green dot) was improved to −5.56 (black dot) after about 4 ns of MD refinement and reached −6.09 (red dot) after a series of MD simulations followed by a short minimization. (*B*) ERRATv2 analysis of the predicted GhCESA1 cytosolic structure (graph 1) and the solved structures of three other GT-2 enzymes used as templates [graph 2, K4CP domains A and B (PDB: 2Z86); graph 3, SpsA (PDB: 1QG8); and graph 4, a putative glycosyltransferase from *Bacteriodes fragilis* (PDB: 3BCV)]. The histograms show the error value of residues, and the band in the middle of the graph indicates the difference between the lower 95% and the upper 99% value. Of the three crystal structures, the 218 amino acid structure of 3BCV from *B. fragilis* exhibited the best score with only B chain residue 40 showing significant error. Areas possibly in need of further refinement in the GhCESA1 predicted structure include residues that either have high local mobility or are deeply buried: (*i*) N457-V464; (*ii*) D253-V256 that form a β-strand adjacent to the putative UDP binding motif, DCD, in the catalytic core; (*iii*) solvent-exposed P327-I335 that fold back into residues V347-R355 within the P-CR region; (*iv*) P492-G518 that appear to form a loop beside the catalytic site that abuts the QVLRW motif. Even for the SpsA structure, similarly buried residues are nearly impossible to refine fully. For K4CP, core residues around the UDP binding motif of domain "B" shows the greatest error values, probably because they are more mobile and solvent accessible. Similarly, a small region near the UDP-binding motif of SpsA (residues 130–135) also exhibits error values greater than 95% as exemplified by the filled in black bars. (*C*) Resolution of main chain parameters of Gh506 compared with solved crystallographic structures assessed by ProCheck. In the graphs, the value for the predicted GhCESA1 cytosolic structure is shown by the black square relative to values typical for solved structures (gray band): (*a*) Ramachandran plot quality is the percentage of the residues in the most favored regions of the Ramachandran plot where a high quality structure is well over 90%, but becomes less at lower resolutions; (*b*) peptide bond planarity is a measure of the structure's ω-torsion angle where a tight clustering around the ideal 180° represents a planar peptide bond; (*c*) bad nonbonded interactions are defined by the number of bad contacts less than or equal to 2.6 Å per 100 residues; (*d*) C-alpha tetrahedral distortion measures the SD of the zeta torsion angle defined by C-α, N, C and C-ε atoms of a given residue; (*e*) main-chain hydrogen bond energy is derived from the measured SD of the hydrogen bond energies in the main chain by the method of Kabsch and Sanders (1983) (1); (*f*) overall G-factor measures the overall normality of the structure as an average of all of the different G-factors for each residue.

1. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22(12):2577–2637.

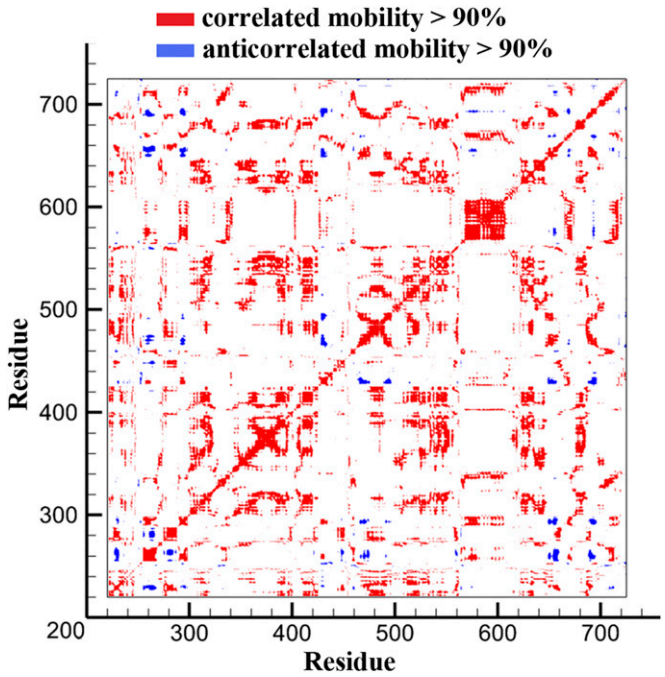


Fig. S4. Correlated residue motions via atomic fluctuations. The CSR region, residues Y540-W658, shows the greatest motion correlation to itself as expected. The P-CR region, residues A295-V420, shows a self-correlation as well, but not as strong because it is less ordered.

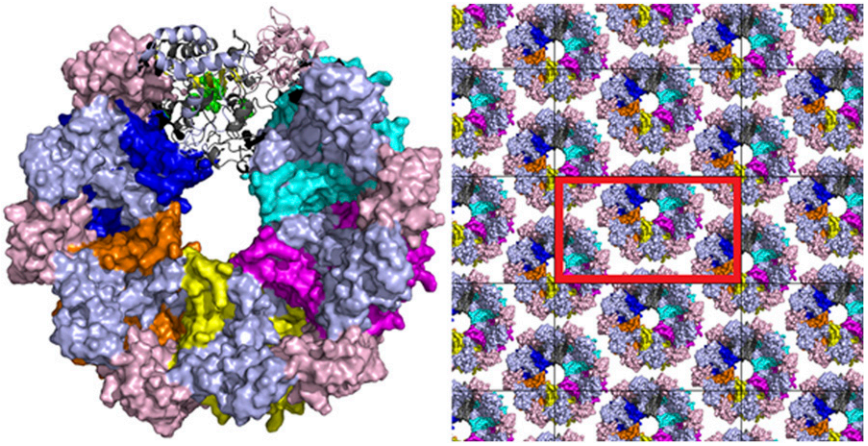


Fig. S5. A possible hexameric assembly of one CESA cytosolic domain isoform (the predicted structure from GhCESA1). One monomer is shown in the ribbon diagram at the top, showing the location of the barely visible β-sheets (yellow) below motifs with conserved D residues (green). The catalytic regions of the other monomers are shown in aqua, magenta, yellow, orange, and dark blue. The light blue and pink regions are the CSR and the P-CR regions, respectively, for all monomers. (*Right*) Possible packing of hexameric assemblies into an orthorhombic unit cell of space group $P2_12_12_1$ (red box). Note that this theoretical possibility for crystallization of hexamers of the predicted GhCESA1 cytosolic region does not imply any preference for hexameric subunits of the rosette CSC in vivo. The number of CESAs in the rosette CSC remains an open question.

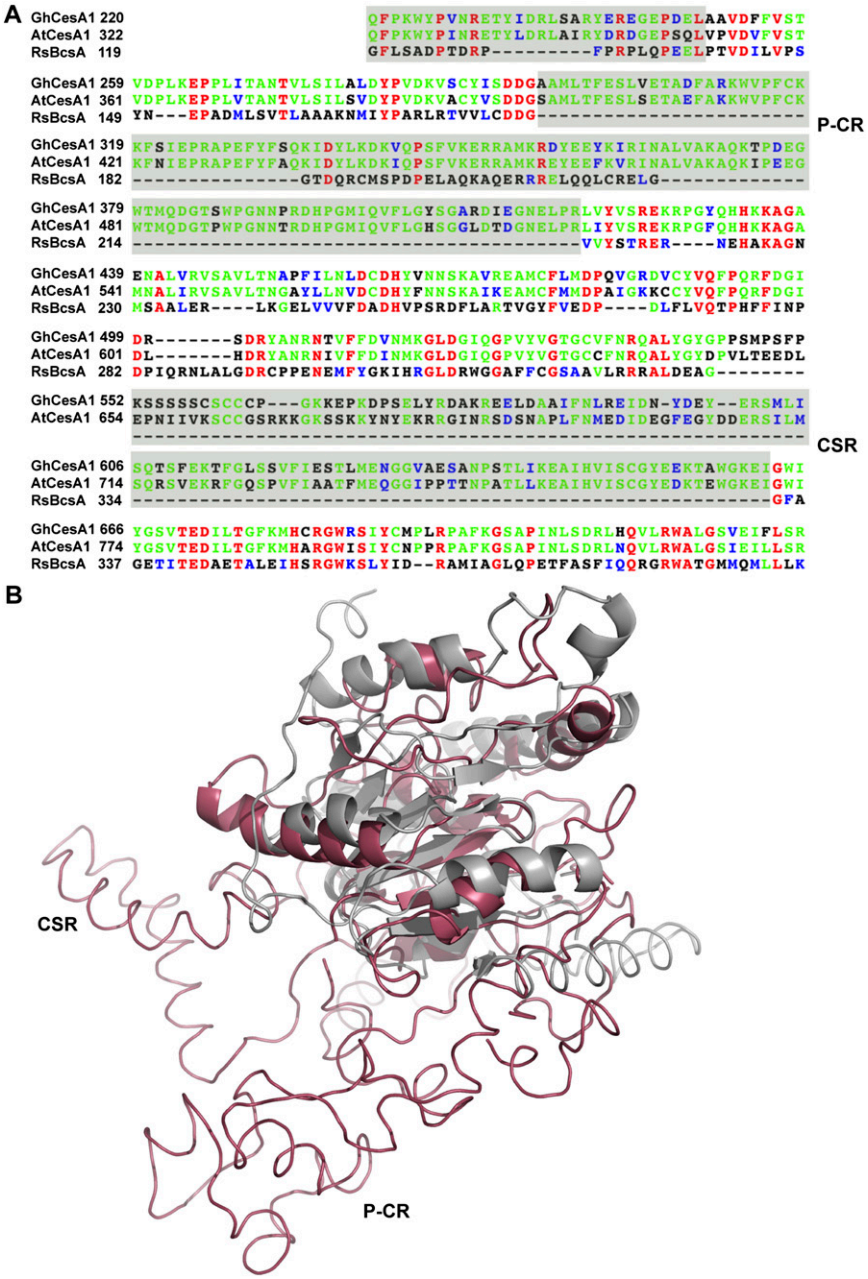


Fig. S6. Sequence and structural alignment of Gh506 and RsBcsA. (*A*) A sequence alignment of the GT-domains of GhCesA1, AtCesA1 and RsBcsA. The alignment is color coded based on sequence similarity. The shaded regions indicate sequences with no template in RsBcsA or weak sequence similarity. (*B*) Alignment of the GT-domains from RsBcsA and Gh506 based on secondary structure matching. Regions used for secondary structure matching are shown as cartoon, omitted regions (shaded gray in *A*) are shown as backbone ribbon. Gh506 and RsBcsA are colored red and gray, respectively.

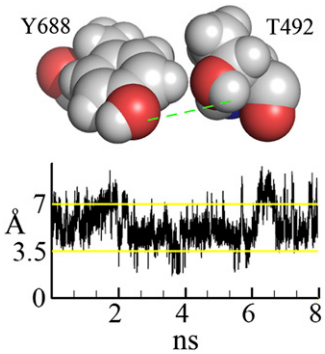


Fig. S7. Hydrogen bonding of P492T to Y688. The distance cut off is 3.5 Å. The strongest interaction during this time interval for *Ghcesa*^{P492T} is before the 4-ns mark. This interaction may serve to stabilize the P492-G518 loop.

Table S1. The PDB identification numbers, E-values, and snapshots of structures used in predicting the structure of the β-sheet region of the GhCESA1 cytosolic region using Hidden Markov chain modeling

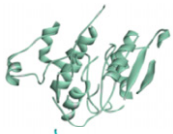
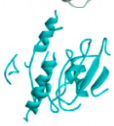
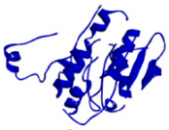
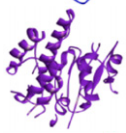
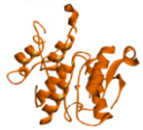
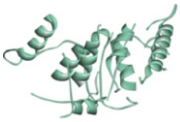
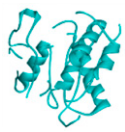
No.	PDB ID	Description	E-value	Snapshot of part of the structure used for prediction
1	1xhb	Crystal structure of UDP-GalNAc:polypeptide alpha-*N*-acetylgalactosaminyltransferase-T1	1.1661e−21	
2	2z86	Crystal structure of chondroitin polymerase from *Escherichia coli* strain K4 (K4CP) complexed with UDP-GlcUA and UDP	1.6825e−20	
3	2ffu	Dynamic association between the catalytic and lectin domains of human UDP-GalNAc:polypeptide alpha-*N*-acetylgalactosaminyltransferase-2	4.7760e−20	
4	3ckj	Essential GT (MAP2569c) from *Mycobacterium avium* subsp. paratuberculosis	1.6086e-18	
5	3bcv	Putative glycosyltransferase from *Bacteroides fragilis*	4.9831e−18	
6	1qg8	SpsA from *Bacillus subtilis*	6.8435e−18	
7	2bo4	Mannosylglycerate Synthase	1.5988e−17	
8	1omz	Alpha 1,4-*N*-acetylhexosaminyltransferase (EXTL2)	2.5389e−16	
9	1fo8	Rabbit *N*-acetylglucosaminyltransferase I	5.2404e−14	
10	2nvx	RNA polymerase II (pol II)	7.9916e−14	
11	2zu9	Mannosyl-3-phosphoglycerate synthase from *Pyrococcus horikoshii*	1.5238e−12	
12	1yro	Bovine beta-1,4-galactosyltransferase I	3.6727e−08	

Table S1. Cont.

No.	PDB ID	Description	E-value	Snapshot of part of the structure used for prediction
13	2fy7	Beta-1,4-galactosyltransferase-I	2.7044e−06	
14	1i52	4-Diphosphocytidyl-2-C- methylerythritol synthetase	2.4158e−01	
15	1fgx	Bovine beta-4-galactosyltransferase catalytic domain	2.7667e−01	
16	2vsh	CDP-activated ribitol for teichoic acid precursors in*Streptococcus pneumoniae*	4.0782e−01	
17	2px7	2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase from *Thermus thermophilus* HB8	1.5287e+00	
18	3cgx	Putative Nucleotide-diphospho-sugar Transferase (YP_389115.1) from *Desulfovibrio desulfuricans* G20	1.7882e+00	
19	1pzt	Beta1,4-galactosyltransferase-I	2.2884e+00	
20	1ezi	Sialic acid-activating synthetase, CMP-acylneuraminate synthetase in the presence and absence of CDP	4.3372e+00	

During the selection of the top models, the SAM-T08 generates pairwise alignments of the target sequence and the best-scoring templates, which are adjudicated by E-value representing how many sequences would score this well in the database. Structures with E-values less than about 1.0E−5 are very likely to have a domain of the same fold as the target. Structures with E-values larger than about 0.1 are very speculative.

Table S2. Structure quality scores

Structure	ProSA *Z*-score	Quality factor (ERRAT2), %	AA Length
GhCESA1	−6.09	86.875	504
SpsA (1qg8)	−7.8	92.411	241
K4CP (2z86)	−9.16	86.067	580
(3BCV)	−6.98	98.082%	196

Table S3. Identity and locations of Gh506 structural features

Gh506 major secondary structure elements	Position in GhCESA1, including additional key motifs	Amino Acid Sequence in GhCESA1 of major secondary structure elements and additional key motifs	GhCESA1 residues analogous to *Arabidopsis* CESA mutations	Structurally coaligned motifs in the BcsA and the Gh506 GT-2 domains
α-1	I233–E241	IDRLSARYE		
Core β-2	D253–S257	DFFVS		VDILVPS148
Core α-2	L267–A278	LITANTVLSIL		ADMLSVTLAAAKN165
Core β-1	S287–S291	SCYIS	S291: *Atcesa3*^{S377F} *ixr1-6* (this paper)	LRTVVLCD179
	D292–G294	DDG		DDG181; D179 coordinates UDP
α-3	E301–K312	ESLVETADFARK		
α-4	P344–K370	PSFVKERRAMKRDYEEYKIRINALVAK, in the P-CR	R351: *Atcesa8*^{R362K} *fra6* (2)	
α-5	I411–V420	IEGNELPRLV, ending the P-CR		
Core α-6	H433–V448	HKKAGAENALVRVSAV; the HKKAGA motif is near DDG.	A436: *Atcesa3*^{A522V} *eli1-2* (3) A447: *Atcesa1*^{A549V} *rsw1-1* (4)	HAKAGN229; A225 and K226 lie on the other side of the pocket that may accommodate Glc when bound to UDP.
Core β-3	F454–D459	FILNLD; including the first D of DCD		LVVVF245
	D459–D461	DCD	D459: *Atcesa7*^{D524N} *irx3-5* (5)	DADH249; D246 coordinates UDP
Core α-7	N466–D479	NSKAVREAMCFLMD; crosses several β-strands leading toward DCD		FLARTVGY262
Core β-5	Y488–F491	YVQF		LVQT274
	P492–G518	PQRFDGIDRSDRYANRNTVFFDVNMKG (loop between β-4,5 and behind QVLRW), contains core α-8	P492: *Atcesa7*^{P557T} *fra5* and *thanatos* (2, 6) G518: *Atcesa1*^{G620E} *lycos* (this paper)	
Core α-8	N508–K517	NTVFFDVNMK, within the P492–G518 loop. A longer sequence N508-I521, (NTVFFDVNMKGLDGI), shares sequence conservation of N..F. . . . . .GLD.. with Rs_BcsA.		Interfacial Helix 1, N298-W312: NEMFYGKIHRGLDRW312,
	V525–G531	VYVGTG531, at the end of β-4	G529: *Atcesa1*^{G631S} *rsw1-2* (7) G531: Atcesa3^{G617E} *cev1* (8)	FFCGS320, binds the terminal disaccharide of the glucan acceptor on the opposite side compared with QRGRW
Core β-4	C532–N535	CVFN, just before the CSR		AVLR325
α-9	P571–R591	PSELYRDAKREELDAAIFNLR, in the CSR		
α-10	Y596–K612	YDEYERSMLISQTSFEK, in the CSR		
α-11	E622–G629	ESTLMENG, in the CSR	S668: *Atcesa8*^{S679L} *irx1-2* (9)	
	T670–D672	TED	E671: *Atcesa1*^{E779K} *rsw1-45* (10) D672: *Atcesa8*^{D683N} *irx1-1* and *Atcesa1*^{D780N} *rsw1-20* (9, 10)	TED343, near the glucan terminus with D343 likely to be the catalytic base. E342 lies on one side of a pocket that may accommodate Glc when bound to UDP
α-12	I673–C681	ILTGFKMHC	H680: *Atcesa7*^{H734Y} *mur10-2* (11)	
Core β-6	S686–C689	SIYC		SLYI360
Core α-13	S705–R725	SDRLHQVLRWALGSVEIFLSR, containing QVLRW		Interfacial Helix 2, F373-R395: FASFIQQRGRWATGMMQMLLLK. Contains QRGRW383. R382 coordinates UDP and W383 interacts with the penultimate glucose at the acceptor site

Entries are in order of appearance in the GhCESA1 cytosolic sequence that was used to generate the Gh506 structure (Fig. 1*B*). Five of these α-helices are designated "core α-helices" because they coalign in the superimposed GT-2 domain of BcsA and the predicted Gh506 structure. Amino acid residue numbers are relative to full-length GhCESA1 (NCBI accession no. P93155) or BcsA (NCBI accession no. Q3J125; PDB ID 4HG6). Functions ascribed to BcsA are from ref. 1. The nomenclature used to identify the Arabidopsis CESA missense mutations here and in the text is as follows. The name of the mutated Arabidopsis AtCESA gene is shown in lower case italics with its superscript showing the affected amino acid. The common names and allele numbers assigned to the mutations are also shown, and some of these are abbreviations as follows: isoxaben resistant (*ixr*), fragile fiber (*fra*), ectopic lignification (*eli*), radially swollen (*rsw*), irregular xylem (*irx*), constitutive expression of VSP (*cev*), and murus (*mur*).

1. Morgan JLW, Strumillo J, Zimmer J (2013) Crystallographic snapshot of cellulose synthesis and membrane translocation. *Nature* 493(7431):181–186.
2. Zhong RQ, Morrison WH, 3rd, Freshour GD, Hahn MG, Ye ZH (2003) Expression of a mutant form of cellulose synthase AtCesA7 causes dominant negative effect on cellulose biosynthesis. *Plant Physiol* 132(2):786–795.
3. Caño-Delgado A, Penfield S, Smith C, Catley M, Bevan M (2003) Reduced cellulose synthesis invokes lignification and defense responses in Arabidopsis thaliana. *Plant J* 34(3):351–362.
4. Arioli T, et al. (1998) Molecular analysis of cellulose biosynthesis in Arabidopsis. *Science* 279(5351):717–720.
5. Liang YK, et al. (2010) Cell wall composition contributes to the control of transpiration efficiency in Arabidopsis thaliana. *Plant J* 64(4):679–686.

6. Daras G, et al. (2009) The thanatos mutation in Arabidopsis thaliana cellulose synthase 3 (AtCesA3) has a dominant-negative effect on cellulose synthesis and plant growth. *New Phytol* 184(1):114–126.
7. Gillmor CS, Poindexter P, Lorieau J, Palcic MM, Somerville C (2002) Alpha-glucosidase I is required for cellulose biosynthesis and morphogenesis in Arabidopsis. *J Cell Biol* 156(6): 1003–1013.
8. Ellis C, Karafyllidis I, Wasternack C, Turner JG (2002) The Arabidopsis mutant cev1 links cell wall signaling to jasmonate and ethylene responses. *Plant Cell* 14(7):1557–1566.
9. Taylor NG, Laurie S, Turner SR (2000) Multiple cellulose synthase catalytic subunits are required for cellulose synthesis in Arabidopsis. *Plant Cell* 12(12):2529–2540.
10. Beeckman T, et al. (2002) Genetic complexity of cellulose synthase a gene function in Arabidopsis embryogenesis. *Plant Physiol* 130(4):1883–1893.
11. Bosca S, et al. (2006) Interactions between MUR10/CesA7-dependent secondary cellulose biosynthesis and primary cell wall structure. *Plant Physiol* 142(4):1353–1363.

Table S4. Summary stability measurements measured as root mean square deviation (rmsd) from the initial structure on the whole structure and on key secondary structure elements of the CESA as a result of mutations over a window of 10 ns

rmsd of a motif, Å	Gh506	P557T	G620E	S377F
All	2.69 ± 0.55	3.14 ± 0.53	3.01 ± 0.70	2.62 ± 0.41
α-2 helix	0.70 ± 0.14	0.88 ± 0.15	0.62 ± 0.25	0.72 ± 0.24
α-3 helix	1.21 ± 0.51	0.93 ± 0.35	0.63 ± 0.33	1.00 ± 0.24
α-7 helix	0.89 ± 0.27	0.93 ± 0.27	0.60 ± 0.10	0.79 ± 0.21
α-9 helix	0.57 ± 0.23	0.50 ± 0.12	0.77 ± 0.32	0.62 ± 0.15
α-11 helix	0.42 ± 0.09	0.44 ± 0.10	0.34 ± 0.10	0.42 ± 0.10
α-13 helix	0.81 ± 0.14	1.14 ± 0.31	0.53 ± 0.12	0.60 ± 0.20
α-6 helix	0.81 ± 0.14	0.44 ± 0.10	0.46 ± 0.11	0.60 ± 0.20
Loop P492-G518	1.65 ± 0.30	1.15 ± 0.14	1.37 ± 0.31	1.55 ± 0.54
Loop S257-P266	0.52 ± 0.10	0.50 ± 0.10	0.45 ± 0.12	0.60 ± 0.13
Loop Y430 -N440	1.10 ± 0.18	0.65 ± 0.17	0.85 ± 0.17	1.18 ± 0.21
Angle formed by residues 598, 608, and 572 with 608 at the vertex (degrees)	79.26 ± 6.59	76.14 ± 4.80	80.32 ± 12.38	81.56 ± 6.68

Other Supporting Information Files

Dataset S1 (TXT)