Stretching to Understand Proteins—A Survey of the Protein Data Bank

Joanna I. Sułkowska and Marek Cieplak
Institute of Physics, Polish Academy of Sciences, Warsaw, Poland

ABSTRACT We make a survey of resistance of 7510 proteins to mechanical stretching at constant speed as studied within a coarse-grained molecular dynamics model. We correlate the maximum force of resistance with the native structure, predict proteins which should be especially strong, and identify the nature of their force clamps.

INTRODUCTION

A common manipulation of single biomolecules involves pulling them by a tip attached to a cantilever moving at a constant speed. Approximately fifty-five proteins have been studied in this way and each protein possesses its own force-displacement pattern, often multipeaked, that reveals mechanical structure of the molecule. The scale of resistance to unraveling is set by a maximum force, F_{max}, which ranges from 35 pN for α-spectrin (1) to 400 pN for the superhelical ankyrin (2) and is close to 200 pN for titin (3) and ubiquitin (4). What are the strong proteins and what makes them strong? The experimental results provide only glimpses of this mechanical landscape that needs to be explored to guide experiments and offer insights into mechanical processes in cells. Here, we provide results of a theoretical survey that determines F_{max} for 7510 proteins and correlates F_{max} with structure.

The set of 7510 proteins, denoted as S7510, contains all nonfragmented structures deposited in the Protein Data Bank (PDB; by August 2005) (5) that are not in complexes and comprise between 40 and 150 amino acids. The CATH-based (6) structure classification scheme is available to a subset of 3813 proteins, S3813, and studies of correlations with the structure are restricted to it. The protein sequence length, N, of 150 extends beyond 120—the size that we observe to be the most probable in the PDB. This length also delimits most single domain units of larger proteins. Approximately twenty of the surveyed proteins have been studied by all-atom simulations. These simulations are demanding computationally and their timescales require considering pulling speeds which are six orders-of-magnitude faster than in experiments. The all-atom models result in large peak forces that are far too large (at ~2000 pN for titin (7)) to yield a detailed understanding of selected proteins. However, they cannot be implemented for PDB-wide surveys. Such a survey can be accomplished by using coarse-grained Gō-like (8) models, as described in Model and Methods. These phenomenological models are well established (9–12) and are defined through the native structure, linking properties to structure. However, this approach restricts the survey to proteins for which the structure is known and is deposited in the PDB. If a given PDB code is represented by many structures, we consider the first of these. If there are several chains corresponding to a code, we take the first listed chain. We restrict the survey to pulling by the termini to restrict the combinatorics of choices.

MODEL AND METHODS

Our molecular dynamics approach is outlined in Cieplak and Hoang (10) and Cieplak et al. (13) and it starts by determining the native contact map. The presence of a contact is decided by checking for overlaps between effective atoms (14). In this procedure, heavy atoms are represented by the van der Waals spheres enlarged by a factor to account for attraction. The potential energy of the system involves the harmonic potential which tethers consecutive C^{α}-values at the equilibrium bond length, d_0, of 3.8 Å. The native contacts are described by the Lennard-Jones potentials,

$$V_{LJ} = 4\epsilon\left[(\sigma_{ij}/r_{ij})^{12} - (\sigma_{ij}/r_{ij})^{6}\right],$$

where r_{ij} is the distance between beads i and j. The length parameters, σ_{ij}, are selected so that the minima of the potentials agree with the native distances between the C^{α} atoms in a contact. The nonnative contacts correspond to a repulsive core of $\sigma = 4$ Å. The i, $i + 2$ native contacts, if detected by the atomic overlaps, are also considered as repulsive cores because they are significantly weaker than hydrogen bonds (as tested by the contact structural units software (15)).

The energy parameter, ϵ, is taken to be uniform and its value represents an effective average of all noncovalent interactions in proteins (800–2300 K). The model also contains a term that favors the native sense of local chirality (16). Contacts in the sulphide bonds (SS) between cysteins are modeled by V_{LJ} with a 20-fold enhanced ϵ to prevent their rupture.

Gō-like models have questionable features in studies of folding but they should be adequate for stretching since the system is nativelike, at least initially. When simulating stretching, both ends of the protein are attached to springs of elastic constant $k = 0.12\ \epsilon/\text{Å}^2$, which is close to the elasticity of experimental cantilevers if one takes 1 kcal/mole as an effective value of ϵ. We perform the survey at $k_BT/\epsilon = 0.3$ since folding is optimal around this temperature for most proteins and because the simulated stretching curves for five domains of titin are similar to experiments (17) at this T. The free end of one of the springs is anchored while that of another moves at a constant speed, v_p, along the initial end-to-end vector. We take $v_p = 0.005$ Å/τ, where the effective characteristic timescale, τ, is of ~0.25 ns (9,18) due to solvent-related effects. This makes v_p two orders-of-magnitude faster than in experimental setups. Thermostatting is provided by the Langevin noise so that equations of motion for each C^{α} read $m\ddot{\vec{r}} = -\gamma\dot{\vec{r}} + \vec{F}_c + \vec{\Gamma}$ and are solved by a fifth-order predictor-corrector scheme. F_c is the net force due

Submitted February 25, 2007, and accepted for publication June 19, 2007.

Address reprint requests to Marek Cieplak, Tel.: 48-22-843-6601, x3365; E-mail: mc@ifpan.edu.pl.

Editor: Ivet Bahar.

0006-3495/08/01/6/08 $2.00

doi: 10.1529/biophysj.107.105973

to the potentials. The damping constant γ is $2m/\tau$ (F_{max} depends on γ only weekly) and the dispersion of the random forces is $\sqrt{2\gamma k_B T}$.

RESULTS AND DISCUSSION

Validation of the model

The defining aspect of the variant of the Gō model we use is that the contact potentials have the Lennard-Jones form with a uniform, i.e., nonspecific, energy amplitude, ϵ. For most proteins, this model yields optimal folding kinetics for temperatures, T, in the range k_BT/ϵ between 0.3 and 0.4, where k_B is the Boltzmann constant. We thus survey proteins at k_BT equal to 0.3 to mimic behavior akin to that expected at a room temperature. The evidence for validation of the model is presented in Fig. 1, which shows a cross plot between the experimental value of F_{max} and its determination, F_{max_th}, within our model (see also (19–21)). We consider proteins for which the PDB structure is available and has no gaps. If several PDB codes correspond to the same protein, F_{max} is averaged over the structures (*open symbols* in Fig. 1). Extracting F_{max} from measurements is complicated by the fact that the force, F, versus displacement, d, curves are often determined for several heterogeneous modules linked in a tandem that need not unwind serially.

Despite such complications, we observe a correlation (the Pearson coefficient is 0.89) corresponding to a linear trend which justifies using the model to compare proteins. We have tried 61 other variants of the Gō model, including those with nonuniform ϵ, and their correlations levels were lower. The models of Clementi et al. (11) and Karanicolas and Brooks (12), however, yield results comparable to ours. It should be noted that the theoretical results for the force come in units of ϵ/Å which complicates making quantitative predictions. The main trend, denoted by the solid line in Fig. 1, can be interpreted as corresponding to ϵ/Å of 67 pN (i.e., with the effective ϵ of ~1 kcal/mole). This translation factor may change when new proteins get added to the testing set whereas the value of F_{max} expressed in ϵ/Å will stay. In the remaining figures, therefore, we use the theoretical units. The bounding slopes, denoted by the dashed lines, correspond to factors 92 and 46 for the lower and upper lines, respectively. When working with single proteins, one could thus adjust the value of ϵ more optimally. For instance, for titin and ubiquitin the factor of 90–100 is adequate. Independent of the choice of the factor, we reproduce the experimentally observed, approximately twofold reduction in F_{max} for ubiquitin when the pulling affects 48-Lys and the C-terminus instead of the two termini (4).

A different kind of validation is provided by pulling bacteriorhodopsin out of a membrane. A Gō-like approach yields a complex F–d pattern (Fig. 10 in (22)) which is like the experimental one (23). Seeking improvement of the model by including the C^β atoms leads to minor shifts in relative values of F_{max} and does not affect the nature of Fig. 1. We could not study the 24-subunit ankyrin since it lacks a deposited structure. Its steel-like properties must be due to its horseshoe-like shape that smaller linkages do not have and generate much smaller forces both in experiment (2,24) and in our model.

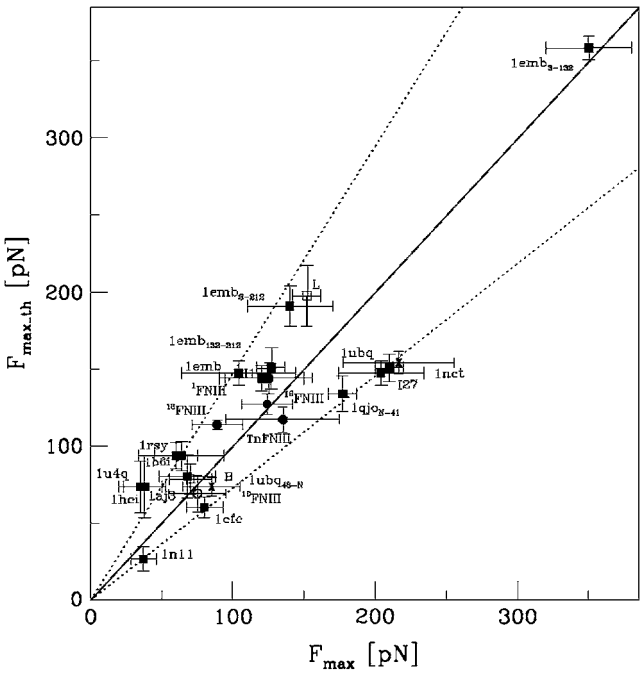

FIGURE 1 F_{max} predicted by our theoretical model versus the corresponding experimental results as listed and referenced in the Appendix. Asterisks correspond to ubiquitin as pulled by the termini (the larger force) or by the K46 and terminus N (the lower force). Symbol *L* denotes protein L (averaged over 2ptl and 1hz5) and *B*, barnase (averaged over 1bni and 1bnr). Circles correspond to fibronectins: solid circles to ^{11}FNIII, ^{12}FNIII, and ^{13}FNIII; open circles to ^{10}FNIII (the latter is averaged over 1fnf, 1ttf, and 1ttg). Generally, the results in this figure are averaged over 10 trajectories. For 1emb$_{3-212}$ however, we show the results for the dominant theoretical pathway (seven trajectories) since an alternative pathway corresponds to a substantially (~50%) higher F_{max}. Three proteins were not included in the figure: 1ksr (the fourth domain of FLN), 1rnh (ribonuclease H), and 1qjo when pulled by the termini (E2lip3). The titinlike structure of the first of these, i.e., with contacts between strands *A* and *G*, is in disagreement with no role of such contacts found in mechanical studies (19,20). The contact map of the second is unstable against small changes in the definition of the contact. There are two reasons to discard the case of 1qjo: the various NMR structures differ significantly in the native direction of the end-to-end vector and the order-of-magnitude smaller experimental F_{max} than for the (N-41) pulling is puzzling (21).

The distribution of F_{max} across the PDB

We now present results of the survey as based on S7510. All results can be accessed at our newly setup web site www.ifpan.edu.pl/BSDB (Biomolecule Stretching Database) by entering the PDB structure code. Fig. 2 shows the distribution of values of F_{max} obtained within our model. They range from 0 to 5.1 (~342 pN) and peak at ~1.4 ϵ/Å (~94 pN). A typical error due to variations between trajectories is ~0.1 (~7 pN). For the I27 domain of titin, F_{max} =

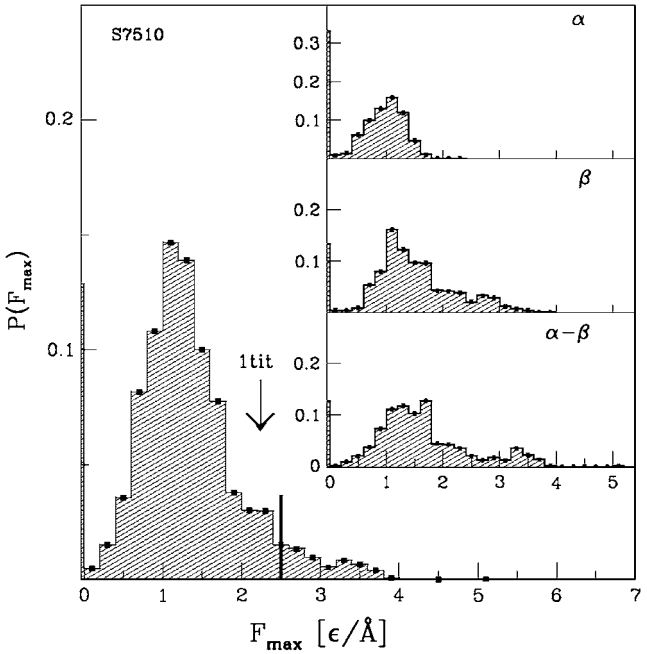

FIGURE 2 Probability distribution of the values of F_{max} of proteins from the set S7510. The arrow indicates the F_{max} for 1tit and the dashed line indicates the threshold above which the proteins belong to the set S134 of the strongest proteins. The inset shows the corresponding distributions for the α, β, and α/β structural classes separately. The maxima in the distributions are near 1.2, 1.5, and 1.2 ϵÅ for the top to bottom panels, respectively. The entries at zero force correspond to cases in which no well-defined force peaks can be identified before covalent bonds become stretched. Such situations arise primarily when the disulphide bonds get involved.

2.1 ϵ/Å—approximately one-half of the largest F_{max} within S7510.

Fig. 3, the top panel, shows values of F_{max} corresponding to a given sequential length. It is seen that, for each N, the forces span comparable ranges of values. Large N proteins may have small F_{max} and smaller proteins may have relatively large F_{max}. However, when one averages entries with the same N, as shown in the bottom panel of Fig. 3, a growing trend is observed. Thus the larger the N, the bigger the probability of generating a large force. This finding is also confirmed by simulations of 239 proteins with N between 153 and 851.

Correlations of F_{max} with structure

The CATH classification scheme divides folds hierarchically into classes, architectures, topologies, and homologies, and assigns a segmented number code to a protein as seen in Table 1. There are marked differences between distributions of F_{max} between the classes (*inset*, Fig. 2). The predicted distributions for the β and α/β proteins have big tails at large forces but the α-proteins are not expected to generate large F_{max}. The weak elastic nature of the α-proteins found is consistent with several experimental results such as those obtained for polycalmodulin (25). However, we expect that

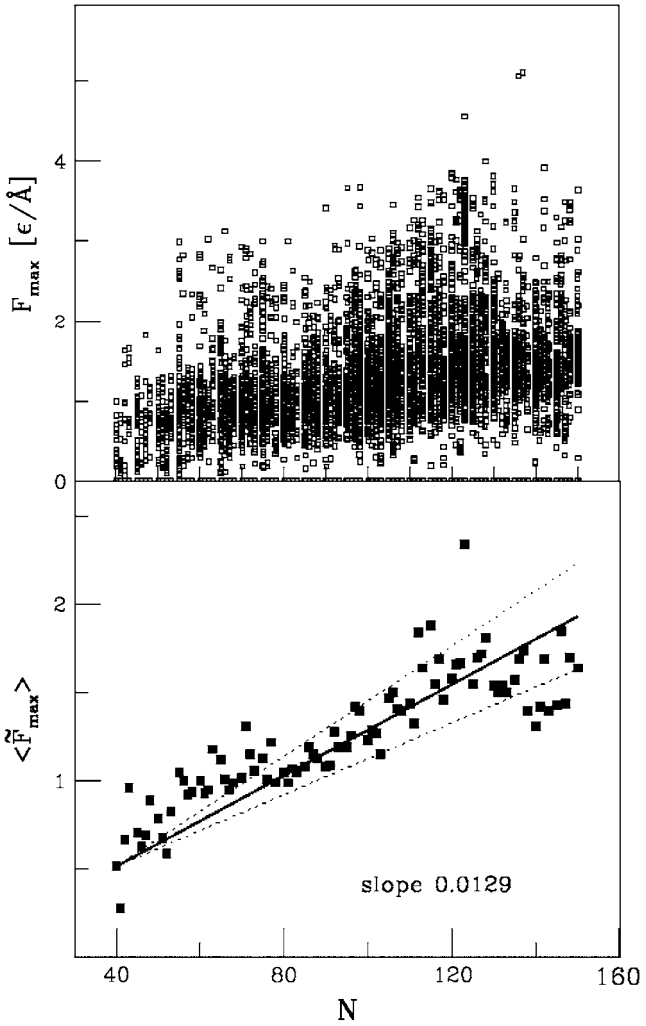

FIGURE 3 The top panel shows values of F_{max} for specific sequential lengths of the protein studied in the survey. The bottom panel shows F_{max} averaged over proteins corresponding to the same value of N.

certain multidomained proteins with α-domains, like 2ng1 and 1cii, may yield substantial forces (2.6 and 1.3 ϵ/Å, respectively).

We now focus on finer characteristics of structure. In the α-class proteins of S3813, 80% have the architecture of an orthogonal bundle. In the β-class, 36% are barrels, 31% are sandwiches, 13% are ribbons, and 13% are rolls. In the α/β class, 40% are α/β rolls and 39% are two-layer sandwiches.

The force distributions corresponding to classes can be further resolved into distributions for specific topologies. This is illustrated in Fig. 4 for the β-sandwiches and α/β rolls. In both cases, the across-the-architecture distributions are broad but with resolvable local maxima. There are three local maxima in the case of the β-sandwich architecture and they all correspond to immunoglobulin-like topology. A further resolution into homologies identifies immunoglobulins and transport proteins as the homologies that yield the larger values of F_{max} (not shown). In the case of the α/β roll architecture, the force distribution has two maxima. The one

TABLE 1 The strongest proteins with $N \leq 150$

n	PDB	F	λ	CATH	n	PDB	F	λ	CATH	n	PDB	F	λ	CATH
1	1c4p	5.1	18	3.10.20	47	1c08*	2.9	6	2.60.40	93	2igd	2.7	7	3.10.20
2	1qqr	5.1	19	3.10.20	48	1i3v	2.9	4	2.60.40	94	4lve	2.7	5	2.60.40
3	1g1k	3.9	6	2.60.40	49	1pgx	2.9	3	3.10.20	95	1igd	2.7	7	3.10.20
4	1c76	3.8	17	3.10.20	50	1l2n	2.9	10	3.10.20	96	1hz6	2.7	10	3.10.20
5	1c77	3.8	25	3.10.20	51	1dfu	2.9	41	2.40.240	97	1a2y	2.7	5	2.60.40
6	1c79	3.8	25	3.10.20	2	1yn4	2.9	5		98	1j1x*	2.7	6	2.60.40
7	1aoh	3.7	6	2.60.40	53	1bmz	2.9	21	2.60.40	99	1fmf	2.7	64	3.40.50
8	1c78	3.7	25	3.10.20	54	1kip*	2.9	6	2.60.40	100	1yn5	2.7	5	
9	2sak	3.7	18	3.10.20	55	1qd0	2.9	4	2.60.40	101	1ap2	2.7	4	2.60.40
10	1nam	3.7	9	2.60.40	56	1ivl*	2.9	7	2.60.40	102	1eo6	2.7	36	3.10.20
11	1so9	3.6	18	2.60.370	57	1tyr	2.9	32	2.60.40	103	1k26	2.7	32	3.90.79
12	1ppx	3.5	40	3.90.79	58	1ic4	2.9	6	2.60.40	104	1ieh	2.7	13	2.60.40
13	1ssn	3.5	34	3.10.20	59	1bz8	2.9	28	2.60.40	105	1hz5	2.6	25	3.10.20
14	1rnz*	3.4	45	3.10.13	60	1vfb*	2.8	6	2.60.40	106	1qp1	2.7	5	2.60.40
15	1ie5*	3.4	14	2.60.40	61	1wtl	2.9	4	2.60.40	107	1sn5	2.7	20	2.60.40
16	1b88	3.4	7	2.60.40	62	1jrk	2.9	32	3.90.79	108	1f2x	2.7	2	2.60.40
17	3rsk*	3.4	45	3.10.130	63	1sok	2.9	25	2.60.40	109	1gb4	2.7	12	3.10.20
18	1npu	3.4	8	2.60.40	64	1bvk	2.9	5	2.60.40	110	1ugm	2.7	33	3.10.20
19	2ncm	3.3	8	2.60.40	65	1ie4	2.8	24	2.60.40	111	1eta	2.7	30	2.60.40
20	1anu	3.3	7	2.60.40	66	2try	2.8	32	2.60.40	112	1tum	2.6	37	3.90.79
21	1rlf	3.3	9	3.10.20	67	1e5a	2.8	26	2.60.40	113	1ufy	2.6	14	3.30.1330
22	1eaj	3.3	8	2.60.40	68	1kir*	2.8	6	2.60.40	114	2rox	2.6	22	2.60.40
23	1oo2	3.3	26	2.60.40	69	1iik	2.8	26	3.10.130	115	1py9*	2.6	4	2.60.40
24	1h5b*	3.2	8	2.60.40	70	1v5o	2.8	16	3.10.20	116	1rbj*	2.55	26	3.10.130
25	1i9e	3.2	4	2.60.40	71	1k53	2.8	26	3.10.20	117	1bzd	2.6	31	2.60.40
26	1mvf	3.1	12	2.60.40	72	1ttr	2.8	26	2.60.40	118	5lve	2.6	5	2.60.40
27	1f5w	3.1	7	2.60.40	73	1kiq*	2.8	6	2.60.40	119	1f86	2.6	23	2.60.40
28	1sp0	3.1	16	2.60.370	74	1oaq	2.8	3	2.60.40	120	1od9*	2.6	9	2.60.40
29	1amx	3.1	31	2.60.40	75	2imm	2.8	5	2.60.40	121	1w19	2.6	39	
30	1i3o	3.1	49	3.40.50	76	1m94	2.8	12	3.10.20	122	1vjk	2.5	23	3.10.20
31	1tfp	3.1	21	2.60.40	77	1kot	2.8	43	3.10.20	123	1ttc	2.5	29	2.60.40
32	1ves	3.1	6		78	1dvy	2.8	22	2.60.40	124	1mfw	2.5	40	3.10.20
33	1sn0	3.1	21	2.60.40	79	1i8k*	2.8	7	2.60.40	125	1pav	2.5	35	3.30.110
34	1oau	3.1	5	2.60.40	80	1h8c	2.8	10	3.10.20	126	1eie	2.5	54	3.10.130
35	1sn2	3.0	21	2.60.40	81	1n4x*	2.6	8	2.60.40	127	1put	2.5	7	3.10.20
36	1oax	3.0	5	2.60.40	82	1wit	2.8	6	2.60.40	128	1mg4	2.5	35	3.10.20
37	1oar	3.0	5	2.60.40	83	1gnu	2.8	44	3.10.20	129	1lm8	2.5	26	3.10.20
38	1pun	3.0	41	3.90.79	84	1em7	2.8	14	3.10.20	130	1ui9	2.5	11	3.30.1330
39	1j05	3.0	5	2.60.40	85	1kgi	2.8	22	2.60.40	131	1nme	2.5	14	3.40.50
40	1lve*	3.0	6	2.60.40	86	1wiu	2.7	5	2.60.40	132	1nvi	2.5	6	3.10.20
41	1fvc	2.9	5	2.60.40	87	1gko	2.7	21	2.60.40	133	1lqb	2.5	23	3.10.20
42	1p7e	2.9	14	3.10.20	88	2dlf	2.7	5	2.60.40	134	1mel	2.5	3	2.60.40
43	1jhl	2.9	6	2.60.40	89	1p4i*	2.7	4	2.60.40					
44	1gke	2.9	27	2.60.40	90	43ca*	2.7	6	2.60.40					
45	1etb	2.9	26	2.60.40	91	43c9	2.7	6	2.60.40		1tit	2.1	4	2.60.40
46	1vhp	2.9	3	2.60.40	92	1tvd	2.7	5	2.60.40		1ubq	2.2	6	3.40.50

Top 134 strongest short proteins as predicted by the Gō-like model used in this article. Titin and ubiquitin are added as a reference. The ordering of proteins corresponding to the same value of F_{max} is arbitrary. F is a shorthand for F_{max} and parameter λ is in percents.
*Indicates proteins that required prohibition of rupture of the SS bonds in the calculations.

at larger forces corresponds to topology of the P-30 proteins. However, this topology spans the entire range of forces and so does the ubiquitin-like topology. Both topologies feed the high F_{max} end of the distribution. The remaining three identified topologies listed in the bottom panel of Fig. 4 are constrained to small forces.

Though structure and force are correlated, we find examples of CATH codes splitting into distinct dynamical behavior suggesting existence of deficiencies in the scheme. For instance, the strong proteins 1p7e (ranked 42nd) and 1pga share the CATH index of 3.10.20.10 with the much weaker (by at least a factor of 4) 1mpe and 1q10. All of these proteins have the sequential length of 56. 1pga and 1q10 differ by 1.9 Å in RMSD and by a three-point mutation that eliminates several long-range contacts and reduces the force.

The set of the strongest proteins

The set S134 of the top 134 (top 1.8%) strongest proteins is listed in Table 1. It comprises structures with $F_{max} > 2.5$

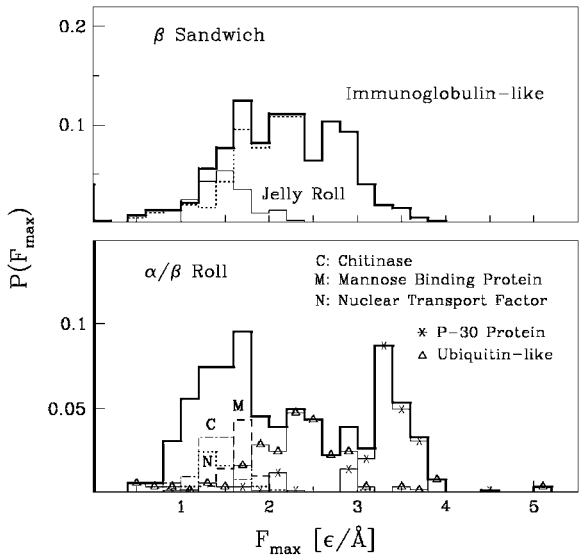

FIGURE 4 Decomposition of the distribution of F_{max} corresponding to a given architecture into contributions related to specific topologies that are indicated in the figure. The top panel is for the β-sandwich architecture whereas the bottom panel is for the α/β roll. Both architectures are represented by ~470 proteins each.

ϵ/Å. (When using the slope of the main trend in Fig. 1, the threshold would correspond to ~170 pN but the I27 domain of titin is not in S134 since its F_{max} is 2.1 ϵ/Å.) The table displays values of N, F_{max}, the relative location of the main maximum, λ, and the symbol of structural CATH classification if available. The parameter $\lambda = (L_m - L_n)/(L_f - L_n)$ is defined in terms of characteristic end-to-end distances L: L_n is the native value, L_m corresponds to the location of the tallest force peak, and L_f to full extension of $(N - 1)$ 3.8 Å. We find that the distribution of the values of λ is peaked at ~10% for the set S134 whereas it is rather flat generally, indicating that large forces often come with rupture events near the termini as in titin (7,13). We find that 72% of the strong proteins has the *F–d* pattern in which a major peak is followed by a minor peak, 19% have also some number of preceding peaks, and 7% a preceding peak and no after-peak. Only four proteins, including the top two, have just one force maximum.

The distribution of forces across architectures is changed significantly relative to the general case when one focuses on the strongest proteins. None of the strong proteins belongs to the α-class. The proteins in S134 belong to six architectures. Two of them are especially well represented: β-sandwiches (60%) and α/β rolls (30%). The strong proteins belong to 11 topologies, and Immunoglobulin-like (2.60.40 in the CATH scheme) and Ubiquitin-like (UB roll; 3.10.20) are the most frequent of these (the remaining CATH codes are 3.90.79, 3.60.370, 3.40.50, 3.30.110, 3.10.130, 3.10.1330, 3.10.50, 2.40.40, and 2.40.240).

The proteins in Table 1 are not necessarily distinct biologically and there could be several PDB codes corresponding to nearly the same protein. Stretching is sensitive to structural details and thus to the particular PDB code. We have found that 41 proteins in S134 are unrelated homologically whereas the remaining 93 belong to 33 different groups of at least two elements each. In particular, the top two proteins, 1c4p and 1qqr, are both streptokinase β-domain proteins (UB roll topology) but involved in different functions (blood clotting and hydrolase activation, respectively). The third-ranked protein, 1g1k, the seventh-ranked 1aoh and the 20th-ranked 1anu are all cohesin domains of the cellulosome from *Clostridium thermocellum*. However, only the latter two show close homology.

The types of mechanical clamps

We now ask what makes proteins strong. The mechanisms of unfolding can be elucidated through scenario diagrams (13) that show at what displacement d_u a given native contact is broken for the last time. A contact is identified by the sequential distance $|j - i|$. It is declared to be broken if $r_{ij} > 1.5\sigma_{ij}$. An accumulation of contacts unfolding at a value of d_u indicates a force peak. A scenario diagram for the

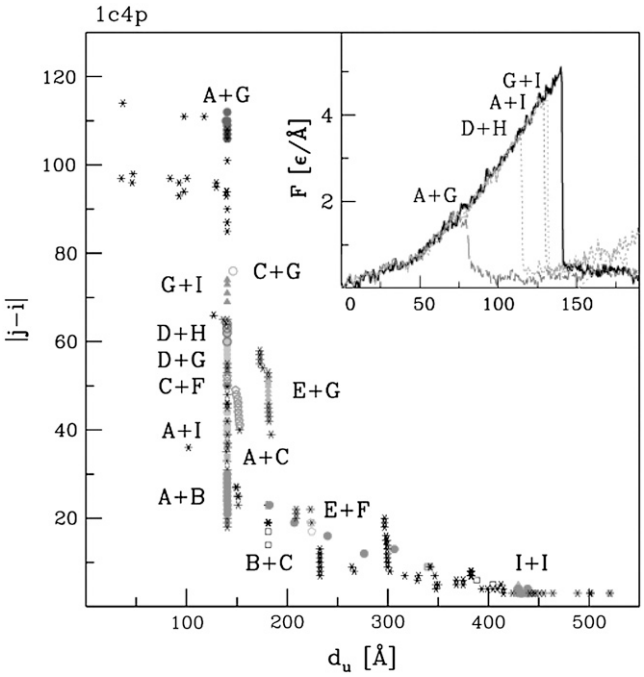

FIGURE 5 Unfolding scenario for 1c4p. The asterisks correspond to contacts which do not involve any secondary structures. The remaining symbols are diversified and the letter symbols placed next to them indicate which secondary structures are involved in a contact that is broken at a distance d_u. The inset refers to the *F–d* curves. The solid line with notation all corresponds to a situation in which all contacts are present. The remaining lines correspond to a situation in which the indicated contacts are removed. This protein may also unfold along a different pathway with F_{max} of 4.8 ϵ/Å and a secondary maximum due to a delayed rupture of C+F, D+H, D+G, and G+I.

strongest protein 1c4p is shown in Fig. 5. This protein consists of four chains which we find to possess similar elastic properties and the figure refers to the first chain. The chain contains an α-helix (196–210), denoted as I, and eight β-strands, denoted by A–H, as labeled consecutively from the N- to C-termini. Fig. 5 shows that many sets of contacts (e.g., between A and G, between A and B, etc.) rupture around d of 140 Å. However, their elastic contributions are strongly uneven. Which of them correspond to the mechanical clamp that holds the protein the most? One can identify the most relevant set of contacts by removing the sets one at a time and by inspecting the resulting F–d curves. The inset of Fig. 5 indicates that the contacts between parallel strands A (amino acids 158–168) and G (266–278) contribute 50% to $F_{\max}$ and thus form the heart of the clamp.

In 90% of structures in S134, the mechanical clamp is found to be due to long parallel β-strands that are shear-ruptured on pulling (7,26,27). Examples of such clamps, marked in solid representation, are seen in the upper six panels of Fig. 6. The top peak forces arise when at least one of these strands is near a terminus. The mechanical clamp may act at the beginning of folding. Oftentimes, however, especially when the termini are on the same side of the native protein, a prior unwinding of the surrounding layers is required which results in a rotation and minor preceding peaks. The strength of the clamp is governed primarily by the number of contacts within the clamp and then by any cross-linking (and usually hydrophobic) stabilizing interactions that may encase the clamp. The bottom three panels of Fig. 6 show examples of nontypical mechanical clamps that are found in 1amx, 1qp1, and 1pav. In 1amx (also in 1ei5 and 1lm8), the clamp strands are antiparallel; for 1qp1 (also in 1tum and 1f86), the strands are unstructured and do not form a β-sheet. Finally, in 1pav, the clamp is formed by a box made of two antiparallel β-strands placed next to two antiparallel α-helices, all shearing against each other on pulling.

The role of the disulphide bridges

In the basic model, sulphide-bond contacts (SS) between cysteins are not distinguished even though they cannot be ruptured. The SS bonds may affect the rupture process significantly as illustrated in Fig. 7 for 1rnz (ribonuclease A) If the SS contacts were not enhanced then we would get $F_{\max} = 2.7\ \epsilon$/Å occurring at $\sim d = 250$ Å. 1rnz contains four SS contacts. Two of them break before reaching $F_{\max}$ and two contribute to $F_{\max}$. Disallowing for the rupture of the four contacts makes the major force-peak occur earlier and $F_{\max}$ raises to 3.2 ϵ/Å, advancing 1rnz to the top 15 in S134. The scenario diagram shows that the second half of the process evolves very differently once the SS contacts are not allowed to break and the clamp is also distinct. Another example is 1lsl for which one of its six SS bonds intervenes early and confines the stretching to a segment of 37 out of 113 amino acids for which no force-peak develops. The presence of SS bonds, however, need not affect $F_{\max}$ in a major way, especially if their rupture is scheduled to occur past the major

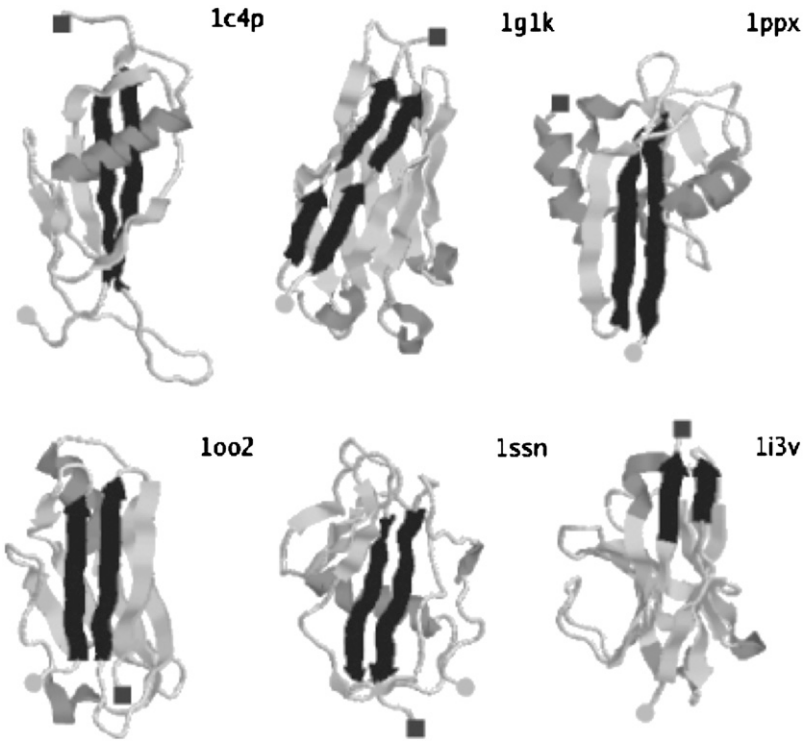

FIGURE 6 Ribbon representation of the strong short proteins as indicated. The elements in solid representation correspond to the mechanical clamp that yields the largest resistance to pull. For 1g1k, the relevant strands are not contiguous.

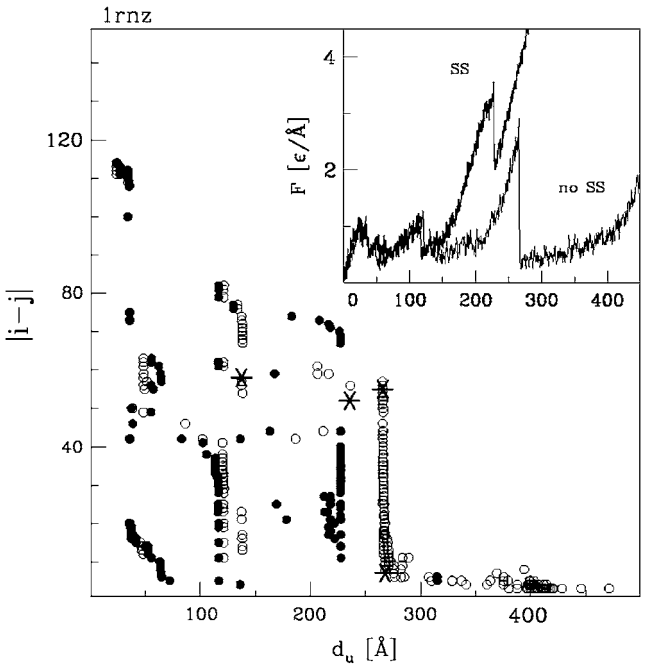

FIGURE 7 Unfolding scenario and the F–d curves, in the inset, for protein 1rnz in two Gō-like models. One model is standard and assumes no special treatment of the SS bonds between the cysteines. It yields the thinner F–d curve and the contact breaking distances corresponding to open circles and stars. The stars indicate contacts between cysteines. The other model does not allow for rupture of the SS bonds. It yields the thicker F–d curve and the data points corresponding to solid circles in the scenario diagram.

peak. It should be noted that comparing stretching in the models with and without the energy enhancement is meaningful physically since the SS bonds can be converted to weaker SH bonds by an application of the reducing agent DTT. Such experimental studies have been performed, e.g., with the cell adhesion molecule Mel-CAM (28).

CONCLUSIONS

We have used a simple geometry-based coarse-grained model and found the distribution of $F_{\max}$ for 7510 proteins to be non-Gaussian. We have correlated the forces with architectures and topologies. In particular, we find no α-proteins in the set that we would expect to be strong.

We make a prediction regarding what proteins are expected to be especially strong. Such proteins belong to a short list of topologies and their strength arises from a clamp, which usually consist of long and parallel β-strands along the force vector but other mechanisms also exist. Taking into account refinements in the model, such as the presence of side groups and of the sulphide bridges, mostly reshuffles the ranking without affecting the top of the list. This suggests that the Gō-model-based and PDB-wide identification of strong proteins may find support in all-atom simulations and experiments.

APPENDIX

Proteins used in the validation of the model and the experimental values of $F_{\max}$

TABLE 2 The listing is from the highest to the lowest experimental value of $F_{\max}$

Protein		References
1v9e carbonic anhydrase II	1100 pN	(29)
1n11 Ankyrin*24	450 pN	(24,30)
1emb green fluorescent protein (3-132)	350 pN	(19,43,44)
1wit I28 titin	230 pN	(31)
1ubq ubiquitin N-C	220 pN	(4,32)
1nct M5	210 pN	(33,34)
1tit titin	210 pN	(3)
1qjo E2lip3 N-41	177 pN	(21)
1hz6/2ptl protein L	152 pN	(35)
1ten $fnIII^{3}$	135 pN	(36)
1emb (3-212)	130 pN	(19,43,44)
1g1c M5 titin	127 pN	(31)
1fnh $fnIII^{12}$	124 pN	(37)
1emb (132-212)	120 pN	(19,43)
1emb (N-C)	104 pN	(19,43)
1vsc Mel-CAM	95 pN	(28)
1fnh $fnIII^{13}$	89 pN	(37)
1ubq ubiquitin 48-C	76 pN	(4,32)
1cfc poly-calmodulin	80 pN	(38)
1fnf/1ttf/1ttg $fnIII^{10}$	78 pN	(37,39)
1bni/1bnr-barnase	70 pN	(40)
1b6i T4 lysozyme	60 pN	(41)
1rsy/1dqv calcium binding	80 pN	(38)
1aj3 spectrin R16	54 pN	(42)
1ksr/1whl DdFLN	45 pN	(19,43,44)
1u4q spectrin r13-r18	35 pN	(1,45,46)
1hci spectrin α-actin	38 pN	(1,46)
1n11 ankyrin, one subunit	37 pN	(24,30)
1rnh/2rn2 ribonuclease H	20 pN	(47) (not included)
1qjo E2lpi3 (N-41)	20 pN	(21) (not included)

This project was started by an inspiring conversation with Julio M. Fernandez and benefited from many of his later suggestions. This project was also helped by plenty of advice from Jayanth R. Banavar. We appreciate useful comments of Trinh Xuan Hoang, Harald Janovjak, and Piotr Szymczak.

This work was funded by the Ministry of Science and Informatics in Poland (grant No. 202-021-31-0739).

REFERENCES

1. Rief, M., J. Pascual, M. Saraste, and H. G. Gaub. 1999. Single molecule force spectroscopy of spectrin repeats: low unfolding forces in helix bundles. *J. Mol. Biol.* 286:553–561.
2. Lee, G., K. Abdi, Y. Jiang, P. Michaely, V. Bennett, and P. E. Marszalek. 2006. Nanospring behavior of ankyrin repeats. *Nature*. 440:246–249.
3. Li, H., W. A. Linke, A. F. Oberhauser, M. Carrion-Vazquez, J. G. Kerkviliet, H. Lu, P. E. Marszalek, and J. M. Fernandez. 2002. Reverse engineering of the giant muscle protein titin. *Nature*. 418:998–1002.
4. Carrion-Vazquez, M., H. Li, H. Lu, P. E. Marszalek, A. F. Oberhauser, and J. M. Fernandez. 2003. The mechanical stability of ubiquitin is linkage dependent. *Nat. Struct. Biol.* 10:738–743.

5. Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235–242.
6. Orengo, C. A., A. D. Micchie, S. Jones, D. T. Jones, M. B. Swindels, and J. M. Thornton. 1997. CATH—a hierarchical classification of protein domain structures. *Structure.* 5:1093–1108.
7. Lu, H., B. Isralewitz, A. Krammer, V. Vogel, and K. Schulten. 1998. Unfolding of titin immunoglobulin domains by steered molecular dynamics simulation. *Biophys. J.* 75:662–671.
8. Abe, H., and N. Gō. 1981. Noninteracting local-structure model of folding and unfolding transition in globular proteins. II. Application to two-dimensional lattice proteins. *Biopolymers.* 20:1013–1031.
9. Veitshans, T., D. Klimov, and D. Thirumalai. 1997. Protein folding kinetics: time scales, pathways and energy landscapes in terms of sequence-dependent properties. *Folding Des.* 2:1–22.
10. Cieplak, M., and T. X. Hoang. 2003. Universality classes in folding times of proteins. *Biophys. J.* 84:475–488.
11. Clementi, C., H. Nymeyer, and J. N. Onuchic. 2000. Topological and energetic factors: what determines the structural details of the transition state ensemble and ''on-route'' intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* 298: 937–953.
12. Karanicolas, J., and C. L. Brooks III. 2002. The origins of asymmetry in the folding transition states of protein L and protein G. *Protein Sci.* 11:2351–2361.
13. Cieplak, M., T. X. Hoang, and M. O. Robbins. 2004. Thermal effects in stretching of Gō-like models of titin and secondary structures. *Proteins Struct. Funct. Biol.* 56:285–297.
14. Tsai, J., R. Taylor, C. Chothia, and M. Gerstein. 1999. The packing density in proteins: standard radii and volumes. *J. Mol. Biol.* 290:253–266.
15. Sobolev, V., A. Sorokine, J. Prilusky, E. E. Abola, and M. Edelman. 1999. Automated analysis of interatomic contacts in proteins. *Bioinformatics.* 15:327–332.
16. Kwiecinska, J. I., and M. Cieplak. 2005. Chirality and protein folding. *J. Phys. Cond. Mat.* 17:S1565–S1580.
17. Cieplak, M., A. Pastore, and T. X. Hoang. 2005. Mechanical properties of the domains of titin in a Gō-like model. *J. Chem. Phys.* 122:054906.
18. Szymczak, P., and M. Cieplak. 2006. Stretching of proteins in a uniform flow. *J. Chem. Phys.* 125:164903.
19. Schwaiger, I., A. Kardinal, M. Schleicher, A. A. Noegel, and M. Rief. 2004. A mechanical unfolding intermediate in an actin-crosslinking protein. *Nat. Struct. Mol. Biol.* 11:81–85.
20. Schweiger, I., M. Schleicher, A. E. Noegel, and M. Rief. 2005. The folding pathway of a fast-folding immunoglobulin domain revealed by single-molecule mechanical experiment. *EMBO Rep.* 6:46–51.
21. Brockwell, D. J., E. Paci, R. C. Zinober, G. S. Beddard, P. D. Olmsted, D. A. Smith, R. N. Perham, and S. E. Radford. 2003. Pulling geometry defines mechanical resistance of β-sheet protein. *Nat. Struct. Biol.* 10:731–737.
22. Cieplak, M., S. Filipek, H. Janovjak, and K. A. Krzysko. 2006. Pulling single bacteriorhodopsin out of a membrane: comparison of simulation and experiment. *Biochim. Biophys. Acta.* 1758:537–544.
23. Janovjak, H., M. Kessler, D. Oesterhelt, H. G. Gaub, and D. J. Muller. 2003. Unfolding pathways of native bacteriorhodopsin depend on temperature. *EMBO J.* 22:5220–5229.
24. Li, L. W., S. Wetzel, A. Pluckthun, and J. M. Fernandez. 2006. Stepwise unfolding of ankyrin repeats in a single protein revealed by atomic force microscopy. *Biophys. J.* 90:L30–L32.
25. Carrion-Vazquez, M., A. F. Oberhauser, T. E. Fisher, P. E. Marszalek, H. Li, and J. M. Fernandez. 2000. Mechanical design of proteins studied by single-molecule force spectroscopy and protein engineering. *Prog. Biophys. Mol. Biol.* 74:63–91.
26. West, D. K., D. J. Brockwell, P. D. Olmsted, S. E. Radford, and E. Paci. 2006. Mechanical resistance of proteins explained using simple molecular models. *Biophys. J.* 90:287–297.
27. Li, P. C., and D. E. Makarov. 2004. Simulation of the mechanical unfolding of ubiquitin: probing different unfolding reaction coordinates by changing the pulling geometry. *J. Chem. Phys.* 121:4826–4832.
28. Carl, P., C. H. Kwok, G. Manderson, D. W. Speicher, and D. E. Discher. 2001. Force unfolding modulated by disulphide bonds in the Ig domains of a cell adhesion molecule. *Proc. Natl. Acad. Sci. USA.* 98:1565–1570.
29. Baumann, C. G., V. A. Bloomfield, S. B. Smith, C. Bustamante, M. D. Wang, and S. M. Block. 2000. Stretching of single collapsed DNA molecules. *Biophys. J.* 78:1965–1978.
30. Lee, G., K. Abdi, Y. Jiang, P. Michaely, V. Bennett, and P. E. Marszalek. 2006. Nanospring behavior of ankyrin repeats. *Nature.* 440:246–249.
31. Li, H., A. F. Oberhauser, S. B. Fowler, J. Clarke, and J. M. Fernandez. 2000. Atomic force microscopy reveals the mechanical design of a modular protein. *Proc. Natl. Acad. Sci. USA.* 97:6527–6531.
32. Chyan, C.-L., F.-C. Lin, H. Peng, J.-M. Yuan, C.-H. Chang, S.-H. Lin, and G. Yang. 2004. Reversible mechanical unfolding of single ubiquitin molecules. *Biophys. J.* 87:3995–4006.
33. Watanabe, K., C. Muhle-Goll, M. S. Z. Kellermayer, S. Labeit, and H. L. Granzier. 2002. Different molecular mechanics displayed by titin's constitutively and differentially expressed tandem Ig segments. *Struct. Biol. J.* 137:248–258.
34. Watanabe, K., P. Nair, D. Labeit, M. S. Z. Kellermayer, M. Greaser, S. Labeit, and H. L. Granzier. 2002. Molecular mechanics of cardiac titins PEVK and N2B spring elements. *J. Biol. Chem.* 277:11549–11558.
35. Brockwell, D. J., S. Godfrey, S. Beddard, E. Paci, D. K. West, P. D. Olmsted, D. Alastair Smith, and S. E. Radford. 2005. Mechanically unfolding small topologically simple protein L. *Biophys. J.* 89:506–519.
36. Leahy, D. J., W. A. Hendrickson, I. Aukhil, and H. P. Erickson. 1992. Structure of a fibronectin type III domain from tenascin phased by MAD analysis of the selenomethionyl protein. *Science.* 258:987–991.
37. Oberhauser, A. F., C. Badilla-Fernandez, M. Carrion-Vazquez, and J. M. Fernandez. 2002. The mechanical hierarchies of fibronectin observed with single-molecule AFM. *J. Mol. Biol.* 31:433–447.
38. Carrion-Vazquez, M., A. F. Oberhauser, T. E. Fisher, P. E. Marszalek, H. Li, and J. M. Fernandez. 2000. Mechanical design of proteins studied by single-molecule force spectroscopy and protein engineering. *Prog. Biophys. Mol. Biol.* 74:63–91.
39. Li, L., H. Han-Li Huang, C. L. Badilla, and J. M. Fernandez. 2005. Mechanical unfolding intermediates observed by single-molecule force spectroscopy in fibronectin type III module. *J. Mol. Biol.* 345:817–826.
40. Best, R. B., B. Li, A. Steward, V. Daggett, and J. Clarke. 2001. Can non-mechanical proteins withstand force? Stretching barnase by atomic force microscopy and molecular dynamics simulation. *Biophys. J.* 81:2344–2356.
41. Yang, G., C. Cecconi, W. A. Baase, I. R. Vetter, W. A. Breyer, J. A. Haack, B. W. Matthews, F. W. Dahlquist, and C. Bustamante. 2000. Solid-state synthesis and mechanical unfolding of polymers of T4 lysozyme. *Proc. Natl. Acad. Sci. USA.* 97:139–144.
42. Lenne, P. F., A. J. Raae, S. M. Altmann, M. Saraste, and J. K. H. Horber. 2000. States and transition during unfolding of a single spectrin repeat. *FEBS Lett.* 476:124–128.
43. Schlierf, M., and M. Rief. 2005. Temperature softening of a protein in single-molecule experiments. *J. Mol. Biol.* 354:497–503.
44. Schlierf, M., and M. Rief. 2006. Single-molecule unfolding force distribution reveals a funnel-shape energy landscape. *Biophys. J.* 90:L33–L35.
45. Law, R., P. Carl, S. Harper, P. Dalhaimer, D. W. Speicher, and D. E. Discher. 2003. Cooperativity in force unfolding of tandem spectrin repeats. *Biophys. J.* 84:533–544.
46. Law, R., P. Carl, S. Harper, D. W. Speicher, and D. E. Discher. 2004. Influence of lateral association on forced unfolding of antiparallel spectrin heterodimers. *J. Biol. Chem.* 279:16410–16416.
47. Cecconi, C., E. A. Shank, C. Bustamante, and S. Marqusee. 2005. Direct observation of the three-state folding of a single protein molecule. *Science.* 309:2057–2060.