SCIENTIFIC REPORTS

OPEN

Received: 22 November 2017
Accepted: 5 March 2018
Published online: 22 March 2018

Dual binding in cohesin-dockerin complexes: the energy landscape and the role of short, terminal segments of the dockerin module

Michał Wojciechowski[1], Bartosz Różycki[1], Pham Dinh Quoc Huy[1,2], Mai Suan Li[1], Edward A. Bayer[3] & Marek Cieplak[1]

The assembly of the polysaccharide degradating cellulosome machinery is mediated by tight binding between cohesin and dockerin domains. We have used an empirical model known as FoldX as well as molecular mechanics methods to determine the free energy of binding between a cohesin and a dockerin from *Clostridium thermocellum* in two possible modes that differ by an approximately 180° rotation. Our studies suggest that the full-length wild-type complex exhibits dual binding at room temperature, i.e., the two modes of binding have comparable probabilities at equilibrium. The ability to bind in the two modes persists at elevated temperatures. However, single-point mutations or truncations of terminal segments in the dockerin result in shifting the equilibrium towards one of the binding modes. Our molecular dynamics simulations of mechanical stretching of the full-length wild-type cohesin-dockerin complex indicate that each mode of binding leads to two kinds of stretching pathways, which may be mistakenly taken as evidence of dual binding.

The fibrous plant cell walls are the major source of carbon and energy on Earth. They are made primarily of cellulose and hemicellulose, which are difficult to degrade into simple sugars. The simple sugars can be transformed to ethanol, to produce biofuels, through fermentation. In nature, the degradation process takes place as a result of the catalytic action of microbially produced enzymes. In particular, many anaerobic organisms have developed special extracellular organelles, known as cellulosomes[1–9], which are very efficient at performing degradation. The efficiency is accomplished through binding of many enzymes to a large noncatalytic protein known as scaffoldin so that the enzymes can act in one location together instead of being dispersed.

Scaffoldin consists of a number of covalently linked cohesins (Coh) and carbohydrate binding modules (CBM) that anchor to polysaccharide chains. The cohesins bind tightly and specifically to dockerin-bearing cellulosomal subunits, which comprise a dockerin domain (Doc), one or more enzymatic domains, and sometimes CBMs. These domains are typically connected by disordered polypeptide segments, which are often termed linkers. The amino-acid sequences of the linkers differ from one dockerin to another. In general, the longer the linker, the less restrictive the tethering constraints, which thus facilitates the interaction of the catalytic domain with the substrate[10–12].

A number of atomic structures of cellulosome-constituting domains and subunits have been solved by X-ray crystallography and NMR. However, there is no single method that could be used to solve structures of the cellulosomes: they are not directly accessible to X-ray crystallography due to the presence of the disordered linkers (although their constituent domains can be crystallized separately); they are not accessible to protein NMR because of their large sizes; and their inherent flexibility makes them still practically inaccessible to cryoEM. Therefore, to delineate conformations of such protein assemblies as the cellulosomes, various complementary methods need to be combined[13,14]. In particular, small angle X-ray scattering (SAXS) in solution is increasingly used to complement protein crystallography in structural studies on multi-protein complexes, including the

[1]Institute of Physics, Polish Academy of Sciences, Al. Lotników 32/46, PL-02668, Warsaw, Poland. [2]Institute for Computational Sciences and Technology, SBI building, Quang Trung Software city, Tan Chanh Hiep Ward, District 12, Ho Chi Minh City, Vietnam. [3]Department of Biomolecular Sciences, The Weizmann Institute of Science, 234 Herzl Street, Rehovot, 7610001, Israel. Correspondence and requests for materials should be addressed to M.C. (email: mc@ifpan.edu.pl)

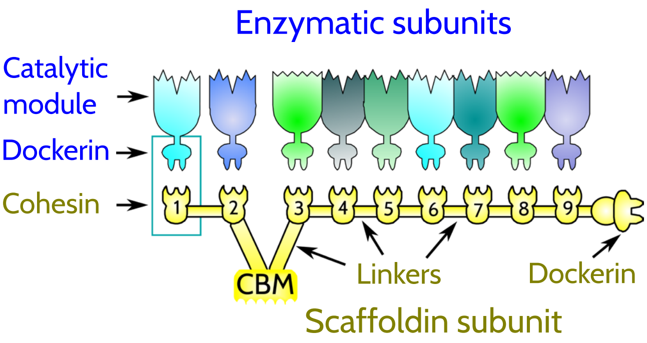

Figure 1. Schematic representation of cellulosome organization in *C. thermocellum*. The cellulosomal structural and enzymatic subunits comprise modular components. The cellulosomal enzymes each carry one or more catalytic modules and a single dockerin module. The scaffoldin is multifunctional, whereby the nine cohesin modules (enumerated) integrate nine dockerin-bearing enzymes into the complex, the carbohydrate-binding module (CBM) binds the complex to the cellulosic substrate, and the C-terminal dockerin module is involved in anchoring the cellulosome complex to the bacterial cell surface. The modules are separated from each other by defined linker segments.

cellulosomal complexes[15–22]. In addition, these structural methods combined with molecular dynamics simulations or other modeling methods can lead to insights into dynamic properties and conformational heterogeneity of the protein complexes under study[14,22,23].

The prototypic cellulosome-producing bacterium is *Clostridium thermocellum*. The typical architecture of the cellulosome that it makes is shown in Fig. 1. Its major scaffoldin, commonly denoted by CipA[24], consists of nine type-I Cohs that show large sequence similarity. The corresponding Docs also exhibit large sequence similarity. However, the similarity is specific to a given organism. Also, the numbers of the Coh modules in various organisms are distinct and range from only two Cohs in the scaffoldin of *Clostridium saccharoperbutylacetonicum*[25] to eleven Cohs in the major scaffoldin of *Bacteroides cellulosolvens*[26,27]. The scaffoldins in many mesophilic bacteria commonly bear five or six Cohs[25].

The question we address here is whether binding of a given Doc to Coh occurs in a variety of ways or just one. Structural analyses of the Coh-Doc complex (derived from three organisms: *C. thermocellum*[28,29], *C. cellulolyticum*[30] and *Ruminococcus flavefaciens*[31]) indicate a possibility of a dual binding that would reduce the conformational constraints within dockerins, facilitating access to the substrate and, thus, enhancing the efficiency of degradation. These structural analyses are further supported by computational and biochemical studies on the contributions of distinct amino acid residues to the Coh-Doc binding affinity[32].

The dual binding[33] bears some similarities to domain swapping[34–36]. It should be noted, however, that these two phenomena are distinct. Firstly, the free energy difference between the monomers and the domain-swapped oligomers are affected by such factors as conformational changes at the inter-domain linker, or hinge[36]. In contrast, the dual binding implies equal free energies in the two binding modes, which precludes crystallization. Secondly, the nature of dual binding is dynamical–the two binding modes may exchange in time. On the other hand, domain swapping is essentially fixed just after two ribosome-borne chains combine[34,35].

The structure of Doc involves three α-helices, which are denoted here as α_1', α_2' and α_3', as shown in panels A and B of Fig. 2. The idea of dual binding stems from a doubled loop-helix sequence in the Doc, wherein both sequence pieces contain a calcium-binding motif and repeated amino acid residues that can interact with the Coh in a similar manner, as illustrated in panels A and B of Fig. 2. As a result of this doubled loop-helix sequence, Doc may bind to Coh through the first and third of its helices but there are two ways in which this could happen, which differ by a rotation of approximately 180°, as shown in Fig. 2. For *C. thermocellum*, the first way, denoted here as mode I, is evidenced by the structure PDB:1OHZ, which has been derived for the wild-type (WT) complex[28] involving the second Coh (c2A) of CipA[37]. The second way, denoted here as mode II, is exemplified for the same Coh by the structure PDB:2CCL, which has been derived for the complex in which Doc has undergone two single-site mutations (S45A and T46A) at the C-terminal helix[29]. The two structures are shown superimposed in panel C of Fig. 2. Despite the existence of the mutations, the two molecules are structurally very close. When the Docs of 1OHZ and 2CCL are sequence-aligned and superimposed, their α-C RMSD is equal to 0.42 Å. A similar alignment and superposition of the Cohs yields the RMSD of 0.40 Å. However, the whole complexes differ in RMSD by about 9 Å.

Two distinct binding modes, which resemble those observed in the crystal structures PDB:1OHZ and PDB:2CCL, have been identified in coarse-grained molecular dynamics simulations of the Coh-Doc association process[38]. An experimental demonstration of the dual binding has been reported by Jobst *et al.*[39]. It involved single molecule force spectroscopy (SMFS) of a Coh-Doc exocellulase Cel48S complex from *C. thermocellum* flanked by other proteins and polymers. The experiment indicated existence of two kinds of stretching trajectories which was interpreted as evidence for the dual binding. However, our theoretical analysis[40] has found a possible alternative explanation of the stretching data: for a given binding mode there are two kinds of stretching trajectories that reflect different structural transformations in Coh and Doc. It may also happen that the dual binding

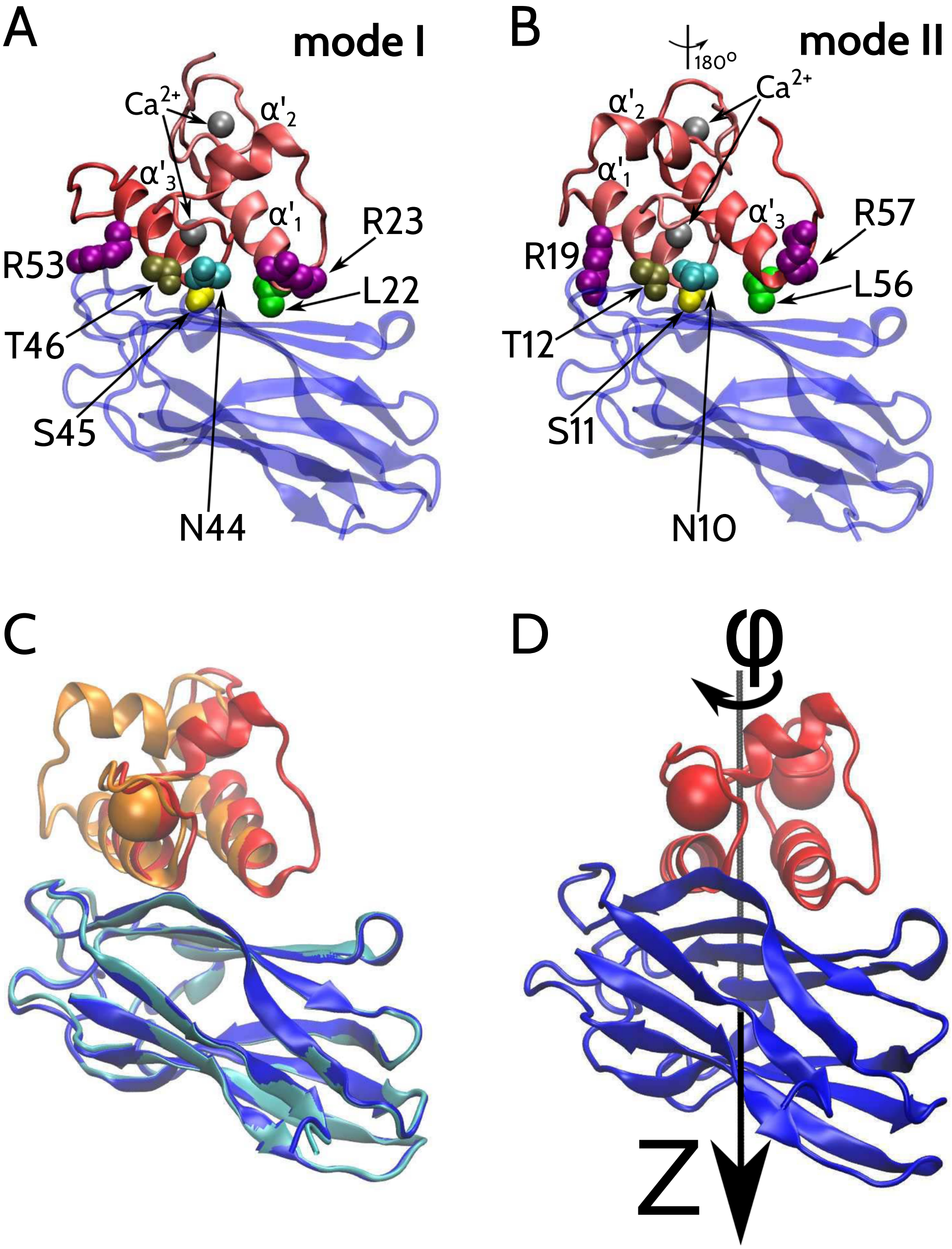

Figure 2. The top panels show structures of the Coh-Doc complex bound in mode I (**A**) and II (**B**). Coh and Doc are shown in blue and red, respectively. The spatial orientation of Coh is the same on both panels. The structure of Doc on panel B (mode II) is rotated by 180° relative to the structure of Doc on panel (A) (mode I). The three α-helices forming the Doc domain are labeled as α_1', α_2' and α_3'. The Doc-bound Ca^{2+} ions are shown as gray spheres. The Doc residues making contacts with the Coh in the binding mode I include R23, L22, N44, S45, T46 and R53. They are shown in the van der Waals representation. The Doc residues making contacts with the Coh in mode II include R57, L56, N10, S11, T12 and R19. The locations of R57, L56, N10, S11, T12 and R19 in mode II are analogous, respectively, to the positions of R23, L22, N44, S45, T46 and R53 in mode I. Panel (C) shows structures of the Coh-Doc complex given by the PDB:1OHZ (Coh in blue, Doc in red) and PDB:2CCL (Coh in cyan, Doc in orange). The Coh structures are superimposed. The spheres represent the dockerin-bound Ca^{2+} ions. The Doc helices are not aligned, indicating the existence of the two different binding modes. Panel D shows the PDB structure 1OHZ. Coh and Doc are shown in blue and red, respectively. The black line shows the derived axis of the symmetry, denoted here as the Z-axis. The sense of the rotation is indicated at the top.

phenomenon does exist, but in each mode there are two kinds of the stretching trajectories. The two effects are mixed in the SMFS experiment since it is done without any control of the starting situation.

Thus we consider the existing evidence for the dual binding mode to be inconclusive and requiring further studies. In this paper, we provide another theoretical approach to the binding problem by analyzing the energy landscapes involved by using an empirical model known as FoldX[41,42]. In particular, we investigate the influence

complex	Coh	Doc	Definition	Index k	ΔG_{min} [kcal/mol] mode I	ΔG_{min} [kcal/mol] mode II
$C_I d_I$	1OHZ	1OHZ	1OHZ	1	−38.2	−28.7
$C_I d_{II}$	1OHZ	2CCL (A45S,A46T)	Doc from 2CCL with reverse mutations	2	−35.1	−31.2
$C_{II} d_I$	2CCL	1OHZ	Doc from 1OHZ	3	−35.6	−30.7
$C_{II} d_{II}$	2CCL	2CCL (A45S,A46T)	Doc from 2CCL with reverse mutations	4	−37.2	−35.5
$C_I d_I^*$	1OHZ	1OHZ (S45A,T46A)	1OHZ with two mutations in Doc	5	−36.0	−28.5
$C_I d_{II}^*$	1OHZ	2CCL	Doc from 2CCL	6	−35.0	−31.9
$C_{II} d_I^*$	2CCL	1OHZ (S45A,T46A)	Doc from 1OHZ with two mutations	7	−34.1	−29.2
$C_{II} d_{II}^*$	2CCL	2CCL	2CCL	8	−31.0	−34.0

Table 1. The Coh-Doc complexes without the terminal tails in the Doc. C_I and C_{II} denote the Coh structures derived from PDB:1OHZ and PDB:2CCL, respectively. Both C_I and C_{II} comprise residues with sequential numbers from 5 to 153. d_I and d_{II} denote the Doc structures derived from PDB:1OHZ and PDB:2CCL, respectively; they both comprise residues with sequential numbers from 1 to 56. The asterisk denotes the two single-point mutations (S45A and T46A) in Doc. The fifth column provides the index, k, associated with the complex. The sixth column lists the values of the minimal free energy, G_{min}, corresponding to binding mode I whereas the seventh column–to mode II.

of the two-site mutation (S45A and T46A) on the energy of Coh-Doc interactions in the two binding modes. We also perform comparative all-atom simulations.

Recent experiments on the ScaA Doc module from *R. flavefaciens* show that physical interactions between an N-terminal Trp and a C-terminal Pro in Doc confer stability on it[43]. However, neither PDB:1OHZ nor PDB:2CCL, encompasses the corresponding residues at the N- and C-terminus of the *C. thermocellum* Doc. Both of these Coh-Doc structures lack the the N- and C-terminal tails in Doc. We modeled these terminal segments and, interestingly, discovered that they indeed have a substantial influence on the energy landscapes of the Coh-Doc complex. In the presence of the full-length N- and C-terminal tails in the WT Doc, the free energy of binding in modes I and II are found to be comparable, which is consistent with the fact that–despite widespread efforts–no crystal structure of the full-length *C. thermocellum* Coh-Doc complex has been ever resolved. In the absence of the terminal tails in the WT Doc, on the other hand, mode I is found to have a substantially lower free energy than mode II, which is consistent with the Coh-Doc structure PDB:1OHZ. The two-site mutation–independent of the presence or absence of the terminal tails in Doc–is found to shift the equilibrium significantly towards mode II, as exemplified by the crystal structure PDB:2CCL.

In addition to the free energy calculations, we also simulate an AFM-like stretching of the tail-extended Coh-Doc complex using a structure-based coarse-grained model[44–49]. Our particular implementation of the structure-based model is related conceptually to several other approaches[50–53]. It has been selected optimally out of 62 variants of structure-based models considered in ref.[54] by making comparisons to experimental data on stretching. For a given binding mode, we still observe two types of force-extension patterns which reflect different scenarios of structural rearrangements during the stretching. Interestingly, the two terminal tails in Doc make the Coh-Doc complex more stable mechanically. In particular, the presence of the tails in Doc affects the probabilities of observing the different types of the force-extension patterns.

Methods

Structural models of Coh and Doc.

The Docs present in the PDB:1OHZ and PDB:2CCL correspond to the C-terminal docking domain of endo-1,4-β-xylanase Y from *C. thermocellum*. According to the Uniprot database, entry P51584, this Doc contains 69 amino acid residues. In keeping with the convention used in the entries PDB:1OHZ and PDB:2CCL, these residues are numbered from −5 to 64. However, the Doc structures given in PDB:1OHZ and PDB:2CCL have different sequential lengths: PDB:1OHZ contains the Doc residues with numbers from 1 to 56 whereas PDB:2CCL–residues from -2 to 59. When analyzing the Docs without the terminal tails, we truncate the Doc structure given in PDB:2CCL to retain only the residues with the sequential numbers ranging from 1 to 56. Otherwise, the residues at sites -5, -4, -3 are P, P, and V respectively; at sites 60, 61, 62, 63, and 64–D, K, F, P, and V respectively. Also the sequences of Cohs given in the two PDB entries have different lengths: PDB:1OHZ contains the Coh residues numbered from 5 to 144 whereas PDB:2CCL comprises the Coh residues from 5 to 153. We thus discard the Coh residues 145 through 153 from PDB:2CCL (when dealing with or without the tails in Doc) so that both Coh structures have identical sequences in our analysis.

We consider eight Coh-Doc complexes that are defined in Table 1. Here, we introduce the following notation: The capital letter C denotes the structure of Coh with the amino acid sequence comprising residue numbers from 5 to 144. The small letter d denotes the truncated Doc with the sequence comprising residue numbers 1 through 56. The asterisk denotes the Doc sequence with the two single-point mutations, i.e., S45A and T46A. The subscripts I and II indicate that the protein structures taken as an initial state for energy calculations are PDB:1OHZ and PDB:2CCL, respectively. This convention implies, for example, that the arrangement denoted by $C_I d_I$ corresponds to the Coh-Doc corresponding to PDB:1OHZ, where the WT Doc makes a complex with Coh in mode I. Similarly, in the arrangement $C_{II} d_I$ the Coh of 2CCL is bound to the Doc of PDB:1OHZ.

We also study systems in which the Doc of PDB:1OHZ undergoes the two-site mutation to make it sequentially identical to the Doc of PDB:2CCL. Such Doc is denoted in Table 1 by d_I^*. In addition, we consider systems

complex	Coh	Doc	Server	Template	Index k	ΔG_{min} [kcal/mol] mode I	ΔG_{min} [kcal/mol] mode II
C_Id_I	1OHZ	1OHZ	Swiss Model	4DH2	9	−25.2	−14.3
$C_{II}d_{II}$	2CCL	2CCL (A45S,A46T)	Swiss Model	4DH2	10	−26.4	−28.4
C_Id_I	1OHZ	1OHZ	iTASER	4DH2	11	−29.5	−18.4
C_Id_I	1OHZ	1OHZ	iTASER	1OHZ	12	−23.6	−15.5
$C_{II}d_{II}$	2CCL	2CCL (A45S,A46T)	iTASER	4DH2	13	−28.0	−28.7
$C_{II}d_{II}$	2CCL	2CCL (A45S,A46T)	iTASER	2CCL	14	−26.0	−29.9
$C_ID_I^*$	1OHZ	1OHZ	Swiss Model	4DH2	15	−24.2	−13.5
$C_{II}D_{II}^*$	2CCL	2CCL	Swiss Model	4DH2	16	−29.6	−29.1
$C_ID_I^*$	1OHZ	1OHZ	iTASER	4DH2	17	−28.0	−18.5
$C_ID_I^*$	1OHZ	1OHZ	iTASER	1OHZ	18	−22.8	−13.0
$C_{II}D_{II}^*$	2CCL	2CCL	iTASER	4DH2	19	−29.0	−30.4
$C_{II}D_{II}^*$	2CCL	2CCL	iTASER	2CCL	20	−26.7	−32.1

Table 2. The Coh-Doc complexes comprising the full-length Doc. As in Table 1, C_I and C_{II} denote the Coh structures derived from PDB:1OHZ and PDB:2CCL, respectively. D_I and D_{II} denote the full-length Doc structures associated with PDB:1OHZ and PDB:2CCL, respectively. They both comprise residues with sequential numbers from -5 to 64. The asterisk denotes the two single-point mutations (S45A and T46A) in Doc. The PDB-unavailable tail parts of the structures are derived either by using the Swiss Model or iTaser server, as listed in the fourth column. The fifth column indicates the PDB structure code of the templates used. The sixth column provides the index, k, associated with the complex. The seventh column lists the values of the minimal free energy corresponding to mode I of binding whereas the eighth column–to mode II.

with the Doc derived from PDB:2CCL in which the original mutation is reverted by implementing the substitutions A45S and A46T. Such Doc is denoted in Table 1 as d_{II}. Therefore, C_Id_{II} corresponds to the Coh-Doc complex in which the Coh of PDB:1OHZ is bound to the Doc of PDB:2CCL in which the reverse mutations A45S and A46T have been made. Finally, the arrangement $C_{II}d_{II}^*$ corresponds to PDB:2CCL, where the mutated Doc makes a complex with Coh in mode II.

Since neither PDB:1OHZ nor PDB:2CCL contains the N- and C-terminal tails of the Docs, we model these terminal segments in two ways: by using the Swiss Model[55] and iTASER[56]. Because of the uncertainty involved in protein structure modeling, in addition to using the two methods, we also consider three different PDB entries as template structures. They are listed in Table 2. The twelve resulting models of the Coh-Doc complex with the tails are summarized in Table 2. We use the convention in which the capital letter D denotes the full-length Doc which comprises the amino acid residues with the sequential numbers from −5 to 64. As before, the asterisk denotes the Doc sequence with the two single-point mutations.

As we explain in the subsequent paragraph, it is possible to determine a symmetry axis, Z, such that a rotation of Doc around Z transforms the Coh-Doc structure between the two binding modes. The geometry involved is illustrated in panel D of Fig. 2. Here, φ is the rotation angle around the Z-axis and the coordinates of Coh-Doc at $Z = 0$ and $\varphi = 0$ correspond to PDB:1OHZ. Similarly $Z \approx 0$ and $\varphi \approx \pi$ correspond to PDB:2CCL. We use the convention in which positive values of Z correspond to shifting the two molecules closer together and negative values–to shifting them further away. The rotations and forward shifts are not implemented when prohibited by steric constraints.

The secondary structure of Doc consist of three helices: $\alpha_1{}'$ constituted by the residues with numbers from 11 to 23, $\alpha_2{}'$ formed by the residues ranging from 28 to 36, and $\alpha_3{}'$ comprising residues from 45 to 56. Here, in keeping with the convention used in refs[40,57] the primed and unprimed symbols indicate the secondary structures belonging to Doc and Coh, respectively. In each of the binding modes only the first and third of these helices couple to Coh, which implies that the symmetry axis can be determined based only on $\alpha_1{}'$ and $\alpha_1{}'$. To determine the symmetry axis we use the following procedure: (i) We superimpose the Cohs of PDB:1OHZ and PDB:2CCL. (ii) We introduce a set of vectors formed by pairs of equivalent α-C atoms in helices $\alpha_1{}'$ and $\alpha_3{}'$ as described in PDB:1OHZ and PDB:2CCL. (iii) We rotate the coordinate system in such a way that the Z-components of these vectors are brought closer to zero. (iv) We iterate step (iii) to bring the Z-components of these vectors to zero. The Z-axis of the final, rotated coordinate system gives the symmetry axis. We find that switching from mode I to mode II requires a rotation around the Z-axis by $\varphi = 174°$. Rotating by 360° brings the system back to mode I.

Free energy calculations using FoldX. The purpose of our calculations is to determine the free energy of the Coh-Doc system as a function of parameters Z and φ, which describe the location of Doc relative to Coh, as depicted in panel D of Fig. 2. One way to determine the energy landscape associated with the protein complex would be by using all-atom simulations and averaging over classes of conformations. However, this procedure would be way too costly numerically when repeating it for various values of Z and φ, even when adopting an implicit solvent approach. Instead, we use an empirical force field known as FoldX[41,42] which has been designed primarily for predicting free energy differences between a WT protein and its mutant. Here, however, the distinct structures correspond to different positions of Doc relative to Coh. We predict the structures and their free energies as a function of Z and φ. The prediction also applies to the Coh-Doc complexes with the mutated (S45A and T46A) or reverse-mutated (A45S and A46T) Doc. We consider the systems both with and without the tails in the

Doc. The effects of the dockerin-bound calcium ions are included in the calculations. FoldX employs an energy function that consists of ten terms that are listed in Supplementary Information (SI).

We perform the free energy calculations for two temperatures: $T = 298$ K, which is the FoldX default parameter, and $T = 308$ K, i.e., the maximum temperature for which the FoldX energy function has been parametrized. The free energy is optimized with respect to the side-chain conformations while the backbone atoms are kept at fixed positions. When not considering the sequence tails in Doc, the initial structures for the free energy minimization are as in the PDB structure files, as listed in Table 1. Otherwise, the structures are predicted by iTaser or the Swiss Model, as specified in Table 2. In all cases, the dockerin-bound calcium ions are included in the structural models and taken into account in the FoldX energy calculations[58].

The resulting ΔG, minimized with respect to the side-chain orientation, is the free energy subject to the constraints on the positions of the backbone atoms of Coh and Doc. Therefore, ΔG depends not only on the coordinates Z and φ but also on the structural model taken as input for the calculations. In particular, selecting the lowest free energy is subject to a substantial uncertainty and, perhaps more importantly, does not provide a proper estimate of the binding strenght because it does not involve any information about the lateral extension of the minimal basins.

In order to estimate the free energy of binding in a more robust way, we perform a second stage of calculations in which ΔG determined for various complexes corresponding to the same sequence (mutated or not) serve as an input. In the case of Docs with tails, we also include the various of determining the structure. Specifically, we define the free energy of Coh-Doc binding in mode I as

$$F_I = -k_BT\ \log\left[\sum_k\sum_Z\sum_{-\pi/2<\varphi<\pi/2}\exp(-\Delta G_k(Z,\ \varphi)/k_BT)\theta(E_c - \Delta G_k)\right]. \tag{1}$$

Here, the sum over index k corresponds to averaging over the input structures listed in Tables 1 or 2. The values of k are written in the last columns of these tables. Accordingly, $\Delta G_k(Z,\ \varphi)$ denotes $\Delta G(Z,\ \varphi)$ computed for the input structure with index k. The unit step function, θ, is defined as follows: $\theta(x) = 1$ for $x > 0$ and $\theta(x) = 0$ for $x < 0$. Thus the configurations $(Z,\ \varphi)$ corresponding to the binding mode I are those with $-\pi/2 < \varphi < \pi/2$ and free energies ΔG_k smaller than a cut-off value E_c. As we show in the Results section, E_I does not depend on the choice of the cut-off as long as E_c is larger than about -25 kcal/mol.

By analogy, the free energy of Coh-Doc binding in mode II is

$$F_{II} = -k_BT\ \log\left[\sum_k\sum_Z\sum_{\pi/2<\varphi<3\pi/2}\exp(-\Delta G_k(Z,\ \varphi)/k_BT)\theta(E_c - \Delta G_k)\right]. \tag{2}$$

The configurations $(Z,\ \varphi)$ corresponding to the binding mode II are those with $\pi/2 < \varphi < 3\pi/2$ and energies ΔG_k smaller than a cut-off value E_c.

If the equilibrium probability of finding the Coh-Doc complex in mode I is denoted by p_I and in mode II by p_{II} then

$$p_I/p_{II} = \exp[-(F_I - F_{II})/k_BT]. \tag{3}$$

The dual binding occures if the probabilities p_I and p_{II} are of the same order of magnitude.

Free energy of binding by using all-atom simulations.

The employment of the empirical models described above allows for the elucidation of the free energy landscape through the usage of the Z and φ variables. It also allows for a considerable probing of the side-chain conformations. However, the drawback involved is the lack of the flexibility of the backbone. It is thus worthwhile to compare the binding energies to those obtained through all-atom molecular dynamics (MD) simulations.

To this end, we used the Amber 14 package[59] with the AMBER force field 99SB[60] and the TIP3P model for the molecules of water[61]. The Newton equations of motion were integrated by using the leapfrog algorithm with the time step of 2 fs. All bonds with the hydrogen atoms were constrained through the SHAKE algorithm[62]. The T was maintained around 300 K by adding terms corresponding to the Langevin dynamics[63] with the collision frequency of 2 ps^{-1}. A 10 Å cut-off was applied to all non-bonded interactions. The Particle Mesh Ewald method[64] was used to treat the long-range electrostatics.

The Coh-Doc complex (shown in Fig. S1 in SI) was solvated in a truncated octahedron box of water that is large enough to avoid interactions between the protein complex and its images in the adjacent cells. The system was neutralized by adding Na$^+$ ions. The energy of the system was then minimized in three stages: the first minimization was done with all atoms of the protein complex being restrained to let the water molecules move to empty places around the proteins; in the second minimization stage, only the backbone heavy atoms are restrained; and in the last stage, no restraints were applied. Each minimization stage involved 2500 steepest descent steps that removed steric clashes and then 2500 conjugate gradient steps to achieve quick convergence. After the energy minimization, the system was heated up from 0 to 300 K in 2.5×10^4 steps with weak restrains applied to the atoms of the Coh-Doc complex. This was followed by 2.5×10^4 relaxation steps at $T = 300$ K (with the restrains kept in place), which brought the system to the density of 1 g/cm^3, and then by 2.5×10^5 steps of equilibration without any restrains. Both the energy minimization and the heating-up simulations were performed at constant volume. The subsequent relaxation and equilibration simulations were performed at constant pressure $p = 1$ atm.

In the production runs, pressure was also maintained at $p = 1$ atm. The coordinates of all atoms of the Coh-Doc complex were recorded every 10 ps. Five independent trajectories of 200 ns were generated for each of the systems under study. For the recorded conformations of the complex, the root-mean-square distance (RMSD) from the native structure was computed by considering only the backbone atoms. An example of the dependence of RMSD on time is shown in Fig. S2 in SI. Based on the RMSD plots we selected, for each of the trajectories independently, a time interval in which RMSD attained a steady state and fluctuations in RMSD did not exceed 2 Å. The Coh-Doc conformations recorded in these time intervals were taken as input for implicit-solvent free-energy calculations, which were performed within the framework of the Molecular Mechanics/Poisson Boltzmann Surface Area (MM/PBSA) method.

The MM/PBSA method is based on calculating the free energy difference, ΔG_{bind}, between the bound and unbound states

$$\Delta G_{bind} = G_{Coh-Doc} - G_{Coh} - G_{Doc} = \Delta E_{ele} + \Delta E_{vdW} + \Delta G_{PB} + \Delta G_{SUR} - T\Delta S.$$

Here, ΔE_{ele} and ΔE_{vdW} denote the electrostatic and van der Waals contributions. These energy contributions are the same as in the AMBER force field used in the MD simulations. ΔG_{PB} and ΔG_{SUR} are the polar and non-polar solvation energy terms. The entropic contribution, $T\Delta S$, is estimated by the normal mode approximation method using the *mmpbsa_py_nabnmode* program implemented in the AMBER package. The solvation energy was calculated by *pbsa*, which is also included in the AMBER package. The polar term, ΔG_{PB}, was obtained by solving linearized Poisson-Boltzmann equation numerically. The non-polar term is defined by $\Delta G_{SUR} = \alpha SASA + \beta$, where SASA is the solvent-accessible surface area that was calculated by the LCPO method[65]. The regression coeficients α and β are set to 0.005 kcal mol^{-1} Å^{-2} and 0, respectively. For a given MD trajectory, ΔG_{bind} is averaged over the selected time interval. The resulting values of time-averaged ΔG_{bind} are summarized in Tables S1–S5 in SI.

Structure-based coarse-grained simulations. We use a structure-based coarse-grained model[44–49] in which amino acid residues are represented by single beads centered on their α-C atoms. The beads are tethered together into chains by strong harmonic potentials with the spring constant $k_{\text{bond}} = 100\,\varepsilon/\text{Å}^2$, where ε is the depth of the potential well associated with the native contacts, which serves as the basic energy scale in our model. We assume $\varepsilon = (110 \pm 30)$ pN Å, as determined by benchmarking against experimental results for 38 proteins[45]. The native contacts are identified using an overlap criterion[49] applied to the coordinates of all heavy atoms in the native structure. Here, the van der Waals radii of the heavy atoms are taken from ref.[66] The effective spheres associated with the atoms, when checking for the overlaps, have radii which are 1.24 times larger[67] (this factor corresponds to the point of inflection in the Lennard-Jones potential). In addition, the amino acid pairs that are very close sequentially, $(i, i+1)$ and $(i, i+2)$, are excluded from the contact map. Examples of contact maps for Doc and Coh-Doc are shown, respectively, in Figs S3 and S4 in SI.

The interactions within the native contacts are described by the Lennard-Jones potential

$$V^{NAT}(r_{ij}) = 4\varepsilon\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]$$

Here, r_{ij} denotes the distance between residue beads i and j. The parameters σ_{ij} are chosen so that each contact in the native structure is stabilized at the minimum of the Lennard-Jones potential. The contacts between the proteins are treated in the same manner as the contacts within the proteins as both sets are dominated by hydrogen bonds. The contacts between the dockerin and the Ca^{2+} ions are determined on the basis of the overlap criterion[49] with the van der Waals radius of Ca^{2+} being 1.53 Å, as in the AMBER force-field[68]. The interactions between the residues that do not form native contacts are purely repulsive and given by the truncated and shifted Lennard-Jones potential corresponding to $\sigma_{ij} = r_0/\sqrt[6]{2}$ with $r_0 = 4$ Å. The energy function comprises also harmonic terms that favor the native values of local chiralities in each amino acid chain[69].

The solvent is implicit and the system evolves in time according to the Langevin dynamics. The overall force acting on a particular bead i is a sum of three terms: (i) the direct force $\vec{F}_i$ that derives from all the potential energy terms, (ii) the damping force that is proportional to the velocity of the bead, and (iii) the random force, $\vec{\Gamma}_i$, that represents thermal noise. The corresponding equations of motion, $m\frac{d^2\vec{r}_i}{dt^2} = \vec{F}_i - \gamma\frac{d\vec{r}_i}{dt} + \vec{\Gamma}_i$, are solved by the fifth-order predictor-corrector algorithm with the time step of 0.005 τ. Here, γ is the damping coefficient, and all beads are assumed to have the same mass m. The dispersion of the thermal noise is given by $\sqrt{2\gamma k_B T}$, where k_B is the Boltzmann constant. The damping coefficient is set to $\gamma = 2m/\tau$. This value corresponds to the overdamped case–practically Brownian dynamics–and the characteristic time scale, τ, is of the order of 1 ns, as argued in refs[70,71].

Thermal stability of dockerin. We use the structure-based coarse-grained simulations to investigate the thermal stability of the Docs with and without the terminal tails. To this end, we compute the probability of finding the Doc in the native state, P_0, as a function of T. The definition of P_0 involves counting the conformations in which all native contacts are present. The native state probability P_0 is thus different from the average fraction of native contacts, Q, which is often used to characterize the deviation from the native state.

To compute $P_0(T)$ and $Q(T)$, we perform MD simulations for T ranging from $0.02\,\varepsilon/k_B$ to $1.16\,\varepsilon/k_B$. Each MD simulation is preceded by a $10^4\tau$ equilibration run and gives $3 \times 10^5\,\tau$ of dynamics. At any given T, we run 101 independent trajectories. P_0 is then determined as an average over the simulation time and over the trajectories. We perform the simulations for the Docs with the terminal tails (index $k = 1$) and without the terminal tails (index $k = 9$, 11 and 12).

Another way to assess thermal stability of proteins is to simulate their unfolding at elevated temperatures. The unfolding simulations start at the native state and finish when all nonlocal contacts get broken, which defines the unfolding time t_{unf}. Specifically, the nonlocality refers to the sequential distance $|i - j| > 4$. At any given T between $0.9\,\varepsilon/k_B$ and $2.1\,\varepsilon/k_B$, we run 301 independent trajectories of $10^5\,\tau$ each and, thus, obtain 301 values of t_{unf}. As the characteristic unfolding time, t_u, we take the median of the distribution of unfolding times t_{unf}. We perform the simulations for the Doc with the tails ($k = 1$) and without the tails ($k = 9$,11 and 12).

Stretching simulations of the Coh-Doc complex. We use the structure-based coarse-grained simulations to investigate also the mechanical stability of the Coh-Doc complex. Stretching of the Coh-Doc complex is implemented by attaching two harmonic springs to the N-terminal amino acids of Coh and Doc. (Other ways of pulling are discuss in ref.[40]). The N-terminus of Doc is denoted as N' and the C-terminus as C'. One of the springs is fixed in space and the other one is moved at a constant speed, v_p, so that the distance it travels in time t is $d = v_p t$. The force constant of the pulling springs is taken as $K = 0.12\,\varepsilon/\mathrm{\AA}^2$, which corresponds to about 1 pN/nm and is close to the elasticity of typical AFM cantilevers[45]. All pulling simulations are performed for $T = 0.3\ \varepsilon/k_B$, which is near-optimal in folding kinetics and is of the order of room temperature.

In our simulations, the response force F acting on the pulling spring is measured and averaged over time periods that correspond to the spring displacements of 0.5 Å[45]. The F-d curves may come with several peaks, and the height of the largest of them is denoted by $Fmax$. As in ref.[40] we perform simulations for the pulling speed $v_p = 5 \times 10^{-5}\,\mathrm{\AA}/\tau \approx 5\,\mathrm{nm/ms}$ which is close to the experimental speeds.

All of the pulling simulations start from the native state. In the course of the simulations, the breaking and re-formation of native contacts is followed in time. The native contact between residues i and j is considered broken if the inter-residue distance r_{ij} exceeds a cutoff length of $1.5\ \sigma_{ij}$. Due to thermal fluctuations, the broken contacts may get re-established. To characterize the unfolding and dissociation patterns, we record the spring displacements at which the native contacts break for the last time.

Data availability. All data generated or analysed during this study are included in this published article (and its Supplementary Information files).

Results

Energy landscapes of Coh-Doc interactions. *Docs without the tails.* Figure 3 illustrates the FoldX-derived free-energy landscapes for systems $C_I d_I$ (panels A, B and C) and $C_I d_I^*$ (panels D, E and F). The panels A and D are color-maps showing ΔG_k for $k = 1$ and $k = 5$, respectively, as a function of Z and φ. (Here, ΔG is minimized with respect to the side-chain orientations for the input structures with indices $k = 1$ and $k = 5$). The value of ΔG is indicated by colors according to the scale bar on the right-hand side. The regions with smallest/largest values of ΔG are marked in blue/red. We note that there are two pronounced local minima in ΔG. The first one is located around $\varphi = 0$ (binding mode I), whereas second one is around $\varphi = \pi$ (binding mode II). More precisely, for system $C_I d_I$, the two minima are located at $\varphi = 3°$ and $\varphi = 173°$. They have different depths, −38.2 and −28.7 kcal/mol, and the parameter region corresponding to mode I has a larger area in the (Z, φ) plane. For system $C_I d_I^*$, the two minima are located at $\varphi = 3°$ and $\varphi = 173°$. Their depths are −35.2 and −28.1 kcal/mol.

Panels B and E in Fig. 3 show the Z-dependence of ΔG for two fixed values of φ, i.e., for $\varphi = 3°$, corresponding to mode I (upper subpanels), and for $\varphi = 173°$ (panel B) or $\varphi = 172°$ (panel D), corresponding to mode II (lower subpanels). The data points are fitted to parabolas around the minima to help estimate their depth. Panels C and F in Fig. 3 show the φ-dependence at Z equal to the value at the deepest minimum, Z_{min}. The solid lines are curves that are fitted to the data points.

The values of ΔG_{min}, corresponding to the two binding modes for the eight complexes under study, are listed in Table 1, columns number five and six. For a given binding mode (I or II) and a given sequence (WT or two-site mutant), the values of ΔG_{min} depend on k because the positions of the backbone atoms are not identical in the structural models that are taken as input for the FoldX calculations. For the systems comprising the WT Doc, for which $k = 1, \ldots, 4$, the values of ΔG_{min} are observed to be scattered between −38.2 and −35.1 kcal/mol in mode I and between −35.5 and −28.7 kcal/mol in mode II. For the systems comprising the mutated Doc, with $k = 5, \ldots, 8$, they are seen to be between −36.0 and −31.0 kcal/mol in mode I and between −34.0 and −28.5 kcal/mol in mode II.

The probability that Doc binds Coh in mode I or II is not determined only by the corresponding values of ΔG_{min} but also by the widths of the two valleys in the Coh-Doc energy landscape. For this reason we consider the binding free energies F_I and F_{II} as well as the binding probabilities p_I and p_{II} as defined by Eqs (1–3) with the condition that $p_I + p_{II} = 1$. The binding free energies F_I and F_{II} for the WT sequence are given by Eqs (1 and 2) with $k = 1, \ldots, 4$. By analogy, the binding free energies F_I and F_{II} for the two-site mutant are given by Eqs (1 and 2) with $k = 5, \ldots, 8$. Figure 4A shows F_I and F_{II} as a function of the energy cut-off E_c for the WT case. We observe that F_I and F_{II} saturate above a threshold of about −25 kcal/mol, and we take the saturation values as estimates of the binding energies (from now on F_I and F_{II} denote the saturation values).

For the WT system, we obtain $F_I = -38.5$ kcal/mol and $F_{II} = -35.3$ kcal/mol which yields $p_I/p_{II} \approx 200$, indicating a strong preference for binding in mode I. On the other hand, for the mutated system we get $F_I = -35.9$ kcal/mol and $F_{II} = -34.0$ kcal/mol, as can be seen in Fig. 4B. According to Eq. (3), the probability of binding mode I to

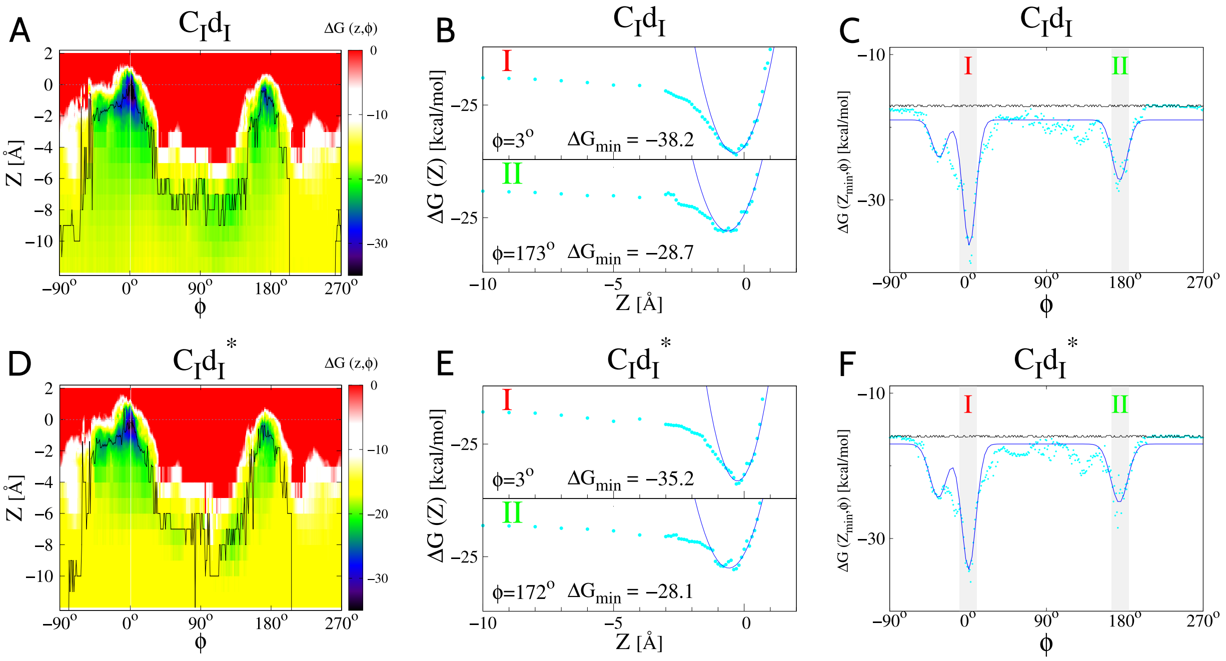

Figure 3. Results of the FoldX-based free-energy calculations for system C_ID_I (panels A,B and C) defined in Table 1 for index $k = 1$, and for system $C_Id_I^*$ (panels D–F) corresponding to index $k = 5$ in Table 1. (**A**) ΔG as a function of coordinates Z and φ. The values of ΔG are indicated by the color scale shown on the right hand side. (**B**) ΔG as a function of Z for angles $\varphi = 3°$ (top sub-panel, mode I) and $\varphi = 173°$ (bottom sub-panel, mode II) at which $\Delta G(Z, \varphi)$ is found to take the minimal values, $\Delta G_{min} = -38.2$ kcal/mol and $\Delta G_{min} = -28.7$ kcal/mol, respectively. The data points in cyan show the results of the FoldX-based calculations. The solid blue lines correspond to a harmonic approximation at the minimum of $\Delta G(Z)$. (**C**) ΔG minimized with respect to Z and plotted as a function of φ. The data points in cyan show the results of the FoldX-based calculations. The solid blue line represents a fit to the data points. This line is a to guide the eye. The fitting function used here involves three Gaussians. Panels (D–F) are analogous to panels (A–C), respectively.

the probability of binding mode II is $p_I/p_{II} \approx 20$. Thus mode I is still more preferred than mode II, but the equilibrium shifts towards mode II. It should be noted, however, that FoldX does not strictly endorse mode II for the mutated sequence that is evidenced by PDB:2CCL.

The MM/PBSA calculations yield the following results: $\Delta G_{bind} = -36.3 \pm 2.9$ kcal/mol for the binding of the WT Doc to the Coh in mode I, i.e., for the Coh-Doc system with index $k = 1$, and $\Delta G_{bind} = -34.5 \pm 2.9$ kcal/mol for the binding of the mutated Doc the Coh in mode II, i.e., for the Coh-Doc system with index $k = 8$. These binding energies are comparable within the statistical error. They are also consistent with the binding free energies obtained from the FoldX calculations.

Docs with the tails. Figure 5 illustrates the FoldX-derived free-energy landscapes for systems $C_{II}D_{II}$ ($k = 10$; panels A, B and C) and $C_{II}D_{II}^*$ ($k = 16$; panels D, E and F). Overall, they resemble the energy landscapes obtained for systems C_ID_I and $C_Id_I^*$, in which the Docs lack the terminal tails (compare with Fig. 3). The noticeable differences are slight shifts in the minimum-energy angles φ. In the case of system $C_{II}D_{II}$, ΔG_{min} favors binding mode II by 2 kcal/mol. In the case of system $C_{II}D_{II}^*$, however, ΔG_{min} for mode I is smaller by 0.5 kcal/mol than ΔG_{min} for mode II. The values of ΔG_{min} corresponding to modes I and II for k between 9 and 20 are listed in Table 2.

The binding free energies for the WT system ($k = 9, \ldots, 14$) are $F_I = -29.2$ kcal/mol and $F_{II} = -30.2$ kcal/mol, as can be seen in Fig. 4C. Thus mode II has a somewhat lower free energy than mode I: F_I–F_{II}, is only of order of 0.5 kcal/mol. This implies that $p_I/p_{II} \approx 0.4$, which suggests that the WT full-length Doc should bind its partner in modes I and II with comparable probabilities. This result rationalizes the fact that no crystal structure of the full-length Coh-Doc WT complex has ever been solved. When the sequence is truncated, mode I is preferred, and that is why it was crystallized as PDB:1OHZ. It should be noted that none of our predicted models indicates any presence of flexible or disordered parts in the tails. Such a flexibility would provide the usual mechanism of not having a well defined structure.

For the system with the double-site mutation ($k = 15, \ldots, 20$), we get $F_I = -29.6$ kcal/mol and $F_{II} = -32.4$ kcal/mol, as can be seen in Fig. 4D, which implies $p_I/p_{II} \approx 0.01$ according to Eq. (3). This result means that the two-site mutation shifts the equilibrium towards mode II, from $p_I/p_{II} \approx 0.4$ to $p_I/p_{II} \approx 0.01$. Thus mode II is clearly dominating in this case and probably could be crystallized, if attempted.

We also use the MM/PBSA method to compute the binding free energy between the Coh and the Doc with the tails. For the system in which the Coh binds in mode I to the WT Doc, we obtain $\Delta G_{bind} = -39.4 \pm 7.2$ kcal/

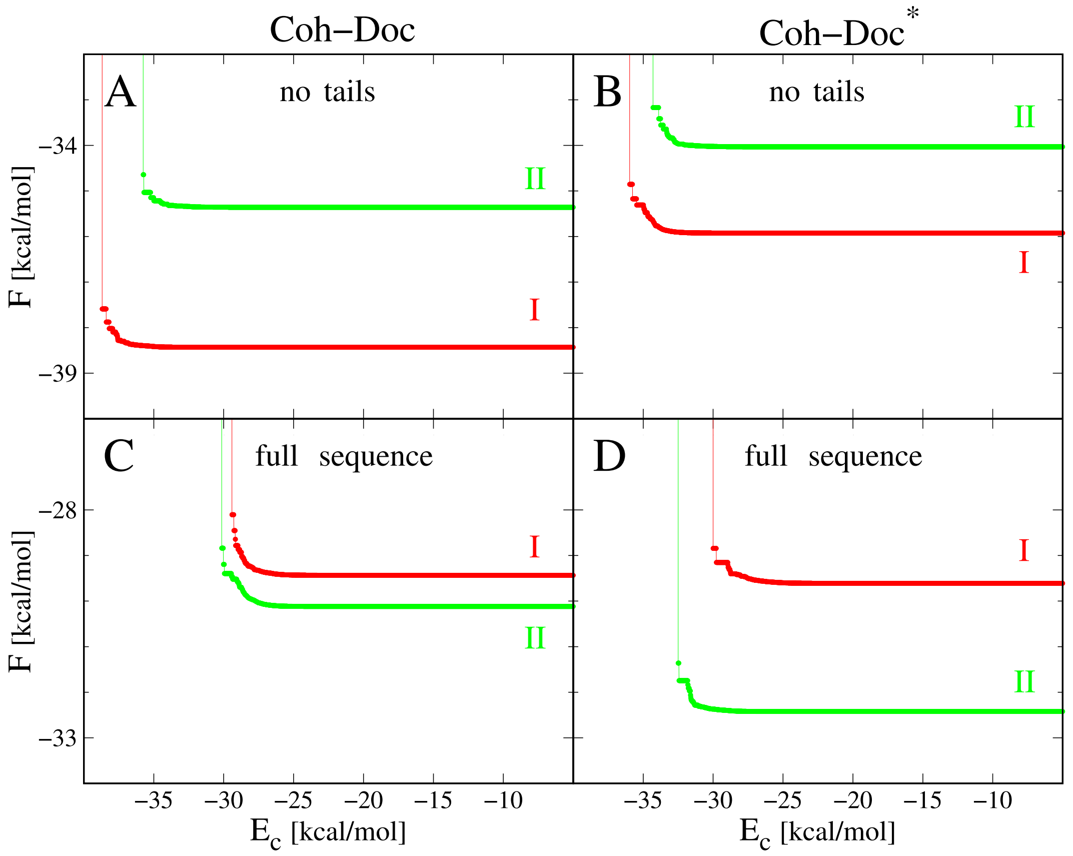

Figure 4. Free energies of binding in mode I (red) and II (green) as functions of the energy cut-off, E_c, which is used to compute F_I and F_{II} according to Eqs (1) and (2). The top panels, A and B, correspond to the Doc without the tails (sequences comprising residues 1 through 56). The bottom panels, (C and D), correspond to the full-length Doc (sequences comprising residues −5 through 56). Four cases are considered: (**A**) Coh-Doc in which Doc is WT and without the tails. Here, F_I and F_{II} are computed by taking $k = 1, \ldots, 4$. (**B**) Coh-Doc* in which Doc* is the two-point mutant without the tails. The binding energies are computed by taking $k = 5, \ldots, 8$ in this case. (**C**) Coh-Doc in which Doc is WT and full-length, i.e., comprising residues −5 through 56. This case corresponds to averaging over $k = 9, \ldots, 14$ in Eqs (1) and (2). (**D**) Coh-Doc* in which Doc* is the two-point mutant with the tails, i.e., comprising residues −5 through 56. Structural models with $k = 15, \ldots, 20$ have been used for calculating F_I and F_{II} in this case.

mol. For the system in which the Coh binds in mode II to the mutated Doc, the MM/PBSA method yields $\Delta G_{bind} = -43.6 \pm 6.0$ kcal/mol. The difference in the binding free energies is only about 4 kcal/mol and, thus, is comparable to the statistical error on ΔG_{bind}. Therefore, one can not state with confidence which of the two systems is bound more tightly. However, the difference between F_{II} for the full-length Coh-Doc* complex and F_I for the full-length Coh-Doc complex, as obtained from the FoldX calculations, is about 3 kcal/mol, which compares well with the results obtained in the framework of the MM/PBSA method.

For the WT Coh-Doc complex in the binding mode I and II, the MM/PBSA method yields $\Delta G_{bind} = -39.4 \pm 7.2$ kcal/mol and $\Delta G_{bind} = -45.5 \pm 7.3$ kcal/mol, respectively. The difference in these two energies is smaller than the statistical error of the MM/PBSA calculations, which supports the hypothesis of dual binding. We also note that truncation of the terminal tails from the Doc leads to an increase inΔG_{bind} from -39.4 ± 7.2 kcal/mol to -36.3 ± 2.9 kcal/mol for the WT system in mode I, and from -43.6 ± 6.0 kcal/mol to -34.5 ± 2.9 kcal/mol for the mutated system in mode II. These results indicate that the presence of the tails in the Doc enhances the Coh-Doc binding affinity.

The effects of heating. We have used FoldX to perform analogous free-energy calculations for $T = 308$ K, which is 10 K higher than the room temperature considered so far. The results of these calculations are shown in Fig. S5 in SI. For the Coh in complex with the Doc without the tails, we obtain the following free-energy values: $F_I = -30.8$ kcal/mol and $F_{II} = -26.4$ kcal/mol for the WT system (Fig. S5A) and $F_I = -28.6$ kcal/mol and $F_{II} = -26$ kcal/mol for the mutated system (Fig. S5B). Therefore, mode I is seen to be dominating in both cases. For the Coh in complex with the Doc containing the tails, we obtain the following results: $F_I = -20.8$ kcal/mol and $F_{II} = -21.9$ kcal/mol for the WT system (Fig. S5C) and $F_I = -22.8$ kcal/mol and $F_{II} = -23.7$ kcal/mol for the mutated one (Fig. S5D). These results show that mode II is somewhat more favorable than mode I but the

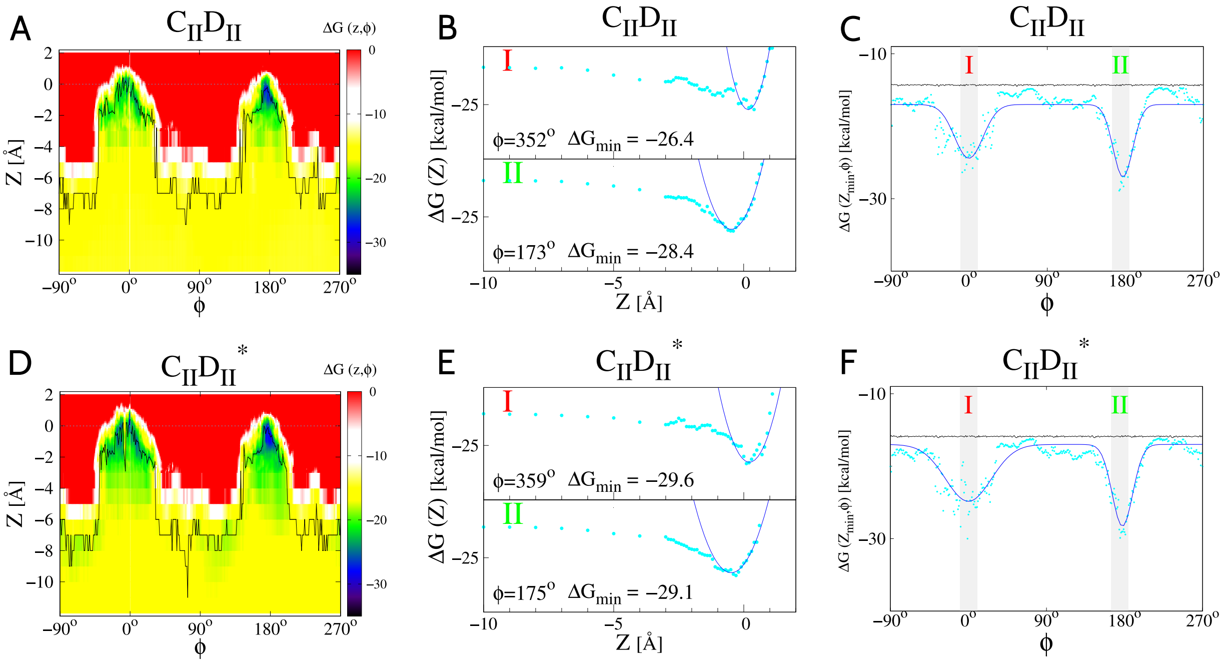

Figure 5. Analogous to Fig. 3 but for system $C_{II}D_{II}$ (panels A–C) corresponding to index $k = 10$ in Table 2, and for system $C_{II}D^{*}_{II}$ (panels D–F) corresponding to index $k = 16$ in Table 2.

free-energy difference between the two modes is only of the order of 1 kcal/mol. We thus conclude that the dual binding persists at $T = 308$ K.

The results of coarse-grained simulations shown in Fig. S6 indicate that the WT Doc with the tails is thermally more stable than without the tails. For instance, the characteristic time of thermal unfolding, t_u, at a given temperature, is longer for the Doc with the tails than for the Doc without the tails. We observe this dependence both for the WT Doc and for the Doc with the two-site mutation. The corresponding simulation data are shown in the upper and lower panels of Fig. S6. In addition, the characteristic temperature at which the average fraction of native contacts, Q, takes the value $Q = \frac{1}{2}$ is higher for the Doc with the tails than for the Doc without the tails (data not shown). However, another characteristic temperature, T_0, at which P_0 crosses $\frac{1}{2}$ does not distinguish between the systems outside of the error bars (data not shown). Nevertheless, these observations, taken together, are consistent with recent fluorescence experiments on the ScaA dockerin from *R. flavefaciens* indicating that interactions between the N- and C-terminal tails in Doc have a significant influence on thermal stability[43].

The differences in the thermal stability can be explained, to some extent, in terms of the contact map shown in Fig. S3. The Doc in PDB:1OHZ has 137 native contacts, according to the overlap criterion that underlies the coarse-grained simulations. By building the terminal tails into the dockerin structure, the number of native contacts is increased by 20 to 23, depending on the procedure used to generate the full-length structure (20 contacts for iTaser with the 1OHZ template, 23 contacts for iTaser with the 4DH2 template, and 22 for Swiss modeller with the 4DH2 template). Similarly, the mutated Doc in PDB:2CCL has 134 native contacts and including the tails adds between 29 and 30 new contacts (30 contacts for iTaser with the 2CCL template, 29 contacts for iTaser with the 4DH2 template, and 30 for Swiss modeller with the 4DH2 template). In both cases, the increased number of the native contacts results in the enhanced thermal stability. These new contacts are not related to the presence of the Ca^{2+} ions.

Stretching of the Coh-Doc complex with tails. We follow the procedure and notation as described in ref.[40] except that now we consider Docs with the tails. We illustrate our findings only for C_ID_I ($k = 9$) and $C_{II}D^{*}_{II}$ ($k = 16$). Examples of the force-displacement (F–d) curves are shown in Fig. 6, the left and right panels respectively.

The figures also show displacements at which particular types of contacts are breaking. For instance, $\beta_1 - \beta_{2,9}$ indicates contacts between β_1–β_2 and β_1–β_9 in Coh, whereas $N' - \alpha_{1,2}'$ means contacts between the N-terminal region on Doc and the first two helices in Doc. The former contacts break first in $C_{II}D^{*}_{II}$ and the latter in C_ID_I.

The interesting observation is that for both binding modes, there are short and long trajectories. For C_ID_I, about 66% of the trajectories (100 trajectories were considered) are of the long type and 34% of the short type. The sequence of the unravelling events in the short trajectory is like in the short trajectory for PDB:1OHZ. In the long trajectory, it is similar to the sequence in the dominant trajectory for PDB:1OHZ (the middle panel in Fig. 4 in ref.[40]).

For $C_{II}D^{*}_{II}$ most of the trajectories are short. Only 3% were long, but for the corresponding system without the tails, we have not observed any long trajectories. Both kinds of the trajectories start with unravelling of $\beta_1 - \beta_{2,9}$,

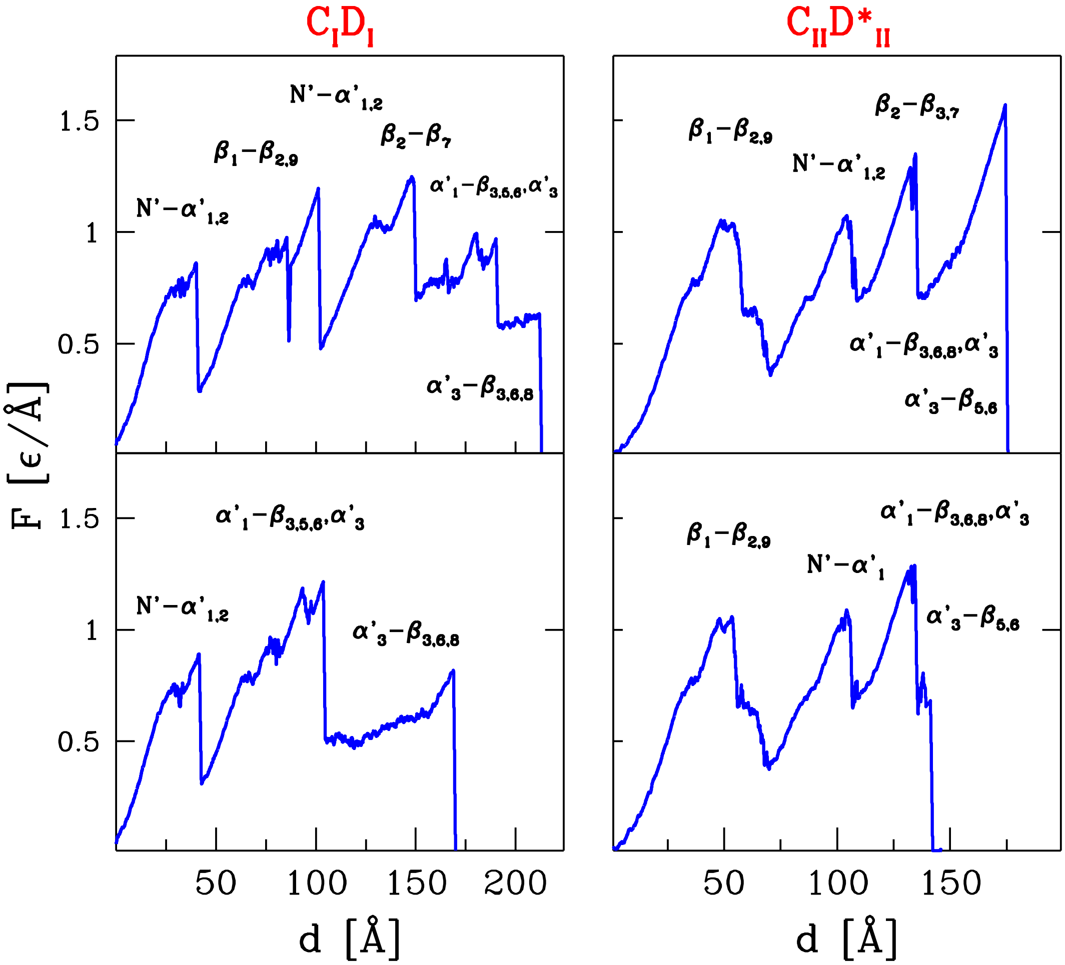

Figure 6. Force-displacement curves for the WT system $C_I D_I$ with $k = 9$ (left panels) and for the mutated system $C_{II} D_{II}^*$ with $k = 19$ (right panels). The upper panels show typical long trajectories. The lower panels correspond to short trajectories. The symbols indicate the contacts that break at particular displacements. The primed symbols refer to Doc and the unprimed to Coh. N' is the region near the N-terminus in Doc. For the short trajectory in the lower panel, the contacts $\alpha_1' - \beta_{3,6,8}$, α_3' break both around 130 and 140 Å, i.e., at the two last force peaks.

whereas for PDB:2CCL, i.e. without the tails, they do with unravelling of $N' - \alpha_1'$ (as shown in ref.[40]). Whenever a trajectory for the system with the tails has a corresponding (long or short) trajectory for the system without the tails, the force peaks are found to be identical within the thermal noise of of 0.1 ε/Å. However the locations of the peaks and of the points of dissociation are often shifted. For instance, long trajectories in $C_I D_I$ result in dissociation around 215 Å but in $C_I d_I$–around 180 Å.

Conclusions

Our computational studies show that the full-length WT Coh-Doc complex exhibits dual binding at room temperature, and the ability to bind in the two modes persists at temperatures elevated by 10 K. At the same time, our MD simulations indicate that each mode of binding leads to two kinds of dissociation pathways in AFM-like stretching trajectories. In our opinion, experimental tests of dual binding require other kinds of spectroscopy experiments. In addition, it would be interesting to find other examples of protein complexes with dual binding. Our MD simulations show also that the full-length Doc is thermally more stable than the truncated Doc missing the terminal tails. This result is fully consistent with recent fluorescence experiments[43]. Taken together, our computational studies provide a detailed analysis of the Coh-Doc energy landscape and of the role of the short, terminal segments in the Doc module.

References

1. Bayer, E. A., Kenig, R. & Lamed, R. Adherence of *Clostridium thermocellum* to cellulose. *J. Bacteriol.* **2**, 818–827 (1983).
2. Beguin, P. & Lemaire, M. The cellulosome: an exocellular, multiprotein complex specialized in cellulose degradation. *Crit. Rev. Biochem. Mol. Biol.* **31**, 201236 (1996).
3. Bayer, E. A., Chanzy, H., Lamed, R. & Shoham, Y. Cellulose, cellulases and cellulosomes. *Curr. Opin. Struct. Biol.* **8**, 548557 (1998).
4. Bayer, E. A., Belaich, J. P., Shoham, Y. & Lamed, R. The cellulosomes: multienzyme machines for degradation of plant cell wall polysaccharides. *Annu. Rev. Microbiol.* **58**, 521–554 (2004).

5. Doi, R. H. & Kosugi, A. Cellulosomes: plant-cell-wall-degrading enzyme complexes. *Nat. Rev. Microbiol.* **2**, 541551 (2004).
6. Demain, A. L., Newcomb, M. & Wu, J. H. Cellulase, clostridia, and ethanol. *Microbiol. Mol. Biol. Rev.* **69**, 124154 (2005).
7. Fontes, C. M. G. A. & Gilbert, H. J. Cellulosomes: Highly efficient nanomachines designed to deconstruct plant cell wall complex carbohydrates. *Ann. Rev. Biochem.* **79**, 655–681 (2010).
8. Gunnoo, M. *et al*. Nano-scale engineering of designer cellulosomes. *Adv. Mat.* **28**, 5619–5647 (2016).
9. Smith, S. P. & Bayer, E. A. Insights into cellulosome assembly and dynamics: From dissection to reconstruction of the supramolecular enzyme complex. *Curr. Opin. Struct. Biol.* **23**, 686–694 (2013).
10. Caspi, J. *et al*. Effect of linker length and dockerin position on conversion of a *Thermobifida fusca* endoglucanase to the cellulosomal mode. *App. Environ. Microbiol.* **75**, 7335–7342 (2009).
11. Vazana, Y. *et al*. A synthetic biology approach for evaluating the functional contribution of designer cellulosome components to deconstruction of cellulosic substrates. *Biotechnology for Biofuels* **6**, 182 (2013).
12. Różycki, B., Cazade, P.-A., O'Mahony, S., Thompson, D. & Cieplak, M. The length but not the sequence of peptide linker modules exerts the primary influence on the conformations of protein domains in cellulosome multi-enzyme complexes. *Phys. Chem. Chem. Phys.* **19**, 21414–21425 (2017).
13. Sali, A. *et al*. Outcome of the first wwPDB hybrid/integrative methods task force workshop. *Structure* **23**, 1156–1167 (2015).
14. Peti, W., Page, R., Boura, E. & Różycki, B. Structures of Dynamic Protein Complexes: Hybrid Techniques to Study MAP Kinase Complexes and the ESCRT System. In: Protein NMR (375–389). Humana Press, New York, NY. *Methods Mol. Biol.* **1688**, 375–389 (2018).
15. Ossowski, I. *et al*. Protein disorder: Conformational distribution of the flexible linker in a chimeric double cellulase. *Biophys. J.* **88**, 2823–2832 (2005).
16. Hammel, M. *et al*. Structural basis of cellulosome efficiency explored by small angle X-ray scattering. *J. Biol. Chem.* **280**, 38562–38568 (2005).
17. Noach, I. *et al*. Inter-modular linker flexibility revealed from crystal structures of adjacent cellulosomal cohesins of *Acetivibrio cellulolyticus*. *J. Mol. Biol.* **391**, 86–97 (2009).
18. Adams, J. J. *et al*. Insights into higher-order organization of the cellulosome revealed by a dissect-and-build approach: Crystal structure of interacting Clostridium thermocellum multimodular components. *J. Mol. Biol.* **396**, 833–839 (2010).
19. Czjzek, M., Fierobe, H. P. & Receveur-Brechot, V. Small-angle X-ray scattering and crystallography: A winning combination for exploring the multimodular organization of cellulolytic macromolecular complexes. *Methods in Enzymology* **510**, 183–210 (2012).
20. Currie, M. A. *et al*. Scaffoldin conformation and dynamics revealed by a ternary complex from the *Clostridium thermocellum* cellulosome. *J. Biol. Chem.* **287**, 26953–26961 (2012).
21. Currie, M. A. *et al*. Small angle X-ray scattering analysis of *Clostridium thermocellum* cellulosome N-terminal complexes reveals a highly dynamic structure. *J. Biol. Chem.* **288**, 7978–7985 (2013).
22. Różycki, B., Cieplak, M. & Czjzek, M. Large conformational fluctuations of the multi-domain Xylanase Z of *Clostridium thermocellum*. *J. Struct Biol.* **191**, 68–75 (2015).
23. Chalupska, D. *et al*. Structural analysis of phosphatidylinositol 4-kinase III β (PI4KB)–14-3-3 protein complex reveals internal flexibility and explains 14-3-3 mediated protection from degradation *in vitro*. *J Struct. Biol.* **200**, 36–44 (2017).
24. Gerngross, U. T., Romaniec, M. P. M., Kobayashi, T., Huskisson, N. S. & Demain, A. L. Sequencing of a *Clostridium thermocellum* gene (cipA) encoding the cellulosomal SL-protein reveals an unusual degree of internal homology. *Mol. Microbiol.* **8**, 325–334 (1993).
25. Dassa, B. *et al*. Pan-cellulosomics of mesophilic clostridia: Variations on a theme. o sl. *Microorganisms* **5**, 74–92 (2017).
26. Ding, S.-Y., Bayer, E. A., Steiner, D., Shoham, Y. & Lamed, R. A scaffoldin of the *Bacteroides cellulosolvens* cellulosome that contains 11 type II cohesins. *J. Bacteriol.* **182**, 4915–4925 (2000).
27. Zhivin, O. *et al*. Unique organization and unprecedented diversity of the *Bacteroides (Pseudobacteroides) cellulosolvens* cellulosome system. *Biotechnol. Biofuels* **10**, 211 (2017).
28. Carvalho, A. L. *et al*. Cellulosome assembly revealed by the crystal structure of the cohesin-dockerin complex. *Proc. Natl. Acad. Sci. USA* **100**, 13809–13814 (2003).
29. Carvalho, A. L. *et al*. Evidence for a dual binding module of dockerin modules to cohesins. *Proc. Natl. Acad. Sci. USA* **104**, 3089–3094 (2007).
30. Pinheiro, B. A. *et al*. The *Clostridium cellulolyticum* dockerin displays a dual binding mode for its cohesin partner. *J. Biol. Chem.* **283**, 18422–18430 (2008).
31. Bule, P. *et al*. Assembly of *Ruminococcus flavefaciens* cellulosome revealed by structures of two cohesindockerin complexes. *Sci. Rep.* **7**, 759 (2017).
32. Slutzki, M. *et al*. Crucial roles of single residues in binding affinity, specificity, and promiscuity in the cellulosomal cohesin-dockerin Interface. *J. Biol. Chem.* **290**, 13654–13666 (2015).
33. Nash, M. A., Smith, S. P., Fontes, C. M. G. A. & Bayer, E. A. Single- versus dual-binding conformations in cellulosomal cohesin-dockerin complexes. *Curr. Opin. Struct. Biol.* **40**, 89–96 (2016).
34. Ostermeier, M. & Benkovic, S. J. Evolution of protein function by domain swapping. *Adv. Protein Chem.* **55**, 29–77 (2001).
35. Jaskólski, M. 3D domain swapping, protein oligomerization, and amyloid formation. *Acta Biochem. Polonica* **48**, 807–828 (2001).
36. Liu, Y. & Eisenberg, D. 3D domain swapping: As domains continue to swap. *Protein Sci.* **11**, 1285–1299 (2002).
37. Valbuena, A. *et al*. On the remarkable mechanostability of scaffoldins and the mechanical clamp motif. *Proc. Natl. Acad. Sci. USA* **106**, 13791–13796 (2009).
38. Hall, B. A. & Sansom, M. S. P. Coarse-Grained MD Simulations and Protein-Protein Interactions: The Cohesin-Dockerin System. *J. Chem. Theory Comput.* **5**, 2465–2471 (2009).
39. Jobst, M. A. *et al*. Resolving dual binding modes of cellulosome cohesin-dockerin complexes using single-molecule force spectroscopy. *eLife* **4**, e10319 (2015).
40. Wojciechowski, M. & Cieplak, M. Dual binding mode in cohesin-dockerin complexes as assessed through stretching. *J. Chem. Phys.* **145**, 134102 (2016).
41. Schymkowitz, J. *et al*. The FoldX web server: an online force field. *Nucl. Acids Res.* **33**, W382–8 (2005).
42. Guerois, R., Nielsen, J. E. & Serrano, L. Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. *J. Mol. Biol.* **320**, 369–387 (2002).
43. Slutzki, M. *et al*. Intramolecular clasp of the cellulosomal *Ruminococcus flavefaciens* ScaA dockerin module confers structural stability. *FEBS Open Bio.* **3**, 398–405 (2013).
44. Sułkowska, J. I. & Cieplak, M. Mechanical stretching of proteins–A theoretical survey of the Protein Data Bank. *J. Phys.: Cond. Mat.* **19**, 283201 (2007).
45. Sikora, M., Sułkowska, J. I. & Cieplak, M. Mechanical strength of 17 134 model proteins and cysteine slipknots. *PLoS Comp. Biol.* **5**, e1000547 (2009).
46. Sikora, M. & Cieplak, M. Mechanical stability of multidomain proteins and novel mechanical clamps. *Proteins: Struct. Funct. Bioinf.* **79**, 1786–1799 (2011).
47. Różycki, B. & Cieplak, M. Citrate synthase proteins in extremophilic organisms: Studies within a structure-based model. *J. Chem. Phys.* **141**, 235102 (2014).

48. Różycki, B., Mioduszewski, Ł. & Cieplak, M. Unbinding and unfolding of adhesion protein complexes through stretching: Interplay between shear and tensile mechanical clamps. *Proteins: Struct. Funct. Bioinf.* **82**, 3144–3153 (2014).
49. Wołek, K., Gómez-Sicilia, Á. & Cieplak, M. Determination of contact maps in proteins: a combination of structural and chemical approaches. *J. Chem. Phys.* **143**, 243105 (2015).
50. Ueda, Y., H. Taketomi, H. & Go, N. Studies on protein folding, unfolding and fluctuations by computer simulations. *Biopolymers* **17**, 1531–1548 (1978).
51. Koga, N. & Takada, S. Role sof native topology and chain-length scaling in protein folding: a simulation study with a Go-like model. *J. Mol. Biol.* **313**, 171–180 (2001).
52. Clementi, C., Nymeyer, H. & Onuchic, J. N. Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for proteion folding? An investigation of small globular proteins. *J. Mol. Biol.* **298**, 937–953 (2000).
53. Karanicolas, J. & Brooks, C. L. III The origins of the asymmetery in the folding transition states of proteins L and G. *Protein Sci.* **11**, 2351–2361 (2002).
54. Sułkowska, J. I. & Cieplak, M. Selection of optimal variants of Go-like models of proteins through studies of stretching. *Biophys. J.* **95**, 3174–3191 (2008).
55. Biasini, M. *et al.* SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* **42**, W252–W258 (2014).
56. Yang, J. *et al.* The I-TASSER Suite: Protein structure and function prediction. *Nature Methods* **12**, 7–8 (2015).
57. Wojciechowski, M., Thompsin, D. & Cieplak, M. Mechanostability of cohesin-dockerin complexes in a structure-based model: Anisotropy and lack of universality in the force profiles. *J. Chem. Phys.* **141**, 245103 (2014).
58. Schymkowitz, J. W. H. *et al.* Prediction of water and metal binding sites and their affinities by using the Fold-X force field. *Proc. Natl. Acad. Sci. USA* **102**, 10147–10152 (2005).
59. Case, D. A. *et al.* Amber14, University of California, San Fransisco (2014).
60. Hornak, V. *et al.* Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins: Struct. Funct. Bioinf.* **65**, 712–725 (2006).
61. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).
62. Ryckaert, J.-P., Ciccotti, G. & Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comp. Phys.* **23**, 327–341 (1977).
63. Wu, X. W. & Brooks, B. R. Self-guided Langevin dynamics simulation method. *Chem. Phys. Lett.* **381**, 512–518 (2003).
64. Darden, T., York, D. & Pedersen, L. Particle mesh ewald - an n.log(n) method for ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).
65. Weiser, J., Shenkin, P. S. & Still, W. C. Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO). *J. Comp. Chem.* **20**, 217–230 (1999).
66. Tsai, J., Taylor, R., Chothia, C. & Gerstein, M. The packing density in proteins: Standard radii and volumes. *J. Mol. Biol.* **290**, 253–266 (1999).
67. Settanni, G., Hoang, T. X., Micheletti, C. & Maritan, A. Folding pathways of prion and doppel. *Biophys. J.* **83**, 3533–3541 (2002).
68. Sorin, E. J. & Pande, V. S. Exploring the helix-coil transition via all-atom equilibrium ensemble simulations. *Biophys. J.* **88**, 2472–2493 (2005).
69. Kwiecinska, J. I. & Cieplak, M. Chirality and protein folding. *J. Phys. Cond. Mat.* **17**, S1565–S1580 (2005).
70. Veitshans, T., Klimov, D. & Thirumalai, D. Protein folding kinetics:Timescales, pathways and energy landscapes in terms of sequence dependent properties. *Folding and Design* **2**, 1–22 (1997).
71. Szymczak, P. & Cieplak, M. Stretching of proteins in a uniform flow. *J. Chem. Phys.* **125**, 164903 (2006).

Acknowledgements

The project has been supported by the European Framework Programme VII NMP grant 604530-2 (CellulosomePlus) and by the National Science Centre, Poland, grant numbers 2014/15/B/ST3/01905 (to MC) and 2016/21/B/NZ1/00006 (to BR). PDQH has been partially supported by Department of Science and Technology, Ho Chi Minh City, Vietnam (grant 179/2016/HD-SKHCN).

Author Contributions

All authors designed the research. Michał Wojciechowski, Bartosz Różycki and Pham Dinh Quoc Huy performed the calculations. All authors contributed to writing the paper.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-23380-9.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.