Mechanical stretching of proteins—a theoretical survey of the Protein Data Bank

IOP PUBLISHING JOURNAL OF PHYSICS: CONDENSED MATTER

J. Phys.: Condens. Matter **19** (2007) 283201 (60pp) doi:10.1088/0953-8984/19/28/283201

TOPICAL REVIEW

Mechanical stretching of proteins—a theoretical survey of the Protein Data Bank

Joanna I Sułkowska and Marek Cieplak[1]

Institute of Physics, Polish Academy of Sciences, Aleja Lotników 32/46, 02-668 Warsaw, Poland

E-mail: mc@ifpan.edu.pl

Received 3 November 2006, in final form 28 January 2007
Published 25 June 2007
Online at stacks.iop.org/JPhysCM/19/283201

Abstract

The mechanical stretching of single proteins has been studied experimentally for about 50 proteins, yielding a variety of force patterns and peak forces. Here we perform a theoretical survey of proteins of known native structure and map out the landscape of possible dynamical behaviours under stretching at constant speed. We consider 7510 proteins comprising not more than 150 amino acids and 239 longer proteins. The model used is constructed based on the native geometry. It is solved by methods of molecular dynamics and validated by comparing the theoretical predictions to experimental results. We characterize the distribution of peak forces and investigate correlations with the system size and with the structure classification as characterized by the CATH scheme. Despite the presence of such correlations, proteins with the same CATH index may belong to different classes of dynamical behaviour. We identify proteins with the biggest forces and show that they belong to few topology classes. We determine which protein segments act as mechanical clamps and show that, in most cases, they correspond to long stretches of parallel β-strands, but other mechanisms are also possible.

(Some figures in this article are in colour only in the electronic version)

Contents

[1] Author to whom any correspondence should be addressed.

1. Introduction

Atomic force microscopy (AFM), optical tweezers and related techniques have been developed to a sufficiently high degree of sophistication to allow for the manipulation of single large molecules [1]. The basic mode of such an analysis involves stretching at a constant speed. Another way of manipulation, known as the force clamp, involves adjusting the speed so that the pulling tension is constant. As the molecule is stretched at constant speed, it resists the pull by exerting a force, F, on the pulling device, such as a tip of the AFM cantilever. This force can be monitored by optical means and the outcomes of such experiments are presented as the force versus displacement, d, curves. The F–d curves may show a linear, i.e. Hookean, behaviour at small extensions but generally develop into complex multipeak patterns that contain implicit information about the internal structure of the molecule. In particular, the largest force, F_{max}, in the pattern (determined before reaching a full extension) provides information about the toughest structural unit, or a 'mechanical clamp', that is contained in the system.

Such experiments were performed first on the streptavidin–biotin complex [2], polysaccharides [70], and then on the nucleic acids [3] and proteins. Among the proteins that were stretched experimentally, the giant muscle molecule titin [4–9, 11–15] and ubiquitin [16, 17] have been especially well studied. Both proteins yield a maximum pulling force of about 200 pN but their force–extension patterns differ. As noted by Lu *et al* [15], such values of F_{max} indicate that the mechanical clamp involves a cluster of bonds that unravel simultaneously since breaking of a single hydrogen bond in the two-strand DNA generates F_{max} of only about 13 pN [3]. Separating biotin from streptavidin also involves stretching many bonds together and, within the pulling distance of about 10 Å, the peak force approaches about 300 pN [2].

Titin has a β-sandwich architecture and ubiquitin that of the α/β roll. The maximum force for C2A (which is a three-layer α/β sandwich) just exceeds 60 pN [18] and the peak forces are hard to discern for polycalmodulin [18] which is a mainly α orthogonal bundle type protein. At the other extreme, it has recently been reported [19] that the superhelical ankyrin behaves 'like steel'—it requires a force of around 400 pN to unravel even though individual repeats of the ankyrin modules need about 50 pN to unfold [20]. One can even get peak forces of about 1100 pN when one considers a mutant of bovine carbonic anhydrase II with a disulfide bridge introduced to stabilize a trefoil knot structure of the protein [21, 64]. Are there any other 'steely' proteins that have not yet been studied? What is the distribution of maximum structure-unravelling forces across proteins and what governs the values of the maximal forces? Is titin on the strong side of the distribution, or is it merely typical? What types of protein structures are likely to generate large forces?

In order to answer these and related questions, it is advisable to make a survey of all available information on F_{max} for proteins and chart a 'map' of this unknown territory. The

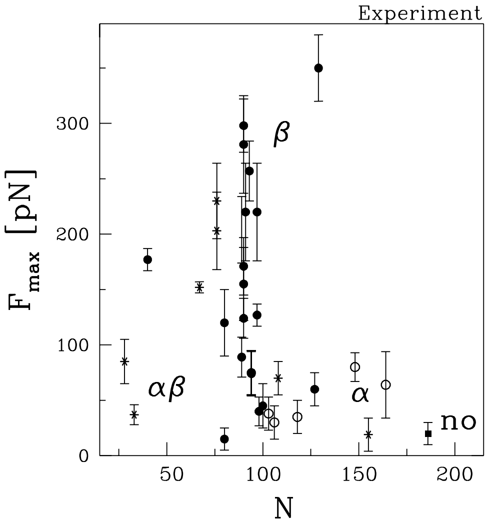

Figure 1. Summary of the experimental results on the peak forces in stretched proteins. Open circles, solid circles, asterisks and solid squares refer to α, β, α/β and no structure proteins, respectively. The proteins are identified by the number of amino acids, N, that they contain.

purpose of such a map would be to guide experimental stretching studies and to offer insights into processes of mechanical deformation that occur in cells, for instance in trans-membrane transfers [22–27] or in molecular motors [28–30].

A good point to start is to review the known experimental results. Table 1 and figure 1 summarize data on the maximal stretching force that are available for about 55 proteins. Table 1 is an updated version of the tables presented in [18], [31] and [32] for 4, 8 and 21 proteins, respectively. The pulling speeds range between 1 and 10^4 nm s^{-1}. The data show that most of the proteins studied mechanically have a number of amino acids, N, not exceeding 162 or involve tandem arrangements of such proteins. The values of $F_{\rm max}$ seem to cluster in a region extending up to 300 pN and then there are some proteins that are strongly resistant to unravelling and lie outside of the scale of figure 1. The experimental set of proteins, however, is too spotty to generate a sense of a landscape in the emerging map and, especially, to uncover dependences on the structural classes. Even fewer proteins, around 20 as listed in table 2, have been studied theoretically by all-atom molecular dynamics simulations. These simulations are limited by nanosecond timescales that are accessible to computations and thus require considering pulling speeds which are six to seven orders of magnitude bigger than those available to experiments. The resulting peak forces are often found to be unrealistically large, for instance of the order of 2000 pN for titin [15, 86], especially when solvent molecules are involved in the simulations. Some of these problems may be remedied by considering variants of the molecular dynamics approach. For instance, Pabo and Amzel [89] proposed a quasi-equilibrium simulation in which the peak force for titin is reduced to under 400 pN. Despite the fact that all-atom simulations may yield a detailed understanding of some small sets of proteins, they are too demanding computationally to generate a more global picture.

Here, we take the position that the sensible way to make a survey of the mechanical properties of proteins is by using approximate coarse-grained Go-like models. These phenomenological models are now well established [94–98] and their defining quality is that the native structure constitutes an input to their construction. This quality links properties to structures directly. However, it necessarily restricts the survey to proteins with known native structures. (A similar comment applies to stochastic models such as the one considered in [94].)

Table 1. Summary of experimental findings on the constant speed stretching of proteins by their termini. The Protein Data Bank (PDB) code is listed if the corresponding structure (or one of the domains from stretched tandem repeats) is deposited in the PDB. The symbols are defined in the main text. The values of N refer to the typical sizes of single domains.

Protein	PDB	N	F_{max} (pN)	v_p (μm s^{-1})	CATH	References
Titins						
$\mathrm{I1}_{I27}$	1g1c	97	127	600	2.60.40.10	[50]
I4		90	171+/26		2.60	[47]
I5		~90	155 ± 33		2.60	[47]
I4–I11		~90	150–200+/30		2.60	[47]
I27	1tit	89	204 ± 30	0.2–1.5 typically 0.6	2.60.40.10	[8, 6, 9, 14, 37] [38, 39, 46, 47] [42, 49, 57, 77]
I28		93	257 ± 27		2.60.40.10	[46, 47, 14]
I27–I28		~91	211–306	1	2.60.40.10	[46]
I27–I30		~90	230		2.60.40.10	[8]
I27–I34		~90	150–330 ± 20	1	2.60.40.10	[8, 46, 47]
I27–I34		~90	231 ± 26	0.5	2.60.40.10	[67, 76]
I32		90	298 ± 24		2.60.40.10	[47]
I34		90	281		2.60.40.10	[47]
Sk47–Sk53		~96	210	0.5	2.60	[67]
I54–I59	1nct	~98	210	0.5–1	2.60.40.10	[76, 78]
Fibronectins						
FNI			70–100	0.6	2.60	[66]
FNII			90–150	0.6	2.60	[66]
FNIII		~90	80–200	0.6	2.60	[61, 66, 69, 75]
$^{1}\mathrm{FNIII}_{\mathrm{2FNIII}}$		97	220 ± 44	0.6	2.60	[61]
$^{2}\mathrm{FNIII}_{\mathrm{1FNIII}}$		91	220 ± 44	0.6	2.60	[61]
$^{1}\mathrm{FNIII}_{\mathrm{I27}}$	1oww	97	120	0.6	2.60	[61]
$^{10}\mathrm{FNIII}$	1fnf	~92	74 ± 20	0.6	2.60	[49, 61]
$^{11}\mathrm{FNIII}$		94	74	0.6	2.60.40.30	[63]
$^{12}\mathrm{FNIII}_{\mathrm{13FNIII}}$	1fnh	92	124 ± 18	0.6	2.60.40.30	[61]
$^{13}\mathrm{FNIII}_{\mathrm{I27}}$	1fnh	89	89 ± 18	0.6	2.60.40.30	[61]
$^{2-14}\mathrm{FNIII}$		~90	145	0.6	2.60	[61]
AFN, 60-A65		~99	180	0.5	2.60	[67]
ConFN, I48–I54		~100	200	0.5	2.60	[67]
Other FN typeIII						
C-tenascin*15		~91	137 ± 12	0.3–0.5	2.60.40.30	[60, 61]
TNFN	1ten	~91	113	0.2–0.6	2.60.40.30	[67]
TNFNAll*15		~90	138 ± 50	0.3–0.5	2.60.40.30	[60]
TNFNA-D*7		~91	138 ± 50	0.3–0.5	2.60.40.30	[60]
Spectrins						
Monomers						
Native	1u4q	~106	30 ± 5	0.3	1.20.58.60	[52, 68]
α-spectrin R16	1aj3	~106	54 + 20	0.3	1.20.58.60	[51]
α-spectrin$_{13-18,18-21}$	1u4q	~106	26 + 15	0.3	1.20.58.60	[52, 54, 68]
β-spectrin$_{1-4}$	1s35	~106	27–13	0.3		[52–54]
α-actin$_{1-4}$	1hci	~106	38	0.3	1.20.58.60	[54, 68]
Dimers						
α, β-spectrin*			54 ± 30			[54]
α-actin$_{1-4}$			50 ± 20		1.20.58.60	[54]

Table 1. (Continued.)

Protein	PDB	N	$F_{\max}$ (pN)	v_p (μm s^{-1})	CATH	References
Other						
PEVK$_{I27}$		186	<20	0.4	4	[47, 48, 56, 74]
N2$_{I27}$		572	<20		4	[47, 78]
Ribonuclease H	1rnh	155	19		3.30.420.10	[39]
E2lip3$_{I27}$	1qjo	80	15 ± 10	0.7	2.40.50.100	[34]
E2lip3(N-41)$_{I27}$	1qjo	40	177 ± 3	0.7	2.40.50.100	[34]
Ankyrin*1	1n11	33	37		3	[19, 20]
Ankyrin*24	1n11	792	450		3	[19]
Mel-CAM		∼100	30		2.60.40.10	[129]
Mel-CAM^{+DTT}		∼100	41		2.60.40.10	[129]
VACM1^{+DTT}	1vcs	∼98	40		2.60.40.10	[130]
$^{1-5}$DdFLN	1wlh	100	40–100		2.60	[71]
$^{4}_{I27-I30}$FLN$_{I31-I34}$	1ksr	100	45 ± 20	0.2–0.4	2.60	[71–73]
C2A	1dqv	127	60		2.60.40.180	[18]
T4 lysozyme	1b6i	164	64		1.10.530.40	[79]
Barnase$_{I27}$	1bnr	108	70	0.1–0.5	3.10.450.30	[33]
Calmodulin	1cfc	148	80		1.10.238.10	[18]
Ubiquitin(48-C)	1ubq	28	85 ± 20	0.3	3.10.20.90	[16, 17]
Ubiquitin(N-C)	1ubq	76	203	0.2–0.4	3.10.20.90	[16, 17]
Ubiquitin	1ubq	76	230 ± 34	1	3.10.20.90	[17]
GFP$_{DdFLN}$	1b9c	238	104 ± 40		2.40.155.10	[40]
GFP$_{Ig}$	1b9c	238	104 ± 40		2.40.155.10	[40]
GFP(3–212)	1emb	219	130 ± 30		2.40.155.10	[41]
GFP(132–212)	1emb	80	120 ± 30		2.40.155.10	[41]
GFP(3-132)	1emb	129	350 ± 40		2.40.155.10	[41]
Protein L	1hz6	67	152 ± 5	0.7	3.10.20.10	[32]
Spider silk pS(S4+1)		∼608	176 ± 73	0.2–1.5	2	[65]
Filamin A*24		96	50–220	0.37	2.60	[43]
Bovine	1v9e	259	1100		3.10.200.10	[36]
Bacteriorhodopsin	1at9	231	350		1.20.1070.10	[62, 44]
Biotin–streptavidin			350			[58]
Proteomer out of hexagonally packed intermediate layer			312 ± 43			[58]
A-macroglobulin			750			[59]
β-fibrils						[45]
DNA			13			[35, 36]
P5abc three helix junction			19			[55]

Fortunately, the number of proteins that are listed in the Protein Data Bank (PDB) [99] is sufficiently large to generate meaningful statistics. Currently, there are more than 29 385 entries in the PDB (we downloaded all structures that had been deposited by 26 July 2005) but many of these correspond to complexes with nucleic acids or with other proteins and some of them correspond to nucleic acids. Figure 2 shows the distribution of the values of N for all proteins, whether in a complex or not, together with our previous assessment [100] of it that was based on 500 proteins, selected randomly from the Swiss-Prot data base. The complexes of proteins are discarded in our stretching studies unless they were clearly resolvable into chains and then the first listed chain was selected. Of around 15 000 proteins that are left, 7510 comprise between 40 and 150 amino acids. The most probable number of amino acids in a protein is found to be

Table 2. Summary of all-atom simulation results on $F_{\max}$.

Protein	PDB	N	$F_{\max}$ (pN)	v_p (Å ps^{-1})	References
Immunoglobulins					
I1 oxidized	1gcg	97	2397	0.5	[82]
I1 reduced	1gcg	97	2090	0.5	[82]
I27	1tit	89	2479	0.5	[82, 17]
I27	1tit	89	2040	0.5	[15, 86–88, 84]
I28		93	2082	0.5	[15]
Fibronectins type III					
^{1}FNIII	1oww	97	1500	0.01	[81]
^{2}FNIII		91	1600	0.01	[81]
^{7}FNIII	1fnf	93	1638	0.5	[80, 88]
^{9}FNIII	1fnf	91	2000	0.1	[80, 84, 85]
^{10}FNIII	1fnf	94	1580	0.5	[80, 84, 88, 85]
^{9}FNIII	1fnf	91	^{9}F < ^{10}F	0.1	[92, 93]
Other					
Ubiquitin (N-C)	1ubq	76	2000	0.1	[16, 90]
Ubiquitin (48-C)	1ubq	28	1200	0.1	[16, 90]
Bovine	1v9e	259	3000	0.5	[36]
Barnase	1bnr	108	500	0.01	[33]
Cad1	1edh	211	1850	0.5	[88]
Cad2	1edh	211	1970	0.5	[88]
Cell adhesion VCAM1	1vsc	89	2050	0.5	[88]
Cell adhesion VCAM2	1vsc	108	1620	0.5	[88]
T4 lysozyme	1b6i	164	75	10^4	[85]
Cytochrome C6 cc6	1cyi	89	No peak	0.5	[88]
Binding protein IGB	1bdd	60	No peak	0.5	[88]
Synaptotagmin (c2)	1rsy	125	No peak	0.5	[88]

close to 120 (figure 2) so the set of 7510 shorter proteins, denoted by S7510, covers the typical sizes. This set is explored fully whereas larger proteins are studied within a set of about 239 proteins extracted from those that were used in threading studies [101] and had no gaps in their structure assignment. The larger proteins are usually multi-domained. Various sets of proteins that are considered in this paper will be identified by the symbol S followed by the number of entries that they contain.

Here we focus on constant speed simulations and find that set S7510 is rich enough to review the types of force–displacement patterns and to determine the nature of the system size dependence. We find that the longer the protein the more likely it is to have a larger value of $F_{\max}$, even though the spread in $F_{\max}$ for a given N depends on N rather weakly. In particular, the strongest protein that we have identified, a streptokinase with the PDB code 1c4p, corresponds to $N = 137$ and has a single domain.

It turns out that only 3813 proteins in set S7510 have a listed CATH-based [102] structure assignment in terms of the structural class, architecture, topology and homology. Thus a search for correlations between dynamics and structure is restricted to this subset—S3813. We find that proteins belonging to the same CATH-based structure index may have distinct dynamical behaviour which points to inadequacies in this classification scheme. Our survey indicates that the α-class proteins tend to show weak force when resisting stretching. However, we have found three cases for which we predict the force to be comparable to that for titin.

We then focus on set S137 of the strongest proteins which are predicted to be in the top 1.8% of the S7510 set in terms of the value of $F_{\max}$. We find that these strong proteins are

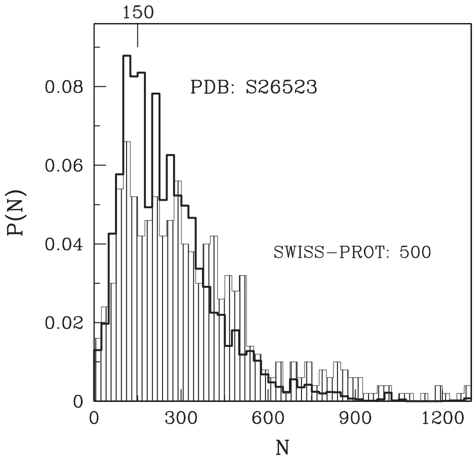

Figure 2. Distribution of sequence lengths across proteins. The shaded histogram is based on 500 proteins picked randomly from the Swiss-Prot data base [100]. The thick-line histogram is based on 26 523 proteins listed in the PDB.

one-domain proteins and belong only to the β and α/β classes. Moreover these strong proteins represent only seven types of architecture, nine types of topology and 17 homological families. Our systematic studies yield several notable predictions. For instance, even though most mechanical clamps consist of two long parallel β strands that undergo shearing, there are other possibilities as well. Mechanical clamps do not need to arise from shearing of β-strands and, even if they do, none of these β strands need to be located near the terminal amino acids. There are some mechanical clamps which involve anti-parallel β strands and we have found one case in which the clamp has a topology of a box with helical walls. We observe that the strongest proteins have mechanical clamps which are reinforced by their surrounding native environment.

Generally, the strength of a protein is found be a sensitive function of the native structure. Thus, in particular, there are proteins, like 1pga and 1p7e, such that a mutation in a few amino acids (leading to structures 1q10 and 1mpe, respectively) may result in an even three-fold reduction in the value of $F_{\max}$.

We first present results obtained within the simplest C$^{\alpha}$-based Go-like model. Later on, we show that the results do not change much when a model with side groups is considered but they provide better insights into the mechanisms of the force clamps. We also comment on the role of contacts made by disulfide bridges.

It should be noted that protein stretching may be sensitive to the location of the application of the stretching force. Throughout this review, we generally consider stretching by the terminal amino acids. Other relevant choices, say pulling by the lysins at various locations, could be physically meaningful but the combinatorics involved would expand the number of cases prohibitively. We consider these other choices only in cases when relevant experimental data are available and then the amino acids that are pulled at are indicated in brackets by their sequence location. For example, 1ubq(N,K43) means ubiquitin pulled by the N-terminus and lysine-43. The choice of location of the application of the stretching force is usually implemented by producing a specific linkage of the chain of domains. The linkage dependence results from various effective directions of the force that disrupts the force clamp. The nature of the force clamp itself may change as well. For green fluorescent protein (GFP), for instance, $F_{\max}$ may vary between 104 and 548 pN [10] depending on the linkage, and for ubiquitin, two different linkages yield 85 and 204 pN [9].

2. The model

Coordinates of atoms in a structure of a protein are downloaded from the PDB [99]. If a given PDB code is represented by many structures, as is often the case in the NMR-derived results, we consider only the first of these. If there are several chains corresponding to a code, we take only the first listed chain. We also discard files in which there are sequential gaps in the determined structure. However, we make exceptions when a protein is studied experimentally and yet its structure contains small gaps, like three amino acids long, as in the case of GFP. In this situation, we use the program called Bioshell [103] to repair the structure. The downloaded structure is considered to be the ground state conformation of the Go-like model [94, 95]. It should be noted, however, that this conformation is determined experimentally, usually at room temperature.

Our approach is outlined in [97, 104–107] and its first step is determination of the native contacts between amino acids pair by pair. The presence of a contact is decided based on checking for overlaps between effective atoms according to a procedure proposed by Tsai *et al* [108]. This procedure is based on representing heavy atoms by spheres with radii which are equal to the van der Waals radii of the atoms multiplied by a factor of 1.24 to account for attractive interactions. In particular, the overlaps may indicate the existence of a contact between amino acids i and $i + 2$, where i is a label along the sequence. In this survey, we treat the $i, i + 2$ native contacts to be similar in strength to other native contacts. In reality, the $i, i + 2$ contacts usually correspond to van der Waals interactions which are much weaker than hydrogen bonds. They arise primarily in α-helices.

The potential energy of the system is given by

$$E_{\mathrm{p}}(\{\mathbf{r}_i\}) = V^{\mathrm{BB}} + V^{\mathrm{NAT}} + V^{\mathrm{NON}} + V^{\mathrm{CHIR}}. \tag{1}$$

The first term, V^{BB}, is the harmonic potential

$$V^{\mathrm{BB}} = \sum_{i=1}^{N-1} \tfrac{1}{2} k (r_{i,i+1} - d_0)^2, \tag{2}$$

which tethers consecutive beads at the equilibrium bond length, d_0, of 3.8 Å. Here, $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ is the distance between the beads and $k = 100\ \epsilon\ \text{Å}^{-2}$, where ϵ is defined below.

The native contacts are described by the Lennard-Jones potentials:

$$V^{\mathrm{NAT}} = \sum_{ij} 4\epsilon \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right]. \tag{3}$$

The length parameters, σ_{ij}, in these potentials are selected so that the minima of the potentials agree with the experimentally determined distances between the C^{α} atoms in a contact. The non-native contacts correspond to a repulsive core of $\sigma = 4$ Å. The energy parameter, ϵ, is taken to be uniform and its value should be in the range 800–2300 K since it corresponds to an effective average of all non-covalent interactions in proteins. Our previous simulations of folding [106, 107] were optimal with the dimensionless temperature $\tilde{T} = k_{\mathrm{B}}T/\epsilon$ of order 0.3 which corresponds to room temperature if ϵ is around 900 K (k_{B} is the Boltzmann constant and T is temperature). Additionally, the simulated stretching curves were similar to experimental curves at $\tilde{T} = 0.3$ [107, 109]. With $\epsilon = 900$ K, the unit of force used in this paper, $\epsilon\ \text{Å}^{-1}$, corresponds to 120 pN. This choice also yields the correct magnitude of the force peak in titin [107] and ubiquitin [110] at room temperature. Therefore, 900 K should be considered to be a representative value of ϵ and we perform the calculations at $k_{\mathrm{B}}T/\epsilon = 0.3$.

When the temperature is raised, the thermal fluctuations aid in the unravelling process. Generally, an increase in T lowers force maxima and makes them occur earlier during the stretching [44, 107, 109, 110]. This point has also been discussed by Hyeon and Thirumalai [111]. At sufficiently high temperature [112], the peaks in the F–d patterns disappear altogether. In a tandem arrangement of protein domains, low-temperature stretching is mostly serial in nature, i.e. it takes place domain by domain, but eventually it becomes more and more parallel [107] as the peak forces become less and less resolvable.

In this survey we treat the disulfide bonds between the cysteines on the same footing as all other contacts, even though such bonds are much stronger and cannot rupture under the usual stretching conditions. Once the set of the strongest proteins is identified, we re-examine these proteins to determine their behaviour when the disulfide bonds are not allowed to break.

The model also contains a four-body chirality term that favours the native sense of chirality. The chirality term chosen in [106] has had the form

$$V^{\mathrm{CHIR}} = \sum_{i=2}^{N-2} \tfrac{1}{2}\kappa\epsilon C_i^2 \Theta(-C_i C_i^{\mathrm{NAT}}), \tag{4}$$

where Θ is the step function (1 for the positive argument and 0 otherwise), but here we adopt a simpler and numerically more efficient expression [113]

$$V_1^{\mathrm{CHIR}} = \sum_{i=2}^{N-2} \tfrac{1}{2}\kappa\epsilon (C_i - C_i^{\mathrm{NAT}})^2, \tag{5}$$

where

$$C_i = \frac{(\mathbf{w}_{i-1} \times \mathbf{w}_i) \cdot \mathbf{w}_{i+1}}{d_0^3}, \tag{6}$$

and C_i^{NAT} is the chirality of residue i in the native conformation. Here, $\mathbf{w}_i = \mathbf{r}_{i+1} - \mathbf{r}_i$. A positive C_i corresponds to right-handed chirality. Otherwise the chirality is left-handed. The values of C_i are essentially between -1 and $+1$. The parameter κ is taken to be equal to 1. The chirality potential acts very much like the bond and dihedral angle potentials considered, for example, in [96] and [114] but we have found it to be computationally more convenient. V^{CHIR} enhances the stability of the model and monitoring of chirality is a part of checking the folding criterion [113]. The full model considered here is kinetically equivalent [115] to the model with the 10–12 contact potentials considered by Clementi *et al* [116] which was demonstrated to have the two-state kinetics of folding for simple proteins. The chirality term plays an important kinetic role when studying folding but is of less relevance for stretching.

In our stretching simulations, both ends of the protein are attached to harmonic springs of elastic constant $k = 0.12\,\epsilon$ Å^{-2} which is close to the values corresponding to the elasticity of experimental cantilevers (this value corresponds to the 'soft' spring case of [107]; the values of $F_{\max}$ do have a certain dependence on the elasticity of the pulling spring). The free end of one of the two springs is anchored while the free end of the second spring is pulled at a constant speed, v_{p}, along the initial end-to-end position vector. We consider two values of v_{p}: 0.005 and 0.05 Å$/\tau$, where $\tau = \sqrt{m\sigma^2/\epsilon}$ is the characteristic time for the Lennard-Jones potentials. Here, $\sigma = 5$ Å is a typical value of σ_{ij} and m is the average mass of the amino acids. The smaller value of v_{p} corresponds to what will be called 'slow' pulling and the larger one to 'fast' pulling.

For an average mass of an amino acid of about 118 Da the value of τ would be $\approx$3 ps. However, it has been argued [117, 96] that the particle that effectively represents an amino acid should have a substantially more extended size than a single atom and is meant to move in an environment with a large friction which overrules inertia-based estimates of the characteristic time. Thus the 'bare' unit of time should be 'renormalized' to an over-damped characteristic timescale [118] τ_{H} of order 3 ns which yields the right order of magnitude for the folding time of

α-helices. A more precise argument based on the Peclet number, *Pe*, for ubiquitin stretched by uniform fluid flows [119] ($Pe = UR_g/D$, where U is the characteristic speed, R_g is the radius of gyration and D is the diffusion coefficient of the protein) suggests τ to be of the order of 0.25 ns. Therefore the slow pulling speed corresponds to about 10^6 nm s^{-1} which is merely two orders of magnitude faster than the top experimental speeds. The force–displacement pattern has been found to be very close to that corresponding to the still smaller velocities of 0.0005 and 0.000 05 Å$/\tau$ [110, 107], so using 0.005 Å$/\tau$ in the survey seems justified.

The thermal fluctuations away from the native state are introduced by means of the Langevin noise, i.e. by random Gaussian forces together with a velocity-dependent damping. This noise mimics the random effects of the solvent and provides thermostatting. The temperature T controls structural fluctuations in the model protein including those which are present under room temperature even though the ground state of our model corresponds to the native state of the protein that was determined at room temperature. In order to account for a finite resolution within which the thermal effects are observed to affect the force–displacement relationship we average the forces over a pulling distance of 0.5 Å.

An equation of motion for each C$^\alpha$ reads

$$m\ddot{\vec{r}} = -\gamma\dot{\vec{r}} + \vec{F}_c + \vec{\Gamma}. \tag{7}$$

F_c is the net force due to the molecular potentials. The damping constant γ is taken to be equal to $2m/\tau$ and the dispersion of the random forces is equal to $\sqrt{2\gamma k_B T}$. This choice of γ corresponds to a situation in which the inertial effects are negligible [106] but the damping action is not yet as strong as in water. The value corresponding to water was estimated to be about 25 times larger [117, 96]. Increasing γ tenfold results in a corresponding tenfold increase in the folding time [96, 97, 106] but it has only a minor effect of the F–d patterns [107]. Thus when studying folding, the simulated folding times should be multiplied by 25 to get agreement with experimental timescales [97]. On the other hand, no adjustment in the timescales is needed in the stretching studies. The equations of motion are solved by a fifth-order Gear predictor–corrector scheme [120] with a time step of 0.005τ. In some places in this paper we shall use the shorthand notation $\tilde{F}$ for the reduced force F Å$/\epsilon$.

The model presented here can also be used for studies of protein stretching by fluid flows [119] even though it comes with no explicit description of the solvent.

In order to quantify the pathways of unfolding, we make use of the so-called scenario diagrams [107]. A scenario diagram shows distance d_u at which a given native contact is broken for the last time. Contacts in the scenario diagrams are identified by the sequential distance $|j - i|$. A contact is said to be broken if the distance between amino acids i and j exceeds $1.5\sigma_{ij}$. An accumulation of many contact unfolding events at a value of d_u indicates the emergence of a force peak.

3. Validation of the approach

Figure 3 shows a cross plot between the experimentally derived value of F_{max} and its theoretical determination within our version of the C$^\alpha$-based Go model. The proteins shown are those for which the structure coordinates are fully available or require minor repairs, as in the case of the GFP (with the code 1emb). Some proteins have several structure assignments. For instance, protein L has three assignments, 1hz5, 1hz6 and 2ptl, whereas barnase has two, 1bni and 1bnr. In this case, we average the results over the structures (the average results are indicated by the symbols L and B, respectively). We observe that, generally, there is a robust degree of correlation corresponding to a linearly growing trend. The Pearson coefficient is 0.85 (removing the $i, i+2$ contacts would increase it to 0.90). This trend is indicated by the solid

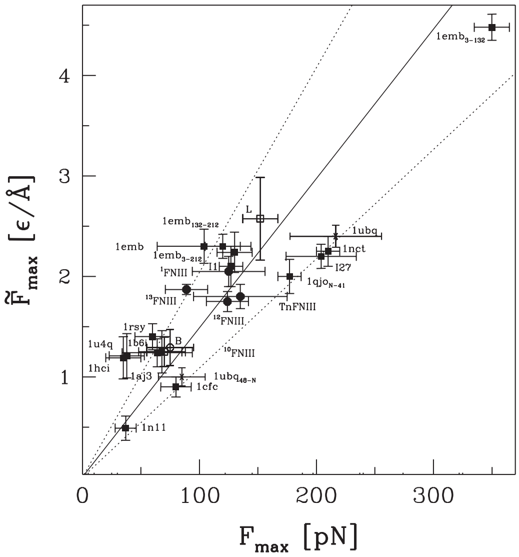

Figure 3. $F_{\max}$ predicted by our theoretical model versus the corresponding experimental results. The proteins are identified, in most cases, by the PDB codes and the data points are shown as square symbols. Special cases have different notations and different data symbols. Asterisks correspond to ubiquitin as pulled by the termini (the larger force) or by the K48 and terminus C (the lower force). Symbol L denotes protein L (averaged over 2ptl and 1hz5) and B denotes barnase (averaged over 1bni and 1bnr). Circles correspond to fibronectins: black to ^{1}FNIII, TnFNIII, ^{11}FNIII, ^{12}FNIII and ^{13}FNIII whereas white to ^{10}FNIII (the latter is averaged over 1fnf, 1ttf and 1ttg). The domains of titin are indicated by the biological names, like I27, I1. Three proteins were not included in the figure: 1ksr (the fourth domain of FLN), 1rnh (ribonuclease H [39]) and 1qjo when pulled by the termini (E2lip3). The titin-like 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 [71]. The contact map of the second is unstable against small changes in the definition of the contact. There are two reasons to mistrust 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 [34]. The solid line has a slope of 0.0122.

line in figure 3. The line is constrained to go through the origin, because if there was an agent which could weaken the contact strengths to zero gradually then in both the experimental and theoretical systems the peak force would have to approach zero. The dotted lines are effective error bars for the linear fit.

Though figure 3 provides encouragement for a more extensive use of the model, it is necessary to point out that extracting a value of $F_{\max}$ for a protein from experimental data is often complicated due to the fact the F–d curves are usually determined for a number of globular modules that are connected in tandem. These modules need not be identical and artificial linkers are often employed for tethering. These circumstances affect the reading of the data. Furthermore, the interpretation of the data involves assumptions such as the serial character of unwinding. Experimental [18] and theoretical [124, 109] studies indicate that a module-by-module unwinding need not hold in general, even if the modules are identical such as in polycalmodulin. The presence or absence of seriality in the unwinding process depends on the protein and on the temperature. The case of fibronectin type III (FNIII) provides a good illustration of the problems with data interpretations [61]: a four-fold repeat of the ^{1}FNIII-I27 linkage, where ^{1}FNIII denotes the first domain of FNIII consisting of around 97 amino acids, appears to unfold in two ways. The first way involves unwinding of the full ^{1}FNIII at 20 pN.

The second way involves a prior formation of an intermediate state in which 53 amino acids are in a native-like structure. Unwinding of the intermediate state leads to a force of 120 pN. Moreover, when a six-fold linkage of the pair ^{1}FNIII and ^{2}FNIII (the second domain of FNIII) is stretched, the force attributed to ^{1}FNIII changes to 220 pN.

In order to provide a more detailed discussion of the comparison of the model predictions to the experimental data it is convenient to divide the proteins into groups. The division is based on the value of F_{max}: soft (F_{max} smaller than 60 pN), soft–medium (between 60 and 100 pN), medium (between 100 and 170 pN) and strong (larger than 170 pN). FNIII spans three of these groups, since the forces range between 20 and 220 pN, and this case will be discussed separately at the end of the review.

The best agreement with the linear trend in figure 3 is observed for the group of strong proteins, especially GFP when pulled at (3–132). The group also includes ubiquitin, I27 (1tit) and representative proteins from tandem linkages of proteins I54–I59 (1nct), E2lip3(N-41) (1qjo) and protein L. We could not study the very strong 24-subunit ankyrin due to the lack of a corresponding PDB structure (a single subunit consists of 33 amino acids). It is clear, however, that its high force is due to the emergence of a natural horseshoe-like structure. Smaller linkages of the subunits do not form such a structure and generate substantially smaller forces both in experiment [19, 20] and in our simulations. In particular, our simulations yield F_{max} comparable to that of the I27 domain of titin for a 12-subunit system.

Proteins with the medium force include fibronectin ^{1}FN (the first domain of 1oww, stretched alone), fibronectin ^{12}FnIII (the first domain of 1fnh), the ^{3}TnFNIII domain of FNIII from tenascin (1ten), the I1 domain of titin (1g1c) and GFP pulled by the termini and at locations (3–212) and (132–212). It should be noted that GFP is stretched experimentally inside a tandem of other proteins such as $^{1-5}$DdFLN or $^{27-34}$IG. This circumstance is likely to affect the geometry of pulling, as defined by the positioning of the direction of the end-to-end vector, the rupturing process itself and the interpretation of allocation of force to specific domains. In the case of the I1 domain of titin (with the PDB code of 1gc1) we obtained a force that is somewhat higher than for I27 instead of being somewhat lower as observed experimentally. This prediction is not changed when various ways of linking I1 with other domains of titin are studied [109]. Nevertheless, they are still rather comparable.

The soft–medium group of proteins includes ubiquitin(48-C) (1ubq), T4 lysozyme (1b6i), barnase, the first C2 domain of the synaptotagmin I (C2A–1rsy), calmodulin and a few domains of fibronectin (the 10th and the 13th). Our theoretical results show a good consistency with the overall trend for most of them. The exceptions are T4 lysozyme and the 13th domain of fibronectin which have higher model forces than expected. However, individual structures for the 13th domain of FNIII are not known. We shall discuss this case later.

Somewhat bigger concerns with the validation of our approach arise when dealing with the soft group of proteins: the spectrin family (1aj3, 1u4q, 1hci) and one subunit of ankyrin (1n11), Experimentally, these proteins are studied in tandem linkages and our model yields lower peak forces when stretching multiple homorepeats of the weak proteins than when considering single units. One subunit of ankyrin has $\tilde{F}_{\mathrm{max}}$ of about 0.5 which agrees with the general trend rather well and the estimated experimental value is about 37 pN [19] or 50 ± 20 pN [20].

The spectrin family contains various subfamilies (α-spectrin, α-actin and dystrophin) and each spectrin protein contains repeated sequences of approximately 106 amino acids with low sequence identity. We found that α-spectrins, β-spectrins and α-actins are represented in the PDB by at least four, one and two structures, respectively. For all of them, we obtain the peak force which is about two times lower than for the I27 domain of titin, instead of the experimentally observed factor of five. However, a direct comparison between the model and experiment is difficult because the experimental stretching has been accomplished for heterolinkages [52–54] and, in addition, there seems to be a higher than usual sensitivity to the pulling speed. A direct comparison is then sensible in the case of one subunit of α-spectrin R16 (the PDB code is 1aj3) for which homolinkages have been used [51]. In this case, the experimental values of $F_{\rm max}$ varied between 30 and 180 pN and the most likely number seems to be around 65 pN. When we use the I27 domain of titin to calibrate the theoretical values of the force, then the prediction is higher than 65 pN. In this case, however, the unstretched long helical pieces provide a relatively high background from which the force peaks emerge. However, if we were to subtract the background in the model calculations, then the predicted value would agree with the experimental one much better than shown in figure 3. We have checked that this occasional background arises when the helical $i, i+2$ contacts in helices are treated on the same footing as the $i, i+3$ and $i, i+4$ contacts. For long helices, this procedure results in a fast change of the conformation to the 3–10 type helices on stretching. These 3–10 type helices stay unravelled until the very end of the process and generate the unphysical background force which should be subtracted in this model.

A different kind of validation of the approach is provided by considering the pulling of bacteriorhodopsin out of a membrane. A Go-like approach to this problem [122] yields a complex F–d pattern (figure 10 in [122]) which is remarkably similar to that obtained experimentally [44]. It should be pointed out, however, that obtaining the agreement in the magnitude of the force has required a fourfold reduction in the value of the effective energy parameter ϵ compared to what is used to model proteins not encased by a membrane.

4. The methods of the survey

Figure 4 shows simulational examples of the F–d curves for single domains of ubiquitin (1ubq) and integrin (1ido) for several values of $v_{\rm p}$. It has been established experimentally [17] and theoretically [107] that $F_{\rm max}$ usually varies with $v_{\rm p}$ merely logarithmically so the exact choice of $v_{\rm p}$ is not very crucial as long as it is not too far off the experimental values. The F–d curve itself, however, is more sensitive to $v_{\rm p}$ and, in addition, it varies somewhat from trajectory to trajectory. We have found that the smaller the value of $v_{\rm p}$, the lesser the dependence on the string of random numbers in the Langevin noise. The smallest $v_{\rm p}$ that we could use to survey the PDB is $0.005\sigma/\tau$—the case of 'slow' pulling. Furthermore, the task could only be accomplished in most cases by first locating the major peak force roughly, by making a fast run—with $v_{\rm p}$ of 0.05 σ/τ—and then by repeating the stretching simulation at the slow speed but without going too far beyond the major peak area.

The structured F–d pattern seen in figure 4 is a result of rupture of various single or multiple contacts in the course of time. The bigger the cluster of bonds that distort simultaneously, the bigger the maximum in the force. The simulations allow us to construct scenarios of unwinding by recording which bond is broken at what time. Once all contacts are ruptured, the C^{α} chain is nearly fully unravelled and, from now on, F grows indefinitely due to stretching of the covalent peptide bonds, as illustrated in the top panels of figure 4. The time evolution ceases once the fully unravelled stage is reached.

For ubiquitin the major peak is followed by a set of smaller peaks (their number depends on the trajectory and speed), whereas for integrin the major peak is preceded by three ascending peaks and then followed by a number of minor peaks. The number of minor after-peaks may vary between trajectories. (The absence of peaks before the major peak in ubiquitin leads to a different behaviour, compared to integrin, in a force-clamp situation [121], i.e. when stretching is accomplished under the conditions of a constant pulling force as opposed to constant speed.) Since the number of peaks before the major maximum seems to be robust, when independent runs are considered, but that of the descending peaks is not, any F–d trace can be meaningfully

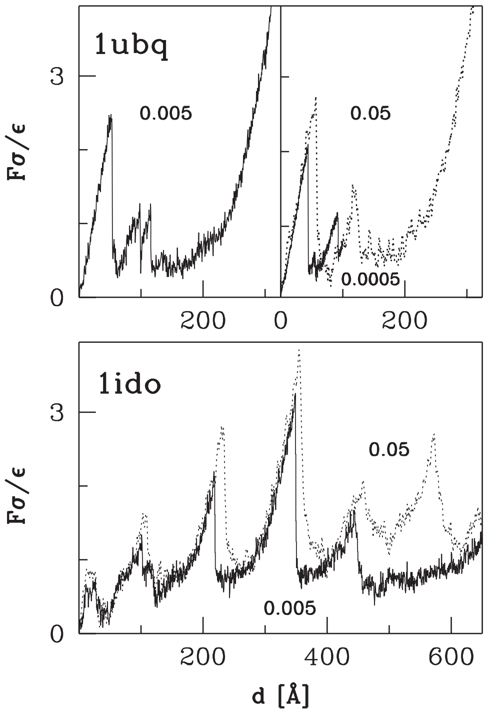

Figure 4. Examples of the force–displacement curves resulting in our model. The top two panels refer to ubiquitin: the one on the left is for $v_p = 0.005\ \sigma/\tau$ ($\tilde{F}_{\max}$ is near 2.45); the one on the right is for an order of magnitude smaller and larger speeds as indicated. The bottom panel compares results for slow pulling (solid line; $\tilde{F}_{\max}$ near 3.3) and fast pulling (dotted line) for integrin.

ascribed to one the following types: M, BnM, BnMA, MA. Here, M (for a major peak) denotes a trace with just one force peak; BnM denotes a situation in which there are n force peaks before the major force peak; BnMA is like BnM but with a number of peaks after the global maximum; finally MA means no peaks before the major maximum and some peaks after it. In our example, 1ubq and 1ido belong to types MA and B3MA respectively.

Even though the precise value of $F_{\max}$ depends on v_p, it is still meaningful to make comparisons across proteins for the same (slow) speed. The survey is restricted to single trajectory calculations in the fast-then-slow mode, since the differences between proteins are usually much more significant than between individual trajectories. However, in hard to resolve cases, multiple trajectories are considered.

5. Results

5.1. Results for proteins with $N \leqslant 150$: set S7510

We first consider all PDB proteins that do not belong to complexes and contain no more than 150 and no less than 40 amino acids—the set S7510. Figure 5 shows the distribution of values of $F_{\max}$ that were obtained during the molecular dynamics simulations within our Go-like model. The values range from 0 to 5.44 and the most probable value is 1.5. A typical error bar due to variations between individual trajectories is of the order of 0.1–0.15. For the I27 domain of titin, the peak force is 2.2. Even though titin is predicted to be almost 50% tougher than a

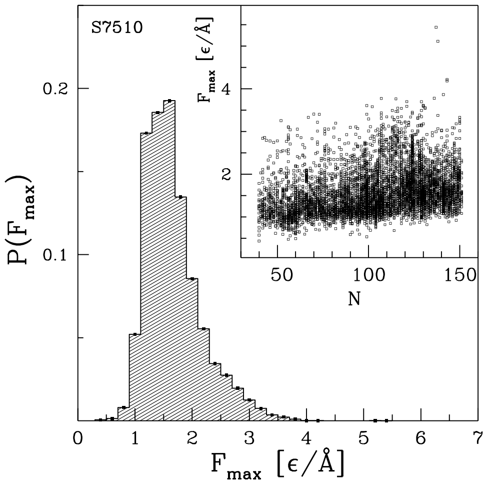

Figure 5. Probability distribution of the values of F_{max} of short proteins from the set S7510. The square data symbols at 5.4, 5.2, 4.2, 4.1 and 4.0 correspond to single entries. The inset shows all values of F_{max} that were obtained for a given sequence length.

typical protein, its strength is only half of the maximum value of $\tilde{F}_{\mathrm{max}}$ that was obtained within S7510. In that sense, titin is strong but not very strong.

The inset of figure 5 shows values of F_{max} that were obtained for each value of N within the studied range. The top five strongest proteins correspond to N around 140, but other than that the range of possible force values is fairly uniform across N. Thus large and small proteins can be comparably weak or strong. Nevertheless, the bigger the N, the bigger the probability that a protein is strong. If one averages over all entries corresponding to a given N, then one gets a growing trend with N, as shown in figure 6. A linear fit yields a slope of 0.0068 ± 0.0018 on the linear scale. Thus a large protein is more likely to give rise to a large F_{max} than a small protein.

The set of the 137 strongest proteins with $\tilde{F}_{\mathrm{max}} > 2.9$, i.e. belonging to the tail of the distribution of the forces, is presented in table 3 together with the entries for titin, ubiquitin and integrin which serve as points of reference. The table displays the values of N, $\tilde{F}_{\mathrm{max}}$, the type of force pattern and the symbol of structural CATH classification if available. It also shows characteristic values of the end-to-end distance L: L_{n} is the native value of L, L_{m} corresponds to the location of the tallest force peak, and L_{f} corresponds to full extension of $(N-1)3.8$ Å. We define a parameter $\lambda = (L_{\mathrm{m}} - L_{\mathrm{n}})/(L_{\mathrm{f}} - L_{\mathrm{n}})$ which describes location of the maximum force as a fraction of $L_{\mathrm{f}} - L_{\mathrm{n}}$. Figure 7 shows that the distribution of the values of λ is peaked around 10% for the set of strong proteins whereas it is rather flat generally. This indicates that large peak forces usually come with rupture events near the termini as happens in titin [84, 107].

We find that 71% of the strong proteins have the F–d (or F–L) pattern which is of the MA type, 20% the BnMA type (in most cases $n = 1$) and 7% the BM type. Only three proteins, including the top two strongest, have patterns with just one force maximum. The F–d plots corresponding to the top four strongest proteins are shown in figure 8.

5.2. Results for proteins with $150 < N \leqslant 851$*: set S239*

We now consider set S239 which comprises three subsets with low sequence homology and covering many different three-dimensional folds out of which only those with $N > 150$ and

Table 3. The top 137 strongest short proteins (with $N \leqslant 150$) as predicted by the Go-like model used in this paper. The symbols are explained in the text. At the end, three weaker proteins are added as a reference. The ordering of proteins corresponding to the same value of $\tilde{F}_{\max}$ is arbitrary. Inclusion of the disulfide bonds in the model (see section on the disulfide bonds) removes 1lsl from the list of strong proteins and advances the ranking of 1rbj, and especially 1rnz. The latter becomes fifth ranked on correcting for the effects of the disulfide bonds. The asterisk indicates proteins for which the disulfide bond would rupture before or at the major peak in the standard Go-like model.

Rank	PDB	N	$F_{\max}$ (ϵ Å^{-1})	L_n (Å)	L_m (Å)	L_f (Å)	Pattern	CATH
1	1c4p	137	5.4	50.4	140.8	516.8	M	3.10.20.180
2	1qqr	138	5.2	52.3	144.7	520.6	M	3.10.20.180
3	1g1k	143	4.2	43.5	83.7	539.6	MA	2.60.40.680
4	1aoh	147	4.0	34.0	71.1	554.8	MA	2.60.40.680
5	1ssn	136	3.8	9.5	187.1	513.0	B1M	3.10.20.130
6	1ie5*	107	3.8	51.5	102.2	402.8	MA	2.60.40.10
7	1c76	136	3.7	29.0	103.5	513.0	MA	3.10.20.130
8	1ppx	129	3.7	36.0	224.9	486.4	B2M	3.90.79.10
9	1c77	136	3.6	27.2	144.1	513.0	MA	3.10.20.130
10	1c79	136	3.6	27.2	140.3	513.0	MA	3.10.20.130
11	1yn4	99	3.6	35.4	62.8	372.4	MA	
12	1c78	136	3.6	27.2	140.3	513.0	MA	3.10.20.130
13	2sak	121	3.6	31.9	108.5	456.0	MA	3.10.20.130
14	1v5o	102	3.6	63.8	127.4	383.8	MA	
15	1sp0	131	3.6	40.6	117.1	494.0	MA	
16	1so9	131	3.5	40.6	122.5	494.0	MA	
17	1sn0	130	3.5	21.5	124.6	490.2	MA	2.60.40.180
18	1oo2	119	3.5	12.9	129.1	448.4	MA	2.60.40.180
19	1i3v	129	3.5	40.4	64.5	486.4	MA	2.60.40.10
20	1i9e	115	3.5	49.5	66.5	433.2	MA	2.60.40.10
21	1nam	116	3.5	35.2	47.5	437.0	MA	2.60.40.10
22	1eaj	126	3.5	41.6	61.9	475.0	MA	2.60.40.10
23	1kiq*	107	3.4	36.2	46.1	402.8	MA	2.60.40.10
24	1f5w	126	3.4	41.4	63.9	475.0	MA	2.60.40.10
25	2ncm	99	3.4	42.1	73.5	372.4	MA	2.60.40.10
26	1pgx	83	3.4	62.7	83.0	311.6	MA	3.10.20.10
27	1m94	73	3.4	27.4	36.9	273.6	MA	3.10.20.90
28	1anu	138	3.4	24.0	34.7	520.6	MA	2.60.40.680
29	1eta	127	3.4	12.6	123.8	478.8	MA	2.60.40.180
30	1kip*	107	3.4	36.3	47.3	402.8	MA	2.60.40.10
31	1sn2	130	3.4	21.6	89.2	490.2	MA	2.60.40.180
32	1tum	129	3.4	33.7	211.7	486.4	B2M	3.90.79.10
33	1nme	146	3.4	52.8	289.4	551.0	B2MA	3.40.50.1460
34	1h5b*	113	3.3	40.5	61.3	425.6	MA	2.60.40.10
35	1npu	116	3.3	40.6	54.0	437.0	MA	
36	1mvf	135	3.3	38.7	73.0	509.2	MA	2.60.40.10
37	43c9*	113	3.3	36.5	46.9	425.6	MA	2.60.40.10
38	1hz6	72	3.3	41.6	59.0	269.8	M	3.10.20.10
39	1bzd	127	3.3	11.0	124.7	478.8	MA	2.60.40.180
40	1a2y	107	3.3	37.0	49.3	402.8	MA	2.60.40.10
41	1km7	100	3.3	35.8	95.7	376.2	MA	3.10.20.90
42	1lve*	122	3.3	39.2	51.0	459.8	MA	2.60.40.10
43	1b88	114	3.3	39.0	64.5	429.4	MA	2.60.40.10
44	1eo6	117	3.2	22.1	152.4	440.8	B1MA	3.10.20.90
45	1ie4	127	3.2	15.3	97.2	478.8	MA	2.60.40.180

Table 3. (Continued.)

Rank	PDB	N	F_{max} (ϵ Å^{-1})	L_{n} (Å)	L_{m} (Å)	L_{f} (Å)	Pattern	CATH
46	1k53	72	3.2	32.5	78.1	269.8	MA	3.10.20.10
47	1kgi	127	3.2	19.2	98.7	478.8	MA	2.60.40.180
48	1oau	122	3.2	43.4	56.9	459.8	MA	2.60.40.10
49	1vhp	117	3.2	40.8	49.2	440.8	MA	2.60.40.10
50	1h8c	82	3.2	39.5	55.3	307.8	MA	3.10.20.90
51	1jf8	131	3.2	17.3	378.3	494.0	B3M	3.40.50.270
52	1jrk	156	3.2	42.2	213.8	589.0	B2MA	3.90.79.10
53	1sn5	130	3.2	21.5	89.6	490.2	MA	2.60.40.180
54	1wtl	108	3.2	40.4	50.9	406.6	MA	2.60.40.10
55	1amx	150	3.1	32.1	197.6	566.2	B1MA	2.60.40.740
56	1qd0	128	3.1	40.2	48.2	482.6	MA	2.60.40.10
57	1ufy	122	3.1	30.8	78.1	459.8	MA	3.30.1330.40
58	1mg4	113	3.1	5.9	117.9	425.6	B1MA	3.10.20.230
59	1p7e	56	3.1	26.3	29.7	209.0	MA	3.10.20.10
60	1j05	111	3.1	36.2	47.2	418.0	MA	2.60.40.10
61	1jhl	108	3.1	39.1	52.7	406.6	MA	2.60.40.10
62	1bmz	127	3.1	11.5	75.2	478.8	MA	2.60.40.180
63	1bvk	108	3.1	39.7	63.3	406.6	MA	2.60.40.10
64	1c08*	107	3.1	35.3	47.9	402.8	MA	2.60.40.10
65	1oar	122	3.1	43.5	55.1	459.8	MA	2.60.40.10
66	1ttc	127	3.1	12.8	126.0	478.8	MA	2.60.40.180
67	1pun	129	3.1	34.2	193.9	486.4	B2M	3.90.79.10
68	1pus	129	3.1	36.1	194.1	486.4	B2M	3.90.79.10
69	1wiu	93	3.1	39.0	48.2	349.6	MA	2.60.40.10
70	1gko	127	3.1	11.6	79.2	478.8	B1MA	2.60.40.180
71	1n4x*	113	3.1	33.7	51.1	425.6	MA	2.60.40.10
72	1nvi	81	3.1	34.1	44.9	304.0	MA	3.10.20.30
73	1fvc	109	3.1	40.7	54.8	410.4	MA	2.60.40.10
74	1ugm	113	3.1	21.9	133.2	425.6	B2MA	
75	1igd	61	3.1	40.4	49.7	228.0	MA	3.10.20.10
76	1ivl*	107	3.1	64.5	52.7	402.8	MA	2.60.40.10
77	1w19	147	3.1	21.6	204.7	554.8	B1MA	
78	1kir*	107	3.1	36.0	48.1	402.8	MA	2.60.40.10
79	1kmt	141	3.1	44.5	71.2	532.0	MA	2.70.50.30
80	1l2n	81	3.1	36.0	52.1	304.0	MA	3.10.20.90
81	1dfu	94	3.1	13.9	120.2	353.4	B1M	2.40.240.10
82	2rox	127	3.1	14.0	90.0	478.8	B1MA	2.60.40.180
83	1oaq	120	3.1	40.6	49.3	452.2	MA	2.60.40.10
84	1rlf	90	3.1	42.4	54.7	338.2	MA	3.10.20.90
85	1tvd	116	3.1	38.0	44.4	437.0	MA	2.60.40.10
86	2dlf	113	3.1	39.3	49.9	425.6	MA	2.60.40.10
87	2imm	114	3.1	39.9	49.4	429.4	MA	2.60.40.10
88	1pga	56	3.1	26.5	30.3	209.0	MA	3.10.20.10
89	1b9r	105	3.0	29.1	37.1	395.2	MA	3.10.20.30
90	1pav	78	3.0	13.0	85.1	292.6	B1MA	
91	1i3o	144	3.0	40.4	279.1	543.4	B2MA	3.40.50.1460
92	1bm7	127	3.0	12.2	73.2	478.8	MA	2.60.40.180
93	1k26	156	3.0	46.1	217.7	589.0	B2MA	3.90.79.10
94	1lqb	118	3.0	38.9	129.4	444.6	B1MA	3.10.20.90
95	1tbe	76	3.0	33.5	47.4	285.0	MA	3.10.20.90
96	1gb4	57	3.0	28.9	35.3	212.8	MA	3.10.20.10

Table 3. (Continued.)

Rank	PDB	N	$F_{\max}$ (ϵ Å^{-1})	$L_{\rm n}$ (Å)	$L_{\rm m}$ (Å)	$L_{\rm f}$ (Å)	Pattern	CATH
97	1rbj*	124	3.0	32.0	252.1	467.4	B2M	3.10.130.10
98	1tfp	130	3.0	12.1	77.1	490.2	MA	2.60.40.180
99	1tyr	127	3.0	10.6	123.9	478.8	B1MA	2.60.40.180
100	1ves	113	3.0	35.1	43.7	425.6	MA	
101	1vfb*	107	3.0	36.6	48.2	402.8	MA	2.60.40.10
102	1lsl*	113	3.0	100.1	277.7	425.6	B1M	
103	1hz5	72	3.0	33.0	77.4	269.8	MA	3.10.20.10
104	1py9*	116	3.0	40.6	46.2	437.0	MA	2.60.40.10
105	1fmf	137	3.0	18.8	316.1	516.8	B3MA	3.40.50.280
106	2try	127	3.0	10.8	124.1	478.8	B1MA	2.60.40.180
107	4lve	114	3.0	39.8	52.9	429.4	MA	2.60.40.10
108	1gke	120	3.0	15.1	98.8	452.2	MA	2.60.40.180
109	1etb	127	3.0	8.6	89.6	478.8	B1MA	2.60.40.180
110	1i8k*	107	3.0	34.3	44.4	402.8	MA	2.60.40.10
111	1ict	127	3.0	9.5	82.8	478.8	B1MA	2.60.40.180
112	1pqe	126	3.0	33.4	73.7	475.0	MA	2.40.40.20
113	1gnu	117	3.0	5.0	201.8	440.8	B3MA	3.10.20.90
114	1kot	119	3.0	12.7	165.1	448.4	B2MA	3.10.20.90
115	1ui9	122	3.0	27.9	60.3	459.8	MA	3.30.1330.40
116	1w29	146	3.0	19.4	205.7	551.0	B1MA	
117	1oax	122	3.0	43.7	55.6	459.8	MA	2.60.40.10
118	1qp1	107	2.9	36.7	47.5	402.8	MA	2.60.40.10
119	1bz8	126	2.9	20.7	117.1	475.0	B1MA	2.60.40.180
120	1mel	148	2.9	38.0	43.8	558.6	MA	2.60.40.10
121	1f2x	135	2.9	40.1	49.2	509.2	MA	2.60.40.10
122	1em7	56	2.9	25.9	30.5	209.0	MA	3.10.20.10
123	1com	127	2.9	25.7	52.2	478.8	MA	3.30.1330.40
124	1lm8	106	2.9	43.3	137.8	399.0	B1MA	3.10.20.90
125	1dvy	124	2.9	12.4	77.1	467.4	B1MA	2.60.40.180
126	1f86	115	2.9	12.1	112.5	433.2	B1MA	2.60.40.180
127	1rnz*	124	2.9	31.7	253.8	467.4	B3M	3.10.130.10
128	1tjn	125	2.9	34.1	247.8	471.2	B1MA	
129	1v80	76	2.9	37.1	50.9	285.0	MA	
130	1vjk	88	2.9	31.0	105.5	330.6	MA	
131	1wit	93	2.9	39.3	50.2	349.6	MA	2.60.40.10
132	1mfw	107	2.9	12.4	136.0	402.8	B1MA	3.10.20.230
133	1ic4	107	2.9	35.4	45.7	402.8	MA	2.60.40.10
134	2igd	61	2.9	40.6	48.7	228.0	MA	3.10.20.10
135	5lve	114	2.9	36.8	52.5	429.4	MA	2.60.40.10
136	1ieh	135	2.9	51.1	115.4	509.2	MA	2.60.40.10
137	1dvt	115	2.9	12.2	77.5	433.2	B1MA	2.60.40.180
	1tit	89	2.2	43.2	54.7	334.4	MA	2.60.40.10
	1ubq	76	2.4	37.1	51.4	285.0	MA	3.10.20.90
	1ido	184	3.2	10.8	306.4	695.4	B3MA	3.40.50.410

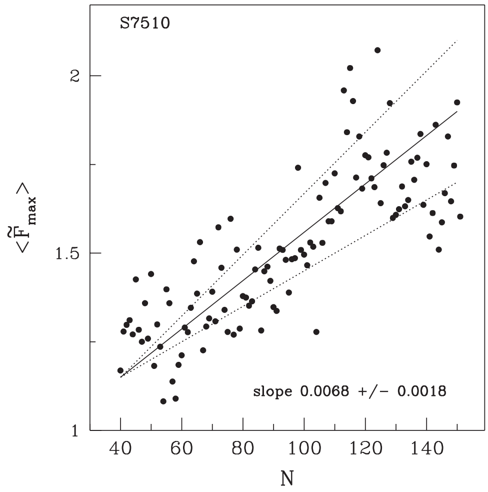

Figure 6. The correlation of the average $\tilde{F}_{\mathrm{max}}$ with N in the set S7510. The average is performed within all proteins corresponding to the same value of N. The solid line corresponds to a slope of 0.0068 and the dotted lines to slopes of 0.0050 and 0.0086.

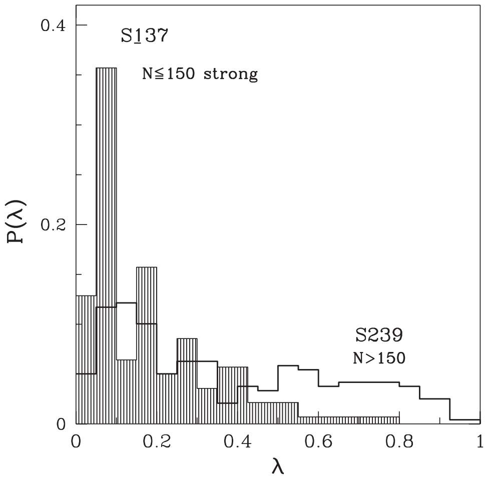

Figure 7. Probability distribution of the parameter λ which specifies at what fraction of the full extension from the native state the peak force arises. The shaded histogram is for the strongest short proteins. The other histogram is based on all long proteins that were studied.

without any gaps in the structure determination are selected. Two of these subsets are the learning (taken from 387 sequences) and testing (taken from 213 sequences) proteins that were used in developing a learning-based threading approach [101] that was linked to measures of the water exposure area. In principle, the sequential lengths of these proteins vary up to 1017 but the largest protein with structure data of a sufficient quality is 1no3 corresponding to $N = 851$. The proteins in these subsets can have between one and up to eight domains which are connected in a chain-like fashion into a well defined three-dimensional structure. The third subset consists of 11 out of 19 proteins that have been selected by the PDB managers [123] as representatives of most designable types of folds, where five of them have $N > 150$. These are listed in table 5. (The proteins 4bcl, 1got, 1tim and 2dnj which also have been selected

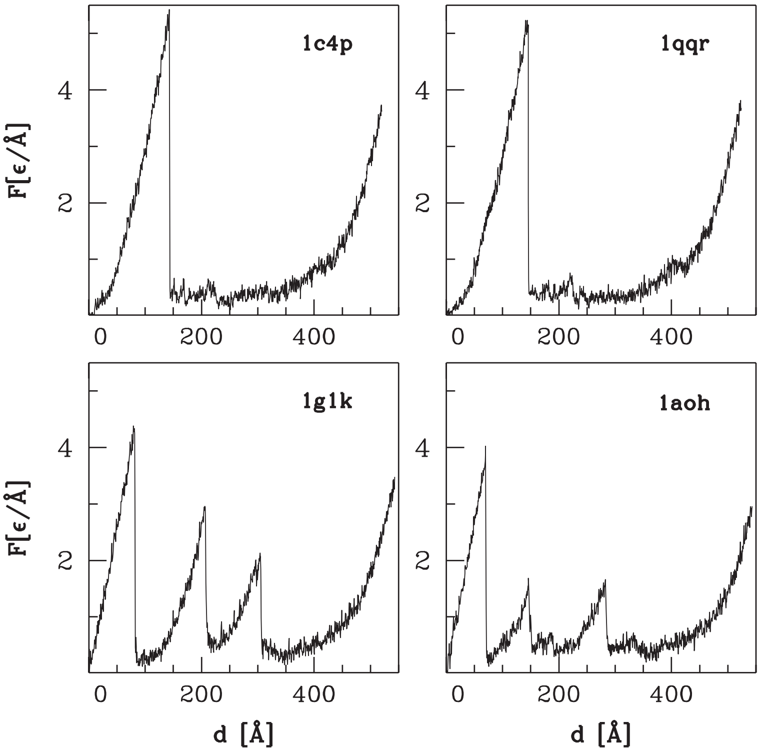

Figure 8. The force–displacement patterns for the four strongest short proteins.

in [123] would qualify but they contain gaps in the structure coordinates.) We have also added carboxypeptidase (1cpy, $N = 421$) as we have studied it before.

We were unable to determine the full F–d traces for the largest proteins at the slow speed of pulling. In order to overcome this limitation, at least approximately, we have investigated a correlation in the values of $F_{\max}$ between the fast and slow runs. The inset in figure 9 shows that, on average, $\tilde{F}_{\max,\text{slow}} = 0.86\tilde{F}_{\max,\text{fast}} - 0.13$. This result is based on all proteins from S7510 (open circles) and about half of the proteins from S239 (solid squares). We use it to rescale results for systems which we could calculate only in the fast way.

The resulting distribution of $F_{\max}$ is shown in the main panel of figure 9 and is seen to be noticeably broader than the distribution shown in figure 5 for the shorter proteins. This observation confirms the relevance of N in setting the scale of the forces. Table 5 suggests that the type of native fold is also important.

Table 6 shows results for proteins with $\tilde{F}_{\max} > 2.9$ from S239. There are 40 such proteins in this set which constitute 9.5% of the whole number—a fivefold increase compared to S7510, again suggesting the importance of the system size. Table 6, together with figure 7, indicates that, in large proteins, the maximum in the force is much less likely to occur at the beginning of pulling than in the case of the shorter proteins.

The strongest force belongs to the single domain, medium sized protein with the PDB code 1bg2, the structure of which is shown in figure 10. The corresponding F–d curve is shown in figure 11 together with the curves for the next three top entries.

The top second (2sli), third (1tmo) and fourth (1bfd) proteins are no longer of the single domain type. They consist of three, four and three domains, respectively. In such multi-

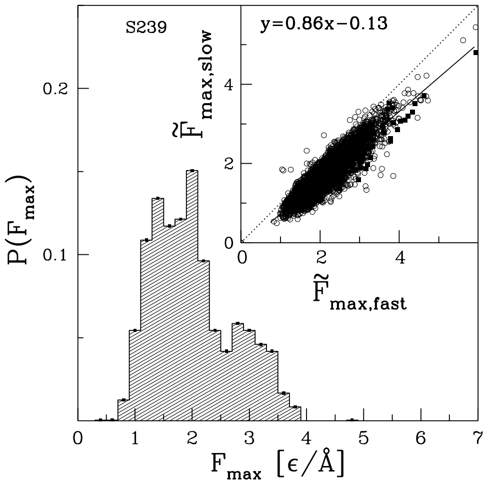

Figure 9. Probability distribution of $F_{\rm max}$ within the set S239. The values of 90 entries were obtained at the fast speed and then rescaled based on the result demonstrated in the inset. The inset shows the correlation between peak forces obtained at the two values of $v_{\rm p}$. Black squares correspond to long proteins $N > 150$ (based on 171 proteins) and open circles correspond to all small proteins. The correlations between the two ways of pulling have slopes of 0.82 and 0.87 in the two sets of proteins separately. The formula written at the top of the panel is the combined result of fitting and it corresponds to the solid line. The dotted line corresponds to a slope of 1.

domained cases, table 6 displays the CATH structure codes of the first two domains only. All top proteins are from the α/β class except for one domain from protein 1tmo which belongs to the α class. It is interesting to notice that among the strong long proteins we found only four other cases which incorporate at least one α class domain. However, the high force in all five cases arises from shearing β-strands from non-α domains.

The three domains of 1bfd correspond to three contiguous segments of the sequence: from 2 to 179, from 180 to 341 and from 342 to 542. We find that the large force comes from nearby parallel β-strands in the first domain. The four domains of 1tmo are not contiguous either. They are defined consecutively as (7–34, 491–508, 539–578), (56–147, 384–490, 516–538), (149–372, 585–610) and (618–790). The biggest force probably turns out to come from shearing of two parallel strands (502–506) and (574–578) which belong to the first domain. Moreover, we see high cooperativity in unfolding of this protein. For instance, the rupture of contacts inside the fourth domain corresponds to simultaneous breaking between the third and fourth domains and unravelling of the first domain proceeds together with breaking contacts between the first and third domains. The three domains of 2sli correspond to the segments (81–276), (277–404, 505–759) and (405–504). In this case, the large resistance to pulling appears to be generated mostly by the second domain, together with the first domains. The second domain unfolds in at least three steps, which are separated by unfolding events taking place in the other two domains.

It is not easy to identify the mechanical clamp in such complicated systems as densely packed domains in large proteins precisely. Our investigation usually relied on separating the individual domains. The exception is 1tmo in which the domains are delocalized along the sequence. Generally, we find that shearing motions between domains is what usually gives rise to large forces.

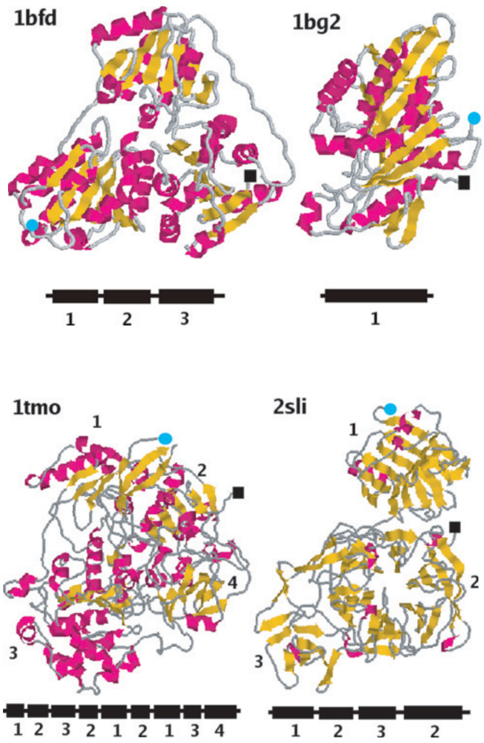

Figure 10. The backbone representation of the four strongest long proteins.

5.3. Force dependence on structural classes: set S3813

We now return to proteins with $N \leqslant 150$ and consider S3813—the subset of S7510 for which the CATH-based [102] assignment of topology to the fold is available. The distribution of F_{max} across S3813 is nearly the same as for S7510 (figure 5) and is, therefore, not shown.

The CATH classification scheme divides fold geometries hierarchically, first into classes (C), then architectures (A), then into topologies ($\mathrm{T_o}$) and finally into homologies (H). There is a numerical code associated with this scheme. For instance, for crambin (1crn) the symbol is 3.30.1350 which means that its class is α–β (C = 3), its architecture is a two-layer sandwich (A = 30) and its topology is crambin ($\mathrm{T_o}$ = 1350). Further extensions of the code refer to specification of the superhomologous family (for crambin, H is that of a plant protein).

We first discuss the role of the structural class in the determination of F_{max}. Figure 12 shows the values of $\tilde{F}_{\mathrm{max}}$ obtained for S3813 when split into four structural classes: α, β, α–β and no structure. The spread in the values for class α is fairly uniform, but it grows with N for class β. These tendencies are more perceptible when one averages F_{max} within the same value of N (we skip Ns for which there are fewer than four proteins available). The results of averaging are shown in figure 13. It is seen that for the α proteins, $\langle F_{\mathrm{max}} \rangle$ is N-independent whereas for the β proteins there is a systematic growth with the slope of about 0.011 ± 0.003. The α–β proteins also lead to growth, but a weaker one. There are too few entries to make an assessment for the 'no structure' class.

Figure 14 shows distributions of F_{max} corresponding to the three structural classes. It is seen that each distribution has a maximum in close vicinity to $1.5\ \epsilon\ \mathrm{\AA}^{-1}$. However, α proteins have a Gaussian looking distribution with a small tail. Thus the α proteins are not likely to

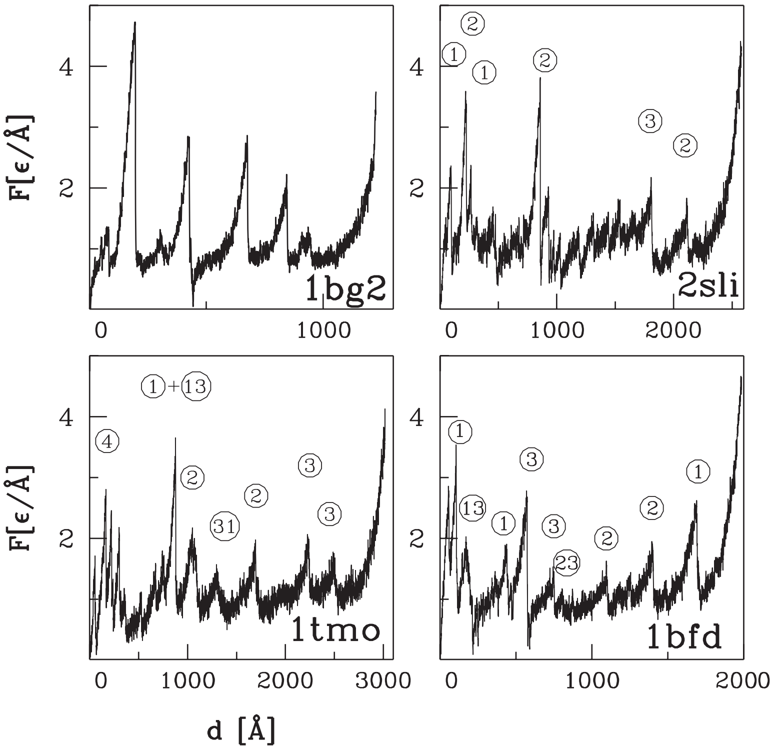

Figure 11. The force–displacement patterns for the four strongest long proteins. The single digits surrounded by circles refer to contacts corresponding to a domain with this label. Two-digit symbols in circles refer to contacts between two domains, as labelled in the circles.

generate a large force. The β and α/β proteins yield asymmetric distributions with substantial tails at large forces. Proteins which are unstructured have a distribution which peaks around $1.15\,\epsilon$ Å^{-1} and is narrow (not shown).

Experiments [18] and simulations [124] have indicated weak peak forces in the helical polycalmodulin. Also our theoretical survey suggests that there should be no short α proteins which could be considered as the strongest ones. Yet, in the set of *long* strong proteins shown in table 5, we find four multi-domained proteins (with N between 293 and 821: 1ile, 2ng1, 1ciy and 1cii) which have one domain that belongs to the α class. The large force in this case is due to interactions of the α-domain with the other domains.

5.4. Force dependence on structural architectures: set S3813

We now consider the finer characteristics of structure. Figure 15 shows architectures that are significantly populated in set S3813. In the α class, the orthogonal bundle is especially well represented—it comes with a weight of 80%. In the β class, 36% are barrels, 31% are sandwiches, 13% are ribbons and finally 13% are rolls. In the α/β class, rolls (40%) and two-layer sandwiches (39%) are especially abundant. The distribution is changed significantly (the histogram outlined by the heavy line in figure 15) when one focuses on the top 1.8% strongest proteins in the set, i.e. on the subset S137. We observe that the majority of the strong proteins are β sandwiches (60%) followed by α/β rolls (30%). None of these strong proteins belongs to the α or 'no structure' classes. We shall analyse properties of S137 further in a later section.

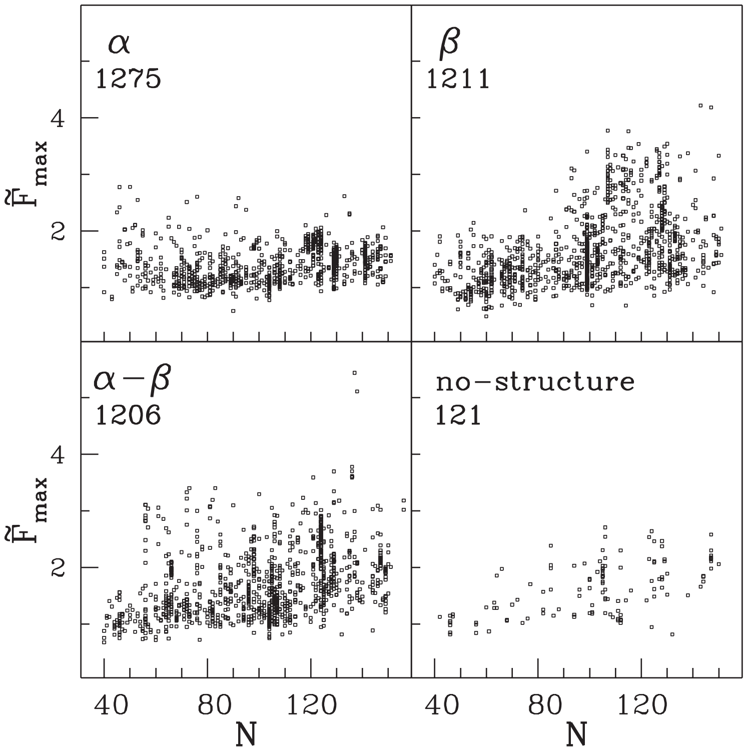

Figure 12. The values of $\tilde{F}_{\max}$ obtained for a given sequence length and split into four structural classes as indicated at the upper left corner of each panel. The numbers indicate the statistics available within S3813.

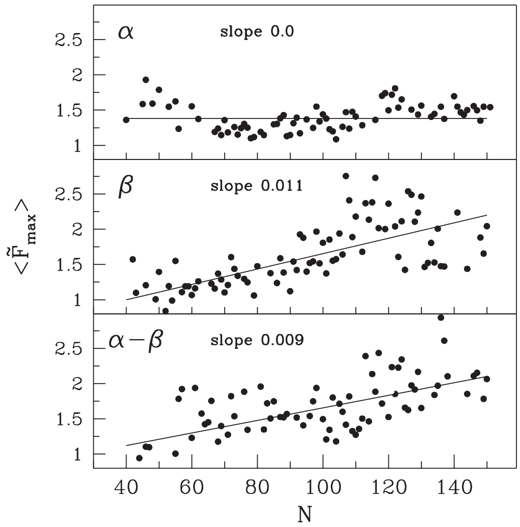

Figure 13. The correlation of the average $\tilde{F}_{\max}$ with N within the three structural classes. The average is performed within all proteins corresponding to the same value of N. The solid line corresponds to the slopes indicated. The spread in each slope is of order 0.002.

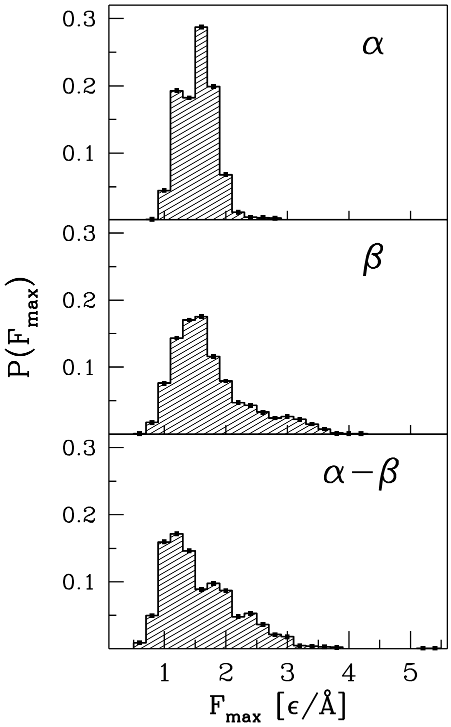

Figure 14. Probability distribution of the values of $F_{\max}$ within the set S3813 as split into the structural classes.

We find that the distribution of $F_{\max}$ for a given architecture peaks at a few selected values. These are listed in table 6. In most cases, there is just one peak. However, for two architectures, roll α/β and complex β, the distribution has two or three maxima. We find that, in this case, each maximum corresponds to a well defined topological group. Figure 16 shows that in the case of the α/β roll architecture the force distribution has two maxima. The first of these, at the weaker force, corresponds to the following topological groups: chitinase, nuclear transport factor and mannose binding proteins. The maximum at the larger force corresponds to topologies ubiquitin-like and P-30 proteins.

In the case of complex α/β architecture the distribution has three maxima shown in the bottom panel of figure 17. These are at $\tilde{F}_{\max}$ equal to 1.3, 1.9 and 3.4 and they correspond to the topologies of cytochrome C3, type 1ii antifreeze and nucleoside triphosphate, respectively. In the case of the α/β sandwich architecture (top panel of figure 17) the two-layer sandwich proteins yield a peak at 1.4 whereas the three-layer sandwich peak is at 2.1, but in both cases the distributions are rather broad.

The peak force distribution for the β sandwich architecture, shown in figure 18, is rather broad. Three local maxima can be resolved in the distribution. They can be understood by first splitting the data into two topologies—jelly roll and immunoglobulin-like—and then by splitting the latter into two homological families of immunoglobulins and transport protein and other. It is the immunoglobulins that account for the high force peak within this architecture.

In the case of the two-layer sandwich, partition into topological groups does not show any special preference for any topology (data not shown). On the other hand, for the three-layer sandwich one topology, that of the Rossmann fold, is responsible for a peak around $\tilde{F}_{\max} = 2.1$ in the distribution (not shown).

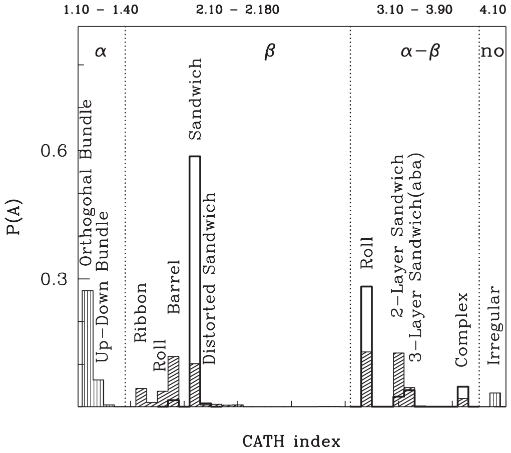

Figure 15. The dashed histogram shows the distribution of proteins across architectures in the set S3813. The thicker line corresponds to the distribution in the set S137 of the strongest proteins. The figure is arranged according to the CATH codes shown at the top. Codes that are not shown are not represented in S3813. The separate classes are divided by vertical dotted lines. 'no' is a shorthand for no structure. The names of well populated architectures are indicated. The histograms are normalized to 1.

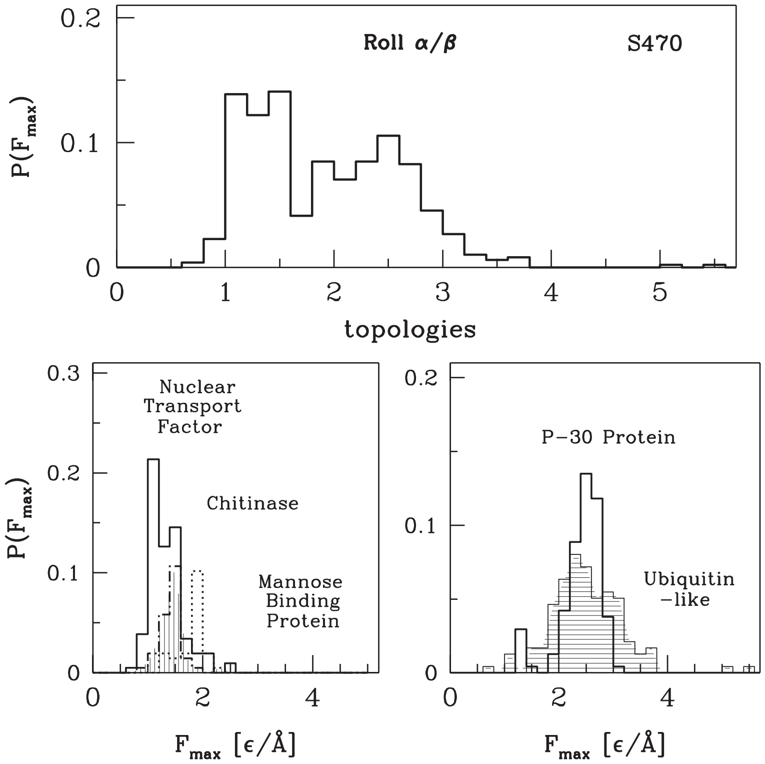

Figure 16. The top panel shows the force distribution within the α/β roll architecture. The bottom panels show the force distributions for various topologies that correspond to this architecture.

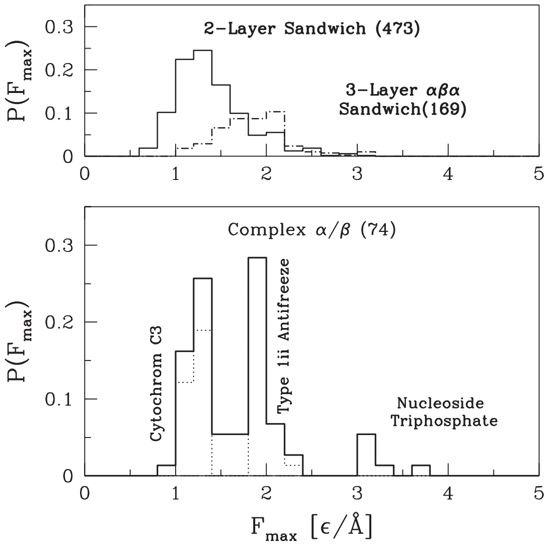

Figure 17. The top panel shows the force distribution for two types of architecture: two-layer sandwich (solid line) and three-layer sandwich. The numbers in brackets indicate the statistics. The bottom panel shows the force distribution for the complex architecture and identifies the contributing topologies.

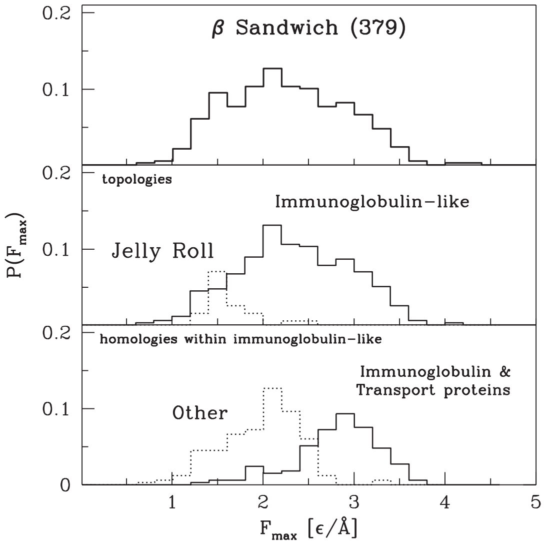

Figure 18. The top panel shows the force distribution within the β-sandwich architecture. The lower panels identify the contributing topologies and homologies as explained in the text.

5.5. Proteins with the same CATH index may differ in resistance to pull

We now focus on proteins with especially interesting dynamical properties. Sequences of some of these are listed in table 9. Consider proteins with the CATH index of 3.10.20.10, i.e. proteins which are α/β roll, ubiquitin-like and immunoglobulin binding. There are 10 such proteins in the strongest set S137 (table 3) and they are all short ($N \sim 56$). In the full set of S3813 we can find four other proteins (1mpe(56), 1jml(72), 1k51(72), 1q10(56)) with the same CATH index and similar length but their $\tilde{F}_{\max}$ is only around 1ϵ Å^{-1} or less. The F–d patterns for the two groups are shown in figure 19. The strong group is represented there by 1pga and 1p7e. They correspond to rank 52 and 56 in table 3. The weak group is represented by 1mpe and 1q10. The ribbon representations of 1pga and 1q10 are also shown in figure 19. For 1q10 we get $\tilde{F}_{\max} = 0.85$, $L_{\rm n} = 20.1$ Å, $L_{\rm m} = 181.6$ Å and $L_{\rm f} = 208.6$ Å. The pattern in the strong group is of the MA kind but in the weak case it is either MA (1q10) or M (the remaining three proteins; for 1mpe one force peak is resolvable in a $T = 0$ run). The main peak in the strong case arises early during the stretching process (the parameter λ is small) but much later in the other case. The scenarios of unfolding, shown at the bottom of figure 19, indicate clearly distinct pathways. One observation is that, in the weak protein case, the long range contacts are missing and the number of native contacts is smaller overall, compared to the strong 1pga and 1p7e.

Table 8 shows sequence alignment for the two weak and two strong proteins discussed here. The sequence alignment was done with the program ClustalW (version 1.82; available at www.ebi.ac.uk/clustalw/#) and it shows that proteins 1pga and 1q10 are almost identical sequentially—the alignment score is equal to 91, as they differ by merely three amino acids. These are: (1pga → 1q10) in positions: 30 F → V (big → small), 33 Y → F (polar → non-polar) and 34 A → F (small → big). The two structures are sufficiently similar to have the same CATH index and yet they are noticeably distinct structurally. They differ by the root-mean-square distance (RMSD) of 1.9 Å and the Z score is 3.9. They are similar in a segment of 38 amino acids where the helical and hairpin pieces are nearly identical.

We have found a similar relation for another pair of proteins 1p7e (strong) and 1mpe (weak) which are also listed in table 8. We conclude that substitution of a very small number of amino acids may dramatically alter the elastic properties of a protein even if the substitution results in no change in the structure classification index.

5.6. *Influence of the temperature*

As discussed in the section that describes the theoretical model used here, the temperature plays an important role in stretching because thermal fluctuations aid in the unravelling process. In most cases, and as long as individual peaks are articulated, the F–d traces at one temperature are scaled versions of the traces at another temperature except that that magnitude of fluctuations could be different.

However, our survey indicates that, for certain proteins, temperature may affect the very nature of the pattern. Figure 20 illustrates this for two proteins: 1nme and 1amx, which are ranked as 33 and 55, respectively, in table 3. The top panels show the F–d curves at $\tilde{T} = 0.3$ (the 'room temperature' value) whereas the bottom panels correspond to $\tilde{T} = 0$, i.e. when thermal fluctuations are not taken into account. It is expected that the peaks get taller and are placed further away from the origin as the temperature is lowered but figure 20 shows that, sometimes, the very identification of the tallest peak is altered. Also, the resolvability of a peak may depend on $\tilde{T}$ and we reemphasize that our tabulated data pertain to $\tilde{T} = 0.3$.

5.7. *Mechanisms of rupture in the strong short proteins of set S137*

The strong short proteins of set S137 belong to seven architectures and nine topologies. Among the architectures, the β sandwich and the α/β roll are especially well represented. Among the topologies, immunoglobins (2.60.40) and ubiquitin-like (UB roll; 3.10.20) are the most frequent. The remaining CATH topology codes are 3.90.79, 3.40.50, 3.10.130, 3.10.1330,

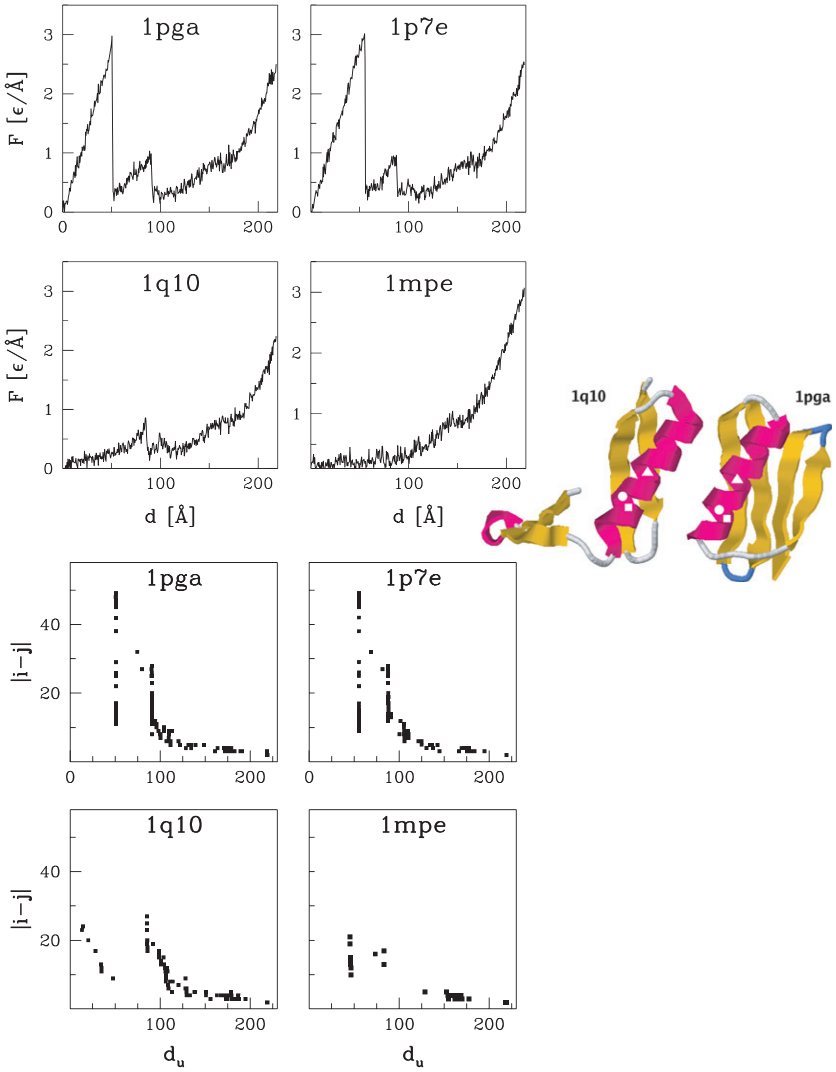

Figure 19. The panel on the right shows the ribbon representation of 1pga and 1q10, as indicated. The white symbols indicate locations which differ in their amino acid content. The squares correspond to Phe34 in 1q10 and Ala34 in 1pga. Similarly, the circles correspond to Phe33 and Tyr33, and triangles to Val30 and Phe30. The top left panel shows the F–d patterns for two strong (top panels) and two weak (bottom panels) proteins corresponding to the same CATH code. The bottom left panel shows the corresponding unfolding scenarios. The y-axis shows the sequential separation between two amino acids that make a native contact. The x-axis shows the pulling tip displacement at which the contact is broken for the last time—thermal fluctuations may reinstate a broken contact temporarily and hence we seek to record the last rupture event.

3.10.50, 2.40.40, 2.40.240. The dominant functions in this set of proteins relate to immune system (16 proteins), binding (14), signalling (14), transport (13) and immunoglobulin (9).

It should be noted that the proteins listed in table 3 are not necessarily distinct biologically and there could be several structure determinations and therefore several PDB codes corresponding to the same or nearly the same protein (or to the same protein but studied under different conditions). The stretching dynamics is sensitive to the structural details and

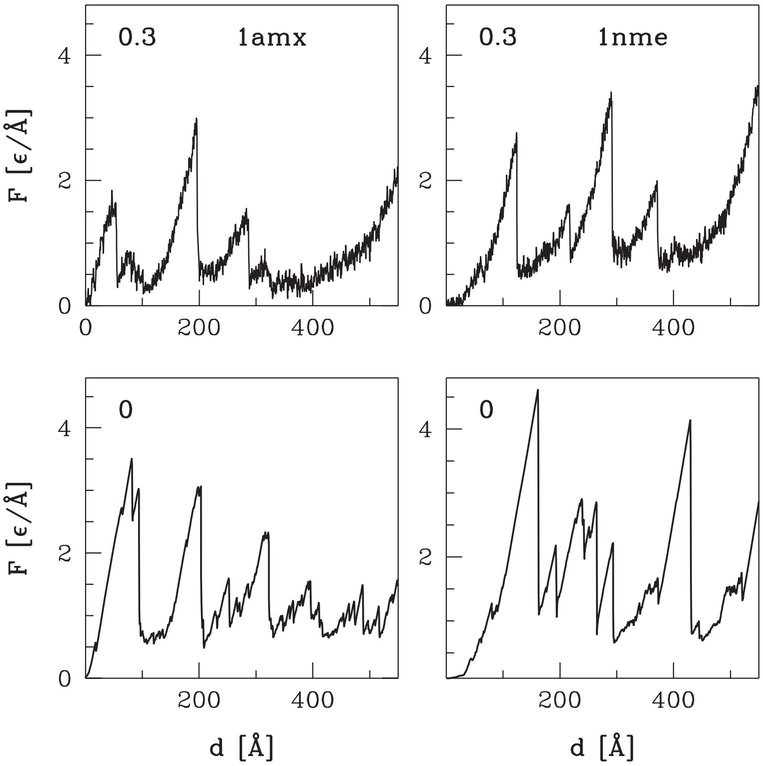

Figure 20. The F–d patterns for two proteins: 1amx (the left panels) and 1nme (the right panels). The top panels refer to $\tilde{T} = 0.3$. The bottom panels correspond to $\tilde{T} = 0$.

thus to the particular code, and the survey has been accomplished on a code by code basis. In order to understand homology-based links between the protein structures listed in table 3, we have performed studies of homology using the FATCAT server [125]. We have found that 43 proteins in the set S137 are unrelated whereas the remaining 93 belong to 32 different groups of at least two elements each. These groups are listed in table 4, where the names of the groups have been assigned based on the biological name of at least one member of the group. The proteins within one group have at least 80% sequence similarity. We find, in particular, the top two proteins, 1c4p and 1qqr, are both streptokinase β-domain proteins (UB-roll topology) but are involved in different functions (blood clotting and hydrolase activation, respectively). The third protein, 1g1k, and the fourth-ranked 1aoh are in the same group as 1anu, ranked 28th, etc.

We observe that there is a correlation between the CATH index and the type of the F–d pattern. Table 7 shows, for instance, that all 48 strong proteins with the 2.60.40.10 index give rise to the MA type of the pattern.

We now consider mechanisms that give rise to the generation of a big force in short proteins, i.e. to the formation of the 'mechanical clamp'. We have found that the very high mechanical resistance to pulling of proteins listed in table 3 is related, in 95% of the cases, to one basic mechanism [15, 126, 90]: shear rupturing of a hydrogen-bonded sheet formed by two *parallel* β strands. The top peak forces arise when at least one of these strands is near a terminus (we continue to consider pulling only by the termini). The mechanical clamp may be involved at the beginning of folding. Often, however, a prior unwinding of the surrounding layers is required which results in a structure rotation and emergence of minor preceding peaks. The

Table 4. Proteins that are close homologically within the set of the top 137 strongest proteins.

Protein	PDB-code
Beta domain of streptokinase	1c4p, 1qqr
Cohesin domain of the cellulosome from *Clostridium thermocellum*	1aoh, 1anu
Staphylokinase, sakstar variant	1ssn, 2sak, 1c76, 1c77, 1c78, 1c79
Mutator mutt protein	1ppx, 1pun, 1pus, 1tum
APOCOX11 (cytochrome C oxidase assembly protein ctag)	1so9, 1sp0
Transthyretin (pre-albumin)	2rox, 1f86, 1bz8, 1oo2, 1gko, 1tfp, 1tyr, dvy
Transthyretin	1ttc, 1eta, 1etb, 1dvt, 1bm7, 1bmz
Human transthyretin	1ict, 1bzd, 2try
Rat transthyretin	1gke, 1kgi, 1ie4
Transthyretin (sea bream)	1sn0, 1sn2, 1sn5
Coxsackie virus and adenovirus receptor (CAR) D1 domain	1eaj, 1f5w
Mouse monoclonal antibody D1.3 (VH or VL domain)	1kip, 1kiq, 1a2y, 1kir
Caspase-3 (large subunit)	1nme, 1i3o
Murine, mouse T cell receptor (TCR)v-alpha domain	1b88, 1nam, 1h5b
B1 domain of protein L from *Peptostreptococcus magnus*	1hz5, 1hz6, 1k53
GABAA receptor associated protein GABARAP	1gnu, 1km7, 1kot
Immunoglobulin E, FV domain, SPE-7	1oau, 1oax, 1oaq, 1oar
Immunoglobulin G, VL domain	1ivl, 2imm
Immunoglobulin VL domain (Bence Jones' protein)	1wtl, 1qp1
Human immunoglobulin K-4 light chain, Len	4lve, 5lve, 1lve
Twitchin immunoglobulin superfamily domain, 18th domain	1wit, 1wiu
Nudix protein from *Pyrobaculum aerophilum*	1jrk, 1k26
Chorismate mutase from *Thermus thermophilus*	1ufy, 1ui9
N-terminal doublecortin domain from DCLK	1mfw, 1mg4
Third IgG-binding domain from streptococcal protein G	1igd, 2igd, 1p7e
B1, B2 IgG-binding domain from streptococcal protein G	1pga, 1pgx
Different variant of B1 domain, streptococcal protein G	1em7, 1gb4
Anti-hen egg lysozyme antibody (HYHEL-10) fragment VL, VH	1jhl, 1c08, 1ic4
Lumazine synthase from *Mycobacterium tuberculosis*	1w19, 1w29
Elongin-B from HIF-pvhl/elongin-c/elongin-b complex	1lqb, 1lm8
Ubiquitin	1v80, 1tbe
Ribonuclease B or A	1rbj, 1rnz

latter situation is nearly a rule when the termini are on the same side of the native protein. The strength of resistance of the clamp is governed primarily by the number of contacts (hydrogen bonds) within the clamp. It can also be enhanced by stabilizing interactions that may encase the clamp.

In the following, we identify the mechanical clamp and analyse the rupturing process in selected short strong proteins. We first discuss six variants of the commonest clamp as found in proteins 1c4p, 1g1k, 1ssn, 1ppx, 1oo2 and 1i3v which are ranked as numbers 1, 3, 5, 8, 18 and 19, respectively. The ribbon representations of their native structure are shown in figure 21. We examine their behaviour by studying unfolding scenarios and by investigating the effect of removal of certain contacts on the resulting F–d curves. The removal is implemented by setting the value of ϵ in the contact under study to zero. The mechanical clamps are identified by the largest resulting reduction in $F_{\max}$ and are listed in table 10.

5.7.1. Protein 1c4p. Protein 1c4p is a streptokinase β-domain which has the ubiquitin-like (UB roll) topology. It consists of four chains which we find to possess similar elastic properties despite somewhat differing sequences. As an illustration, we focus on the first chain here. It

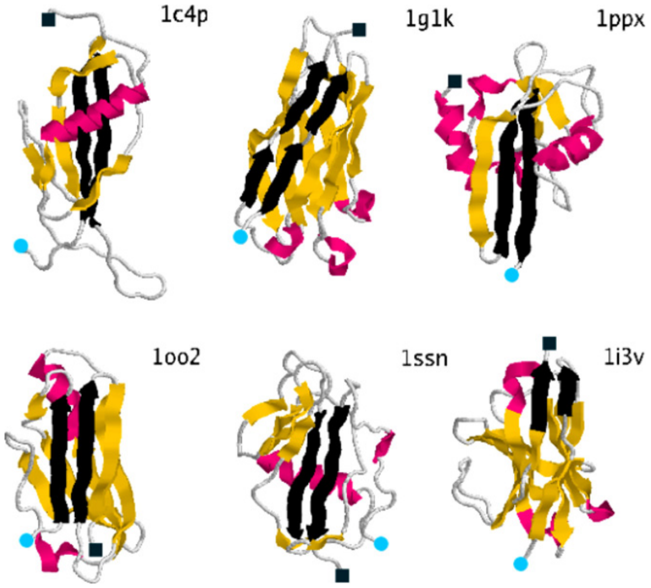

Figure 21. Ribbon representation of the strong short proteins 1g1k, 1oo2, 1i3v, 1c4p, 1ppx, 1ssn of distinct structures. The parallel β-strands shown in black correspond to the 'mechanical clamp' which is responsible for the largest contribution to the peak force.

Table 5. Results of the Go-like model for the proteins selected in [123] as representing typical architectures. The reference contains several more proteins but their structure determination contains gaps.

	PDB	N	$F_{\max}$ (ϵ Å^{-1})	L_{n} (Å)	L_{m} (Å)	L_{f} (Å)	Pattern	CATH	Architecture
1	1stm	141	2.6	23.1	130.7	532.0	B1MA	2.60.120.220	β-sandwich
2	1ndd	74	2.4	30.3	40.2	277.4	MA	3.10.20.90.	α/β roll
3	1air	352	2.0	24.7	859.9	1333.8	B4M	2.160.20.10	β solenoid
4	1fua	206	1.9	28.3	156.3	779.0	B1MA	3.40.225.10	α/β 3-layer sandwich
5	2ccy	127	1.6	28.1	174.3	479.8	B1M	1.20.120.10	α up–down bundle
6	1lrv	233	1.5	64.7	642.9	881.6	B2MA	1.25.10.10	α horseshoe
7	1ppr	312	1.5	36.7	315.3	1181.8	B1MA	1.40.10.10	α solenoid
8	1fbr	93	1.4	56.6	337.3	369.6	B1M		
9	1mbn	153	1.4	24.5	150.1	577.6	B1M	1.10.490.10	α orthogonal bundle
10	1jpc	108	1.4	19.0	57.6	406.6	MA	2.90.10.10	β prism
11	1vie	60	1.1	20.8	125.7	224.2	B1M	2.30.30.60	β roll
12	1rie	127	1.0	10.8	146.5	478.8	B1M	2.102.10.10	β 3-layer sandwich
13	1ccm	46	0.9	11.8	18.7	170.0	MA	3.30.1350.10	α/β 2-layer sandwich
14	1hcd	118	0.8	11.8	44.8	444.6	MA	2.80.10.50	β trefoil
15	1kvd	63	0.6	24.3	194.5	235.6	B1M	4.10.420.10	Irregular

contains an α-helix (196–210), denoted as I, and eight β-strands, denoted by A to H, as labelled consecutively from the N-terminus to the C-terminus. These β-strands form three β-sheets: a, b and c. The a-sheet comprises four strands (A, B, G, E), the b-sheet two (C, F) and the c-sheet also two (D, H). Similar structures (e.g. the second ranked 1qqr) are found in the strong proteins with the CATH index 3.10.20.180, in some proteins (1c76, 2sak) with the index 3.10.20.130 and in most proteins with the index 3.10.20.10 (such as 1pgx, 1hz6, 1k53).

Table 6. Top strongest long proteins (with $N > 150$) as predicted by the Go-like model. The format of this table is similar to that of table 3 except that the number of domains is listed in the column entitled n_D. This column is followed by the CATH index of the first two domains. For multiple domains, a single CATH code means that the code structure is repeated. The force values marked by $*$ are obtained based only on the fast run and using the rescaling formula $F_{\text{max,slow}} = F_{\text{max,fast}}0.86 - 0.13$.

Rank	PDB	N	$F_{\max}$ (ϵ Å^{-1})	$L_{\rm n}$ (Å)	$L_{\rm m}$ (Å)	$L_{\rm f}$ (Å)	Pattern	n_D	D_1	D_2
1	1bg2	323	4.8	17.0	133.3	1223.6	B1MA	1	3.40.850.10	
2	2sli	679	3.8	50.6	224.3	2576.4	B1MA	3	2.60.120.200	2.120.10.10
3	1tmo	794	3.7	44.0	2309.7	3013.8	B4MA	4	3.40.50.740	3.40.228.10
4	1bfd	523	3.6	62.1	124.1	1983.6	MA	3	3.40.50.1220	3.40.50.970
5	2plc	274	3.6	9.4	150.4	1037.8	B1MA	1	3.20.20.190	
6	1clc	541	3.6*	39.6	162.3	2052.0	MA	2	2.60.40.10	1.50.10.10
7	1ile	821	3.6*	40.4	2372.3	3116.0	B5MA	3	1.10.730.10	3.90.740.10
8	5ptd	296	3.5	7.4	199.6	1121.0	B2MA	1	3.20.20.190	
9	1thv	207	3.5	23.4	38.4	782.8	MA	1	2.60.110.10	
10	1cpy	421	3.5*	40.5	1216.3	1596.0	B4MA	2	3.40.50.1820	1.10.287.410
11	1kas	411	3.5*	17.6	54.2	1568.0	MA	2	3.40.47.10	
12	1bif	432	3.5*	54.3	1110.8	1637.8	B4MA	2	3.40.50.300	3.40.50.1240
13	1ctn	538	3.5*	60.1	1024.5	2040.6	B6MA	3	2.60.40.10	3.20.20.80
14	1dot	686	3.4*	28.1	2237.5	2603.0	B4MA	4	3.40.190.10	
15	1cbg	490	3.4*	29.0	268.3	1868.2	B1MA	1	3.20.20.80	
16	1auq	208	3.4	20.9	389.3	786.6	B3MA	1	3.40.50.1820	
17	2ng1	293	3.4	14.0	778.5	1109.6	B5MA	2	1.20.120.140	3.40.50.300
18	1lpp	534	3.4*	49.5	1543.3	2025.4	B3MA	1	3.40.50.1820	
19	1bs9	207	3.3	20.8	475.8	782.8	B3MA	1	3.40.50.1820	
20	1cgt	684	3.3*	55.7	1291.5	2595.6	B6MA	4	3.20.20.80	2.60.40.1180
21	1dmr	779	3.3*	24.8	2249.4	2956.4	B5MA	4	3.40.50.740	3.40.228.10
22	1qpg	415	3.2*	14.3	1210.3	1573.2	B5MA	2	3.40.50.1260	3.40.50.1270
23	8ohm	435	3.2*	43.9	1116.8	1648.2	B3MA	3	3.40.50.300	3.40.50.300
24	1zxq	192	3.2	79.2	451.7	726.2	B3MA	2	2.60.40.10	
25	1a8h	500	3.2*	46.3	1577.4	1996.2	B6MA	3	3.40.50.620	2.170.220.10
26	1ciu	683	3.2*	58.7	1566.4	2591.6	B8M	4	3.20.20.80	2.60.40.1180
27	1cyg	680	3.1*	63.6	1267.1	2580.2	B6MA	4	3.20.20.80	2.60.40.1180
28	1ciy	577	3.1*	41.9	84.4	2188.2	MA	3	1.20.190.10	2.100.10.10
29	1hcz	250	3.1	41.7	160.3	946.2	B1MA	2	2.60.40.830	2.40.50.100
30	1fsz	334	3.1*	62.0	174.6	1265.6	B1MA	2	3.40.50.1440	3.30.1330.20
31	1cex	197	3.1	28.1	436.9	744.2	B3MA	1	3.40.50.1820	
32	1bag	425	3.1*	13.7	508.5	1611.2	B3MA	2	3.20.20.80	2.60.40.1180
33	1gca	309	3.0*	43.4	115.8	1170.8	MA	2	3.40.50.2300	
34	1cii	602	3.0*	38.5	1783.2	2283.2	B2M	3	1.20.250.10	3.30.305.10
35	1avk	620	3.0*	36.3	1051.9	2352.2	B6MA	3	3.10.450.40	3.10.450.40
36	1edg	380	3.0*	31.4	272.5	1443.2	B2MA	1	3.20.20.80	
37	1ra9	159	3.0	13.9	193.2	600.4	B1MA	1	3.40.430.10	
38	1chd	198	3.0	8.5	299.3	748.6	B2MA	1	3.40.50.180	
39	1fts	295	2.9	10.5	779.7	1117.2	B2MA	2	1.20.120.140	3.40.50.300
40	1bk0	329	2.9	29.2	555.6	1246.6	B2MA	1	2.60.120.330	

Table 7. The most probable value of the force within a given CATH-based structure type. The peak forces for a given architecture are denoted as $\tilde{F}^{\mathrm{A}}_{\max}$, for a given topology by $\tilde{F}^{\mathrm{T}_o}_{\max}$, and for a given homology by $\tilde{F}^{H}_{\max}$. The symbol 'Set' refers to the number of proteins inside a given structure type.

Name	$\tilde{F}^{A}_{\max}$	$\tilde{F}^{\mathrm{T}_o}_{\max}$	$\tilde{F}^{H}_{\max}$	Set
Orthogonal bundle α	1.45			1022
Up–down bundle α	1.85			237
Ribbon β	1.35			161
Roll β	1.25			137
Barrel β	1.5			442
Sandwich β	2.1			379
Immunoglobulin, transport protein			2.95	148
Other			2.1	184
Two-layer sandwich α/β	1.45			473
Three-layer sandwich $\alpha/\beta/\alpha$	2.0			169
Roll α/β 1st	1.35			470
Roll α/β 2nd	2.5			470
Ubiquitin, P-30 protein		2.5		237
Other		1.25		233
Complex α/β				74
Cytochrome C3		1.2		23
Type 1ii antifreeze		1.95		27
Nucleoside triphosphate		3.1		6
Few secondary structures	1.15			121

Table 8. Correlation of the CATH classification index with the type of pattern of the *F*–*d* curve. The numbers in the last column indicate the numbers of proteins found that have the listed type (types) of pattern.

CATH	Pattern	Number of cases
2.60.40.10	MA	48
2.60.40.180	MA, BMA	15, 8
2.60.40.680	MA	3
2.60.40.740	B1MA	1
2.40.240.10	MA	1
2.40.40.20	MA	1
2.70.50.30	MA	1
3.10.20.10	MA, M	9, 1
3.10.20.30	MA	2
3.10.20.90	MA, BMA	5, 6
3.10.20.130	MA, BM, M	1, 1, 4
3.10.20.230	BMA	1
3.10.20.180	M	2
3.10.130.10	BM	3
3.90.79.10	BM, BMA	4, 2
3.30.1330.40	MA	3
3.40.50.1460	BMA, BM	4, 1

Figure 22 shows the unfolding scenario for the first chain of 1c4p at $\tilde{T} = 0.3$ together with the *F*–*d* curve in the inset. The force peak at about 140 Å involves a near simultaneous rupture of the contacts which relate to the terminal strands A and G: A + G (meaning between A and G), A + I, G + I (i.e. with the helix) as well as D + H, A + B, E + G, D + G, C + G and

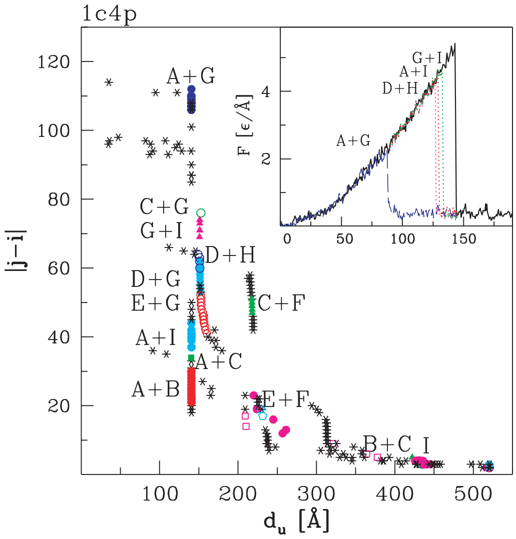

Figure 22. Unfolding scenario for 1c4p at $\tilde{T} = 0.3$. The letter symbols indicate which secondary structures are involved in a contact that is broken at the distance d_u. The data symbols marked by asterisks correspond to contacts which do not involve any secondary structures. The remaining symbols are diversified and have a meaning identified by the letter symbols placed next to them. The inset refers to the F–d curves. The solid line corresponds to a situation in which all contacts are present. The remaining lines, described also by the letter symbols, correspond to a situation in which the indicated contacts (e.g. between strands A and G) are removed.

Table 9. Sequential alignment for two weak (the first two lines) and two strong proteins.

PDB	
1MPE	MQYKVILNGKTLKGETTTEAVDAATFEKV|**V**|KQ|**FF**| NDNGVDGEWTYDDATKTFTVTE 56
1Q10	MQYKVILNGKTLKGETTTEAVDAATAEKV|**V**|KQ|**FF**| NDNGVDGEWTYDDATKTFTVTE 56
1PGA	MTYKLILNGKTLKGETTTEAVDAATAEKV|**F**|KQ|**YA**| NDNGVDGEWTYDDATKTFTVTE 56
1P7E	MQYKLVINGKTLKGETTTKAVDAETAEKA|**F**|KQ|**YA**| NDNGVDGVWTYDDATKTFTVTE 56

A + C. Breaking these contacts results in destroying the *a* β-sheet. Later on, the remaining contacts, such as C + F in the β-sheet are broken and the helix is the last to unravel.

The inset of figure 22 shows that removing contacts between A and G lowers the peak force by about 50%, whereas removing contacts in other regions (illustrated by D + H, A + I, G + I) affects the peak force much less substantially: D + H by 18%, A + I by 10%, A + B by 4% and C + F imperceptibly. Thus the mechanical clamp is formed by the near-terminal strands A and G. These strands are marked in black in figure 21 and are sheared by a force parallel to the end-to-end vector.

The value of the peak force is governed by the number of bonds between A and G. In the case of 1c4p, this number is particularly large (27) (see also table 9) and hence this protein is at the top of the list in table 3. The value of the peak force is, however, enhanced by the nearby contacts such as D + H and with the helix.

We find that this protein can actually unfold in (at least) two ways which is reflected in the values of $F_{\max}$. The first pathway is shown in figure 22. The second pathway yields a force smaller by 0.3 ϵ Å^{-1} and it involves breaking only the A + G, A + I, A + C and A + B contacts in the main peak. Breaking the remaining contacts C + G, G + I, D + H, D + G and E + C generates a second peak which is also smaller.

Table 10. Identification of a mechanical clamp $F_{\max}$ for selected proteins for a full set of the contacts and $F^r_{\max}$ for a situation in which some contacts, shown in the last column, are removed. K_{nat} denotes the number of all native contacts and K_{r} the number of contacts that are removed (aa, amino acid).

PDB	K_{nat}	$F_{\max}$ (ϵ Å^{-1})	K_{r}	$F^r_{\max}$ (ϵ Å^{-1})	aa–aa
1c4p	366	5.4	27	2.1	(158–168, 266–278)
1aoh	456	4.2	21	2.2	(5–10, 137–140); (12–15,142–146)
1g1k	441	4.2	16	2.7	(4–8, 131–135); (10–13, 131–141)
1ssn	382	3.8	22	2.1	(24–32, 124–132)
1ie5	281	3.8	10	2.4	(10–16, 35–40)
1c76	327	3.7	23	1.4	(24–32, 124–135)
1ppx	343	3.7	21	1.2	(2–12, 78–86)
1yn4	295	3.6	29	1.7	(44–53, 132–140)
2sak	334	3.6	30	1.9	(24–32, 109–119)
1sp0	371	3.6	19	1.8	(32–41, 138–144)
1sn0	323	3.5	21	1.4	(12–18, 105–112)
1oo2	316	3.5	17	1.5	(12–18, 105–112)
1i3v	373	3.5	7	2.6	(11–13, 123–128)
1i9e	321	3.5	15	2.3	(9–13, 105–110)
1qp1	315	2.9	13	2.1	(9–13, 102–116)
1amx	456	2.9	24	1.6	(19–26, 105–115)

We have also considered other ways of pulling 1c4p: by 1–108, 1–63 1–89, 22–136, 63–136 and got $F_{\max}$ of 2.5, 2.9, 1.7, 1.9 and 1.7 ϵ Å^{-1}, respectively, indicating that the terminal stretching comes with the strongest clamp.

5.7.2. Protein 1g1k. Protein 1g1k is a cohesin module from *Clostridium cellulolyticum* with a topology described as immunoglobulin-like. 1g1k consists of 11 β-strands which form four β-sheets a, b, c, and d. Proteins 1aoh and 1anu have a very similar structure. In the case of 1g1k, the mechanical clamp is similar to that found in 1c4p except that instead of one long ladder of two β-strands that is additionally stabilized by a helix one observes two short ladders of β-strands that are stabilized by contacts with other β-strands from the same sheet.

Figure 23 shows the unfolding scenario for 1g1k and the corresponding F–d curve in the inset. This protein unfolds in several stages. The highest resistance to pull associated with the first force peak is created by shearing the contacts between A + J and B + K, which involves a simultaneous rupture of the contacts between I + A, I + D, I + K, I + J and H + K. The first after-peak is due to breaking other contacts in the same sheet (C + H, A + C). The next to unravel are the contacts between various sheets, and a β hairpin F + G unravels towards the end of the process.

Our identification of this non-contiguous mechanical clamp is confirmed by noticing that a removal of the contacts in A + J, B + K (shown in black in figure 21) brings the peak force down to about 60% of its value (table 9 and the inset of figure 23). Sixteen contacts are involved in this element so the peak force is not as large as in 1c4p. There are six contacts in A + J and three in B + K so the former stabilizing influence contributes more to $F_{\max}$ than the latter.

This protein, like 1c4p, has at least two ways of unfolding. However, in contrast to the 1c4p we do not observe any changes in the first step connected with the strongest mechanical resistance but in the second force peak. In the 60% of trajectories, the magnitude of the second peak is larger by nearly a factor of two.

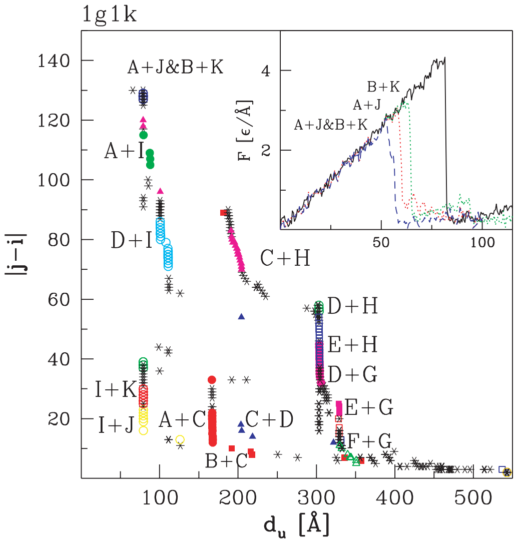

Figure 23. Similar to figure 22 but for protein 1g1k.

One can generate an even bigger force in 1g1k by pulling it by amino acids 1 and 104 instead of by the termini 1 and 143. The resulting $F_{\max}$ is 5.7 ϵ Å^{-1} which makes it bigger than for 1c4p.

5.7.3. Protein 1ppx. Protein 1ppx belongs to the α/β class and has the topology of nucleoside triphosphate pyrophosphohydrolase. It consists of five β-strands forming a single sheet, two α-helices and three 3–10 helices. This proteins has the CATH index of 3.90.79.10 which is also shared by 1jrk, 1k26, 1pun and 1pus.

The identification of the mechanical clamp for this protein is less obvious and is best addressed by looking at the unfolding scenario (figure 24). The F–d curve (the inset of figure 24) indicates that there are two minor peaks before the main peak arises. The first minor peak is due to unravelling of all contacts with the helices (I). The second minor peak is due to breaking B + G. Finally, the main peak involves the terminal strand A and is due mostly to rupturing 21 bonds in A + F. The end events involve unzipping of E + F. Generally, unzipping involves much lower forces than shearing. Thus the mechanical clamp here is again found to be formed by two parallel β-strands, A and F, of which one is terminal and the other is separated from the terminal by another strand. The mechanical clamp is so stable here that its stabilizing bonds unravel before the clamp itself disintegrates.

5.7.4. Protein 1oo2. Protein 1oo2, with CATH index 2.60.40.180, is mechanically very stable, even though it is not easy to recognize what makes it so. This protein consists of single α and 3–10 helices, and two β-sheets: an a-sheet (A, B, I, J-strands) and a b-sheet (C, D, F, H). In this case both termini are located close to each other so that the end-to-end vector does not cut through the protein and is perpendicular to the β-sheets. Moreover, the terminal strands A, J of the protein are not directly hydrogen bonded, but there is one more β-strand, I, between them. All these strands lie in one plane and A is parallel to I.

Figure 25 indicates that around $d = 50$ Å, the force of resistance is rather low and is due to unzipping of the terminal strand J away from strand I. The unzipping results in a rotation

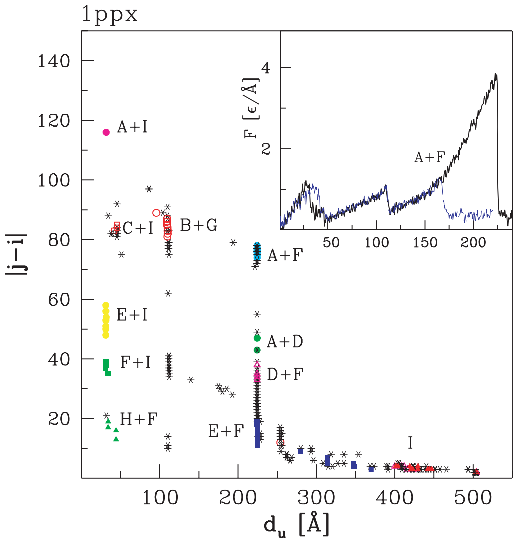

Figure 24. Similar to figure 22 but for protein 1ppx.

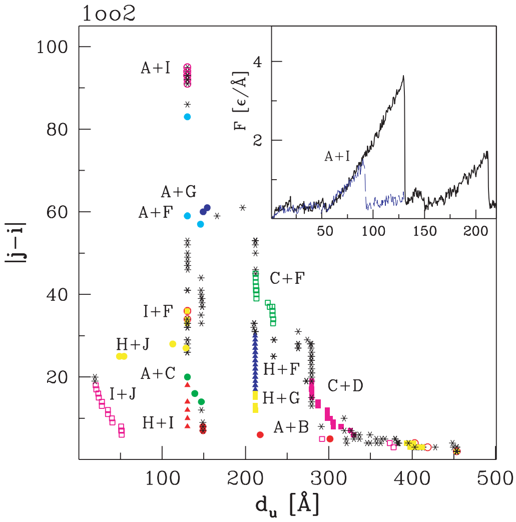

Figure 25. Similar to figure 22 but for protein 1oo2.

so that the β-sheet A + I eventually becomes parallel to the pulling direction. The shearing between strands A and I (30 contacts) is the primary reason for the emergence of the main peak force. Breaking the contacts between A and I causes a simultaneous breaking of the contacts between the a- and b-sheets (A + G, A + F, A + C, I + F, I + H).

5.7.5. Protein 1ssn. Protein 1ssn contains eight β-strands, forming three β-sheets, and two helices: α and 3–10. The end-to-end vector is very short and lies perpendicular to the longest β-sheet. This β-sheet is created by a β-strand from the C-terminus which also forms contacts with strand A. However, there is one short helix before the A strand and one loop at the N-terminus. Figure 26 indicates that the first contacts to break are those between the N-terminal

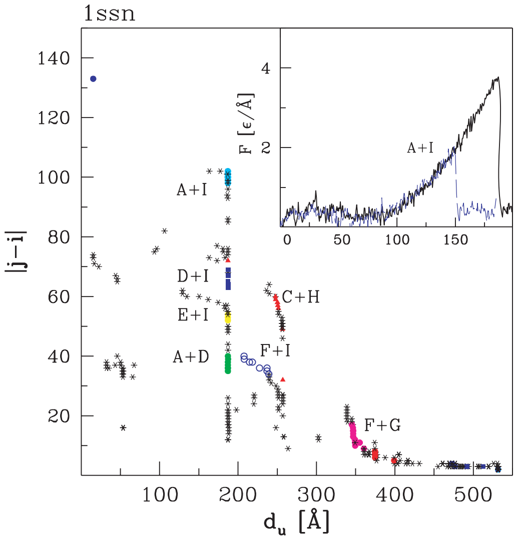

Figure 26. Similar to figure 22 but for protein 1ssn.

loop and strand A. This is seen as a peak with a small force around 30 Å. This results in a rotation. At this stage, the end-to-end vector is parallel to the β-sheet formed by strands A and I. The main peak is due to rupturing of A + I and, to a lesser extent, E + I.

5.7.6. Protein 1i3v. Protein 1i3v is titin-like and corresponds to a CATH index of 2.60.40.10. The end-to-end vector is parallel to one of their β-terminal strands. A core of the mechanical clamp consists two parallel β-strands each emerging from a terminus. In addition, other same-sheet (anti-parallel) strands stabilize the clamp and enhance the resistance to pull. The high resistance to pull of this protein comes mostly from shearing eight contacts between terminal strand K and strand B (which is the second from the N-terminus). The next biggest contribution to the peak force comes from 11 contacts between anti-parallel positioned strands A and C.

5.8. Properties of the strongest proteins: non-typical mechanical clamps

In most cases, the mechanical clamp is made of two parallel β-strands set along the pulling direction. However, we have found that other possibilities exist, such as the three illustrated in figure 27 for proteins 1amx, 1qp1 and 1pav. In the first case, the clamp is made of anti-parallel β-strands. In the second case, the clamp consists of strand-like elements which do not form a secondary structure. Finally, in the third case, the clamp consists of a box motif.

5.8.1. Mechanical clamps made of anti-parallel β-strands. The sixth-ranked protein 1ei5 (CATH code 2.60.40.10), the 55th-ranked protein 1amx (2.60.40.740) and the 124th-ranked 1lm8 (3.10.20.90) belong to the β class. We focus on 1amx that was discussed in the context of the role of the temperature (figure 20). 1amx has two β-sheets, each made of five β-strands that are facing each other. However, in this case the terminal β-strands do not form the same β-sheet but instead belong to separate sheets. Moreover, in the entire structure, there are no two parallel β-strands. We found that the highest resistance of 1amx comes from shearing two anti-parallel β-strands B + I. If these bonds are cut the force is lowered by nearly 50%.

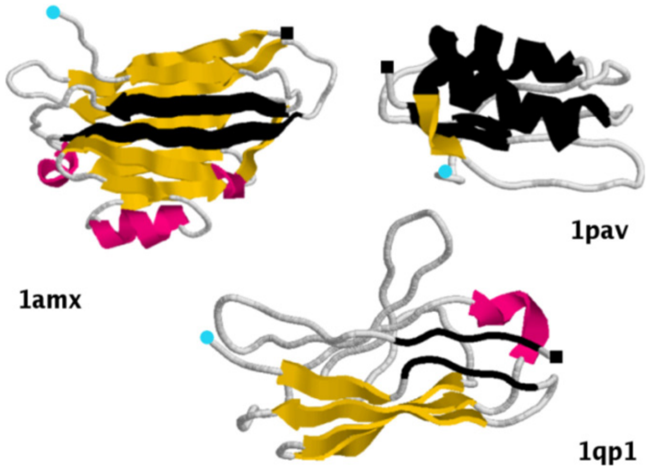

Figure 27. A cartoon representation of 1amx, 1pav and 1qp1. The segments shown in black are responsible for the biggest contribution to the peak force.

Additionally, contacts between the N-terminus and loops, between B and C (170–174, 195–205) and between (170–174, 195–201) have a 30% contribution. In the case of 1ie5, cutting the bonds between anti-parallel strands reduces the peak force by about 30% which does not seem too large at first sight. However, eliminating attractive contacts in other elements of the structure has much lower influence on the force, only around 5%.

5.8.2. Unstructured and delocalized mechanical clamps. In several proteins, such as 1qp1 and 1tum, the relevant mechanical clamps are unstructured, if we assume the strict criteria of what constitutes a hydrogen bond. In particular in the case of 1tum the average backbone distance between segments 2–6 and 78–82 is relatively large, 4.3 Å, and yet elimination of contacts in this region reduces $\tilde{F}_{\max}$ from 3.1 to 2.1. In the case of 1qp1, the relevant mechanical clamp is between segments 9–13 and 102–116 and its removal reduces $\tilde{F}_{\max}$ from 2.9 to 2.1.

Similar observations apply to the (11–19, 104–112) region in the transport protein 1f86 (cutting these bonds reduces $\tilde{F}_{\max}$ from 2.9 to 1.2) and to terminal region in the immunoglobulin-like protein 1b88 (a reduction from 3.2 to 2.1). In the latter case there is also another mechanical clamp between anti-parallel strands A and B—this clamp is weaker as elimination of the corresponding contacts reduces the force from 3.2 to 2.8.

We should point out that the identification of the β-strands used here was based on the information obtained from the PDB website. Our analysis proves, in a few cases, that the mechanical clamp is generated not just by a couple of β-strands but also by their immediate extensions in the sequence. This happens for instance in the case of 1ui9, 1pqe and 1mg4.

5.8.3. The box motif as a mechanical clamp. 1pav belongs to the α/β class. We found that the high mechanical resistance of this protein comes not from parallel or anti-parallel strands but from a box-like motif in which all 'walls' undergo shearing. The box is made of anti-parallel β-strands B (58–60) and E (73–76) and of two helices A (19–28) and C (44–57). The helices lie parallel to each other and are stretched in opposite directions. Removing the A + C, A + E, C + E and B + E bonds reduces the $\tilde{F}_{\max}$ of 3.0, nearly equally, by about 0.50 for each of these sets. In each of these sets of bonds there is a shear, though the contribution from the interhelical shearing to $F_{\max}$ is a bit smaller.

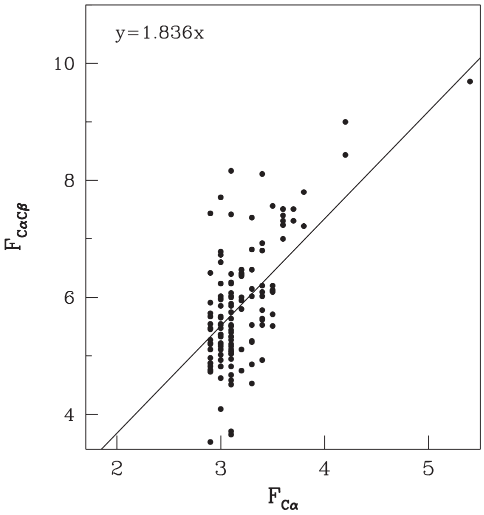

Figure 28. The cross plots between maximal peak forces calculated in the C^{α}-based and (C^{α} and C^{β})-based Go-like models in set S137.

We have found several proteins, such as 1lsl and 1vhp, in which the force of resistance is not associated with a well localized mechanical clamp but instead is spread out throughout the structure. For instance, in the case of 1lsl, the biggest localized contribution appears to be related to the third and fourth β-strands which are anti-parallel. However, the force reduction after the corresponding bond cutting is merely 18%, suggesting that most of the other contributions to resistance are distributed elsewhere and in small pieces.

5.9. Effects of the side groups

The main advantage of using the simple Go-like model is that it allows us to make a survey based on thousands of proteins. We now examine the issue of how robust the predictions remain when the model sheds some of its simplifications. We first examine the role of the side groups. Specifically, we ask what happens when the model based on C^{α} acquires internal degrees of freedom by incorporating the C^{β} atoms.

We restrict the analysis to the set S137 and we find that, in many cases, the change in the model does not affect the physics of stretching, other than shifting the effective energy scale of the model. This is shown in figure 28 which demonstrates a generally linear correlation between the peak forces determined in the two models with the coefficient of proportionality of 1.836. This means that the ranking of the proteins within the set is largely unaffected. In particular, the four top proteins remain at the top four places of the list for the more refined model. Nevertheless, there are certain outlier cases. For instance, proteins 1tum, 1pun and 1kot move from positions 32, 68 and 115 in the C^{α}-based model to positions 5, 6 and 9, respectively, in the model with the side groups. There are also proteins which move down the ladder when including the side groups. These are: 1kiq, 1a2y, 1lve, 1amx, 1qd0, 1j05, 1rbj, 1vfb, 1gke, 1etb, 1ict, 1em7, 1com, 1rnz, 1tjn, 5lve and 1dvt.

Some of the outlier cases will be discussed below and all data for S137 are summarized in table 11. We begin by defining the model with the side groups.

Table 11. The same as table 3 (for proteins with $N < 150$), but predicted by the Go-like model with side group C^{α}–C^{β} used in this paper.

Rank	PDB	N	$F_{\max}$ (ϵ Å^{-1})	L_{n} (Å)	L_{m} (Å)	L_{f} (Å)
1	1c4p	137	9.7	50.4	219.2	516.8
2	1g1k	143	9.0	43.5	158.1	539.6
3	1qqr	138	8.9	52.3	210.5	520.6
4	1aoh	138	8.4	34.0	142.1	520.6
5	1pun	129	8.2	34.2	299.4	486.4
6	1tum	129	8.1	33.7	295.5	486.4
7	1ie5	107	7.8	51.5	167.5	402.8
8	1kot	119	7.7	12.7	283.8	448.4
9	1so9	131	7.6	40.6	120.0	494.0
10	1c76	136	7.5	29.0	155.5	513.0
11	1yn4	99	7.5	35.4	128.3	372.4
12	2sak	121	7.5	31.9	169.5	456.0
13	1v80	76	7.4	37.1	133.1	285.0
14	1rlf	90	7.4	42.4	58.8	338.2
15	1c78	136	7.4	27.2	206.6	513.0
16	1npu	116	7.4	40.6	136.5	437.0
17	1ppx	129	7.3	36.0	280.2	486.4
18	1v5o	102	7.3	63.8	187.5	383.8
19	1c77	136	7.2	27.2	206.7	513.0
20	1c79	136	7.2	27.2	208.7	513.0
21	1ssn	136	7.2	9.5	244.0	513.0
22	1sp0	131	7.0	40.6	180.1	494.0
23	1pgx	83	6.9	62.7	134.9	311.6
24	2ncm	99	6.8	42.1	56.9	372.4
25	1hz6	72	6.8	41.6	127.2	269.8
26	1k26	156	6.8	46.1	289.3	589.0
27	1b9r	105	6.7	29.1	120.5	395.2
27	1pav	78	6.6	13.0	180.5	292.6
29	1jrk	156	6.5	42.2	215.0	589.0
30	1gnu	117	6.6	5.0	263.6	440.8
31	1vhp	117	6.5	40.8	109.1	440.8
32	1b88	114	6.5	39.0	120.8	429.4
33	1vjk	88	6.4	31.0	115.2	330.6
34	1eo6	117	6.4	22.1	235.5	440.8
35	1sn5	130	6.4	21.5	167.5	490.2
36	1pus	129	6.4	36.1	269.2	486.4
37	1oau	122	6.4	43.4	116.2	459.8
38	1h8c	82	6.4	39.5	118.5	307.8
39	1tvd	116	6.3	38.0	112.9	437.0
40	1ves	113	6.2	35.1	108.8	425.6
41	1oar	122	6.2	43.5	112.8	459.8
42	1sn0	130	6.2	21.5	180.0	490.2
43	1nme	146	6.2	52.8	347.1	551.0
44	1h5b	113	6.1	40.5	118.2	425.6
45	1oo2	119	6.1	12.9	166.5	448.4
46	1i3v	129	6.1	40.4	103.6	486.4
47	1i9e	115	6.1	49.5	63.4	433.2
48	1sn2	130	6.1	21.6	90.4	490.2
49	1w19	147	6.1	21.6	348.2	554.8

Table 11. (Continued.)

Rank	PDB	N	$F_{\max}$ (ϵ Å^{-1})	L_{n} (Å)	L_{m} (Å)	L_{f} (Å)
50	1m94	73	6.0	27.4	37.1	273.6
51	1km7	100	6.0	35.8	94.7	376.2
52	1ugm	113	6.0	21.9	136.2	425.6
53	1w29	146	6.0	19.4	206.5	551.0
54	1kgi	127	6.0	19.2	98.3	478.8
55	1wiu	93	6.0	39.0	49.2	349.6
56	1oax	122	5.9	43.7	56.7	459.8
57	1hz5	72	5.9	33.0	77.5	269.8
58	1k53	72	5.9	32.5	77.5	269.8
59	1mfw	107	5.9	12.4	140.4	402.8
60	1bvk	108	5.8	39.7	50.7	406.6
61	1fmf	137	5.8	18.8	319.1	516.8
62	1ivl	107	5.8	36.5	50.8	402.8
63	1wtl	108	5.8	40.4	53.7	406.6
64	1anu	138	5.8	24.0	35.4	520.6
65	1oaq	120	5.7	40.6	49.5	452.2
66	1ieh	135	5.7	51.1	113.3	509.2
67	1eaj	126	5.7	41.6	56.5	475.0
68	1f86	115	5.7	12.1	79.2	433.2
69	1i3o	144	5.7	40.4	0.0	543.4
70	1lqb	118	5.7	38.9	130.4	444.6
71	1bm7	127	5.6	12.2	76.2	478.8
72	1dfu	94	5.6	13.9	121.8	353.4
73	1f5w	126	5.6	41.4	65.9	475.0
74	1eta	127	5.6	12.6	125.4	478.8
75	1lsl	113	5.5	100.1	275.5	425.6
76	1bz8	126	5.5	20.7	121.2	475.0
77	1pga	56	5.5	26.5	30.4	209.0
78	1kip	107	5.5	36.3	46.1	402.8
79	43c9	113	5.5	36.5	47.5	425.6
80	2try	127	5.5	10.8	124.3	478.8
81	1nam	116	5.5	35.2	47.6	437.0
82	1ufy	122	5.5	30.8	80.2	459.8
83	1fvc	109	5.5	40.7	50.9	410.4
84	2igd	61	5.5	40.6	49.0	228.0
85	1tyr	127	5.5	10.6	124.9	478.8
86	1mel	148	5.4	38.0	45.3	558.6
87	1mg4	113	5.4	5.9	120.8	425.6
88	1igd	61	5.4	40.4	50.4	228.0
89	1tbe	76	5.4	33.5	48.0	285.0
90	1gb4	57	5.4	28.9	34.9	212.8
91	1p7e	56	5.3	26.3	31.5	209.0
92	1nvi	81	5.3	34.1	45.5	304.0
93	1ui9	122	5.3	27.9	60.9	459.8
94	1wit	93	5.2	39.3	50.4	349.6
95	1ttc	127	5.2	12.8	126.1	478.8
96	1mvf	135	5.2	38.7	140.3	509.2
97	1bzd	127	5.2	11.0	123.4	478.8
98	1i8k	107	5.2	34.3	46.4	402.8
99	1qp1	107	5.2	36.7	46.5	402.8
100	2dlf	113	5.2	39.3	54.1	425.6

Table 11. (Continued.)

Rank	PDB	N	$F_{\max}$ (ϵ Å^{-1})	L_{n} (Å)	L_{m} (Å)	L_{f} (Å)
101	4lve	114	5.2	39.8	52.9	429.4
102	1lm8	106	5.2	43.3	137.7	399.0
103	1bmz	127	5.2	11.5	74.6	478.8
104	1tfp	130	5.2	12.1	78.9	490.2
105	1dvy	124	5.1	12.4	120.0	467.4
106	1l2n	81	5.1	36.0	54.8	304.0
107	1ie4	127	5.1	15.3	97.3	478.8
108	1pqe	126	5.1	33.4	72.8	475.0
109	1kmt	141	5.0	44.5	71.5	532.0
110	1c08	107	5.0	35.3	41.9	402.8
111	1kir	107	5.0	36.0	44.6	402.8
112	2imm	114	5.0	39.9	52.9	429.4
113	1py9	116	5.0	40.6	51.0	437.0
114	1f2x	135	4.9	40.0	53.3	509.2
115	1gko	127	4.9	11.6	79.5	478.8
116	1kiq	107	4.9	36.2	47.8	402.8
117	1vfb	107	4.9	36.6	44.6	402.8
118	1em7	56	4.8	25.9	31.8	209.0
119	1tjn	125	4.8	34.1	99.7	471.2
120	1lve	122	4.8	39.2	51.6	459.8
121	2rox	127	4.8	14.0	87.8	478.8
122	1gke	120	4.8	15.1	99.8	452.2
123	1etb	127	4.8	8.6	89.7	478.8
124	5lve	114	4.8	36.8	53.4	429.4
125	1ic4	107	4.7	35.4	48.9	402.8
126	1com	127	4.7	25.7	53.0	478.8
127	1jf8	131	4.7	17.3	214.8	494.0
128	1dvt	115	4.7	12.2	79.1	433.2
129	1jhl	108	4.6	39.1	50.5	406.6
130	1ict	127	4.6	9.5	84.2	478.8
131	1n4x	113	4.5	33.7	52.4	425.6
132	1a2y	107	4.5	37.0	157.7	402.8
133	1qd0	128	4.5	40.2	45.7	482.6
134	1rbj	124	4.1	32.0	247.7	467.4
135	1amx	150	3.7	32.1	170.2	566.2
136	1j05	111	3.6	36.2	0.0	418.0
137	1rnz	124	3.5	37.1	247.7	467.4

5.9.1. The Go model with side groups. We generalize our model to include the side groups as represented by the locations of the C$^{\beta}$ atoms in the residues. These C$^{\beta}$ atoms are linked to the C$^{\alpha}$ on the same amino acid by a harmonic tethering term with a minimum at a location $\vec{r}_i^{\,\mathrm{C}^{\beta}}$ as calculated based on the placement $\vec{r}_i^{\,\mathrm{C}^{\alpha}}$ of the corresponding C$^{\alpha}$ atom and of its sequential neighbours along the chain. The distance $l = |\vec{r}_i^{\,\mathrm{C}^{\beta}}|$ is set at 1.5 Å. The directional characteristics are described by

$$\vec{r}_i^{\,\mathrm{C}^{\beta}} = l(\hat{a}\cos\theta + \hat{b}\sin\theta) \tag{8}$$

which was deduced from studies of the peptide geometry [127, 128]. Here, the angle θ is chosen optimally to be equal to 37.6° and

$$\hat{a} = \frac{\hat{s}_{i,i-1} + \hat{s}_{i,i+1}}{|\hat{s}_{i,i-1} + \hat{s}_{i,i+1}|} \qquad \hat{b} = \frac{\hat{s}_{i,i-1} \times \hat{s}_{i,i+1}}{|\hat{s}_{i,i-1} \times \hat{s}_{i,i+1}|} \tag{9}$$

where $\hat{s}_{i,j}$ is a unit vector defined by

$$\hat{s}_{ij} = \frac{\hat{r}_i^{C^\alpha} - \hat{r}_j^{C^\alpha}}{|\hat{r}_i^{C^\alpha} - \hat{r}_j^{C^\alpha}|}. \tag{10}$$

The presence of native contacts between different amino acids i and j is again checked by studying atomic overlaps with the use of the procedure based on the van der Waals radii [108]. The contacts may now arise between C_i^β and C_j^β (if the effective atoms on the side groups overlap), between C_i^α and C_j^β (if the side group on j overlaps with C^α on i), between C_i^β and C_j^α and finally between C_i^α and C_j^α (which arise primarily within secondary structures). Each of the existing contacts is represented by the same Lennard-Jones potential with the energy scale ϵ and with a minimum located at the native distance between the interacting entities (e.g. between C_i^β and C_j^β). Thus interactions between two amino acids in contact involve between one and four Lennard-Jones terms. The pulling speed is kept the same as in the C^α-based model.

The trend shown in figure 28 suggests that the mean number of such terms is close to 1.8 which explains the energy 'conversion factor' between the standard Go model and its version with the side groups when discussing forces. The conversion factor is required because the standard model is defined by one energy parameter ϵ in a contact whereas the finer model may involve several 'subcontacts', each with the strength of ϵ. On the other hand, studies of the melting temperature, T_f, suggest a conversion factor of 1.2. Finally, our studies of the kinetics of folding indicate that a typical reduced temperature of optimal folding shifts from around 0.3 to only around 0.4. These circumstances indicate that a simple conversion of one model into another is not straightforward. We perform the studies of stretching at $\tilde{T} = 0.4$—this is our effective 'room temperature' for the model with the C^β atoms.

Figure 29 is an analogue of figure 3 and displays the experimental results against the theoretical predictions based on the finer model. It is seen that the overall trend is less clearly defined than in the case of the C^α-based model and the Pearson coefficient is equal to 0.80. Thus adding more structure to the model does not necessarily makes it better when confronted with the experiment. Nevertheless, we can still infer what the side groups might do in a real system qualitatively by considering the model with the side groups.

5.9.2. Influence of the side groups on the F–d pattern. We have found that inclusion of the C^β atoms changes the nature of the *F–d* pattern in 52 out of 137 cases in set S137. The typical changes involve (1) vanishing of a minor peak, (2) emergence of an extra minor peak (e.g. in 1eih, 1ivl, and 1vhp) and (3) adjustments in relative heights of the peaks. The first of these is the most common and it usually leads to disappearance of a peak that is located immediately after the major peak. However, even if the pattern does not change, the identification of the mechanical clamp may still shift because the number of the Lennard-Jones terms in a contact undergoes modulations.

The I27 domain of titin (not in the set S137) offers an example of the most typical situation in which a minor peak disappears on including the C^β atoms in the model. The *F–d* curves and the scenarios of unfolding in both models are shown in figure 30. It is seen that the separate events that yield two peaks in the C^α-based model (the second peak is due to breaking the C + F and B + E contacts, as defined in the caption of figure 30) merge together in the C^α–C^β-based model and yield just one peak. This indicates that the latter model leads to more engrossing elastic couplings and thus a larger cooperativity in behaviour than the simpler model.

It is interesting to note that the disappearance of the minor peak does not involve any change in the contact map *per se* but is only due to an effective enhancement of certain contacts

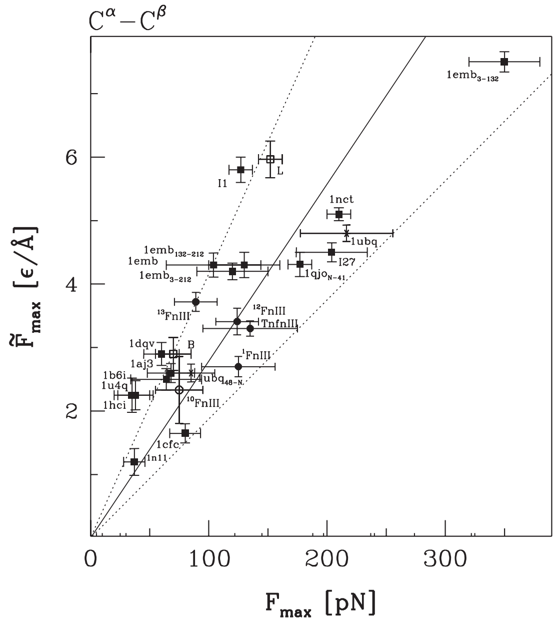

Figure 29. The same as figure 3 but for the C^α–C^β model.

relative to the other. This is illustrated in figure 31 which compares the uniform strength contact map of the C^α-based model to the inhomogeneous strength contact map of the C^α–C^β-based model. It should also be noted that, in the finer model, the A + B contacts, responsible for the so-called intermediate state, break clearly ahead of the main peak compared to the simple model (in which this happens only at very low temperatures). Also the F + G contacts (the pentagons in figure 30) unravel much later. The C^α–C^β scenarios do not depend much on which kind of atoms, α or β, are involved in making the contact.

There are 27 proteins, e.g. 2sak, in set S137 which behave in a similar way to titin. A multi-peak variant of it is realized by eight proteins. Among them, there is 1oo2, presented in figure 32. In this case, switching to the finer model results in the disappearance of two minor peaks: the second and the fourth. As discussed in the context of figure 25, the mechanical clamp arises primarily due to the A + I contacts and making the clamp work involves unzipping the I + J contacts. The second peak in the simple model is due to interactions between side groups of A + G, A + F and A + C. In the C^α–C^β model these interactions unravel simultaneously with the mechanical clamp. Likewise, the C + D contacts responsible for the fourth peak get ruptured together with C + F, H + F, H + G, and A + B to form a second peak in the finer model.

A protein known as human GABA receptor is an especially interesting case as a prediction of its elastic properties within the finer model seems to be very sensitive to the precise knowledge of its structure. This protein has a UB-roll topology and two determinations of its structure have been deposited in the PDB: 1kot and 1gnu. The former was obtained by NMR and the latter by x-ray techniques. The contact maps of the two structures are somewhat different. In particular, 1gnu lacks the E + I and F + I contacts, where E(58–71), F(79–81), I(110–112) and their RMSD is 2.86 Å. In the C^α-based models of 1kot and 1gnu, the mechanical clamps seem to be dominated by the C + H contacts (30–34, 106–108) and the two *F*–*d* curves for the two systems are nearly identical. When we consider the 1kot structure then the inclusion of the C^β atoms pushes the force rank of 1kot from place 115 to place 8 whereas

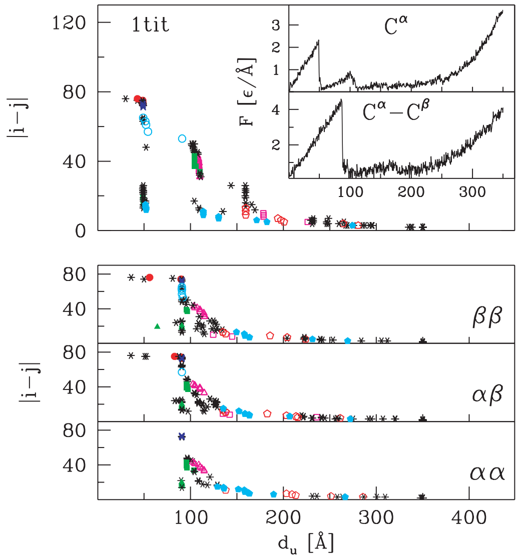

Figure 30. Unfolding scenarios for 1tit obtained in the C^α and C^α–C^β-based models is shown in the top and bottom panels, respectively (at their corresponding temperatures of optimal folding). Additionally, the bottom panel is divided into three subpanels, which show the scenarios of unfolding for the specific types of contacts: C^β–C^β, C^α–C^β and C^α–C^α, top to bottom respectively. The right-hand corner of the top panel shows the F–d curves for the two models. The A, A′, B, C, D, E, F and G strands in titin correspond to the sequential segments 4–7, 11–15, 18–25, 32–36, 47–52, 55–61, 69–75 and 78–88, respectively. The symbols assigned to specific contacts are the same in all scenario panels. Open circles, open triangles, open pentagons, solid circles and solid squares correspond to contacts B–G, B–E, D–E, A–G and C–F, respectively. The stars denote all other contacts.

for 1gnu the ranking is practically unchanged. Observing the F–d curves may then offer ways for an independent experimental determination of the structure.

The dynamical properties, as deduced from the 1kot structure, are shown in figure 33, together with the unfolding scenarios. It is seen that switching to the C^α–C^β-based model does not change the pattern before the major peak but it merges the major peak with one after-peak. The events generating the after-peak in the C^α-based model are now linked to the events leading to the major peak itself, very much like what we have observed for titin.

In the case of the unstructured 1tum, the two variants of the Go-like model yield F–d curves of a similar shape (not shown). However, the nature of the relevant mechanical clamp changes. Instead of the clamp generated by the U1 + C contacts (2–8, 78–86) in the simple model, the clamp in the finer model is primarily due to A + U2 (46–55, 20–25) which are primarily of the α–β and β–β kind. This switch affects the ranking in a forward way. We observe similar phenomena for 1pun, 1ppx, 1puq and 1pus. A backward motion in the ranking also takes place in a few cases such as 1amx. For this protein, the order of unfolding events is affected by the choice of model. Choosing the C^α–C^β model results in more unzipping than shearing and leads to a relative lowering of the maximum unravelling force.

The C^α–C^β-based description, being more refined, should be more trustworthy than the C^α-based modelling when discussing the microscopic mechanisms of rupture. However, in most cases the simpler model appears to correlate forces with the experimental results better, at least when no differentiation between the values of the effective energy parameter ϵ for various kinds of coupling is made.

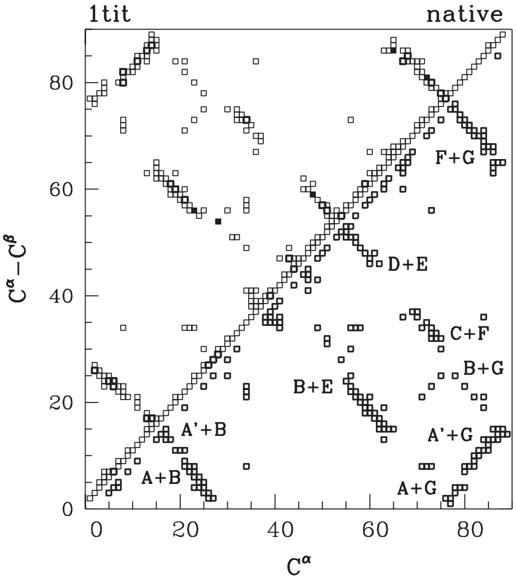

Figure 31. The contact map of the I27 domain of titin as represented by the C^α-based (below the diagonal) and the C^α–C^β-based (above the diagonal) models. The thickness of the symbol, in the latter case, is proportional to the number of interactions involved in the coupling between a given pair of amino acids. The strands participating in the coupling are indicated by letters.

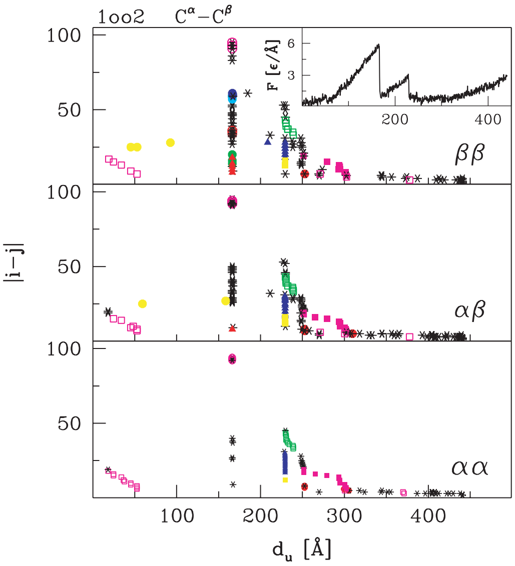

Figure 32. The C^α–C^β version of figure 25 for protein 1oo2.

5.10. The role of the disulfide bonds

In the basic C^α-based model used in the survey, no provision is taken of the fact that contacts formed between cysteines may correspond to disulfide bonds. Such bonds are covalent in nature

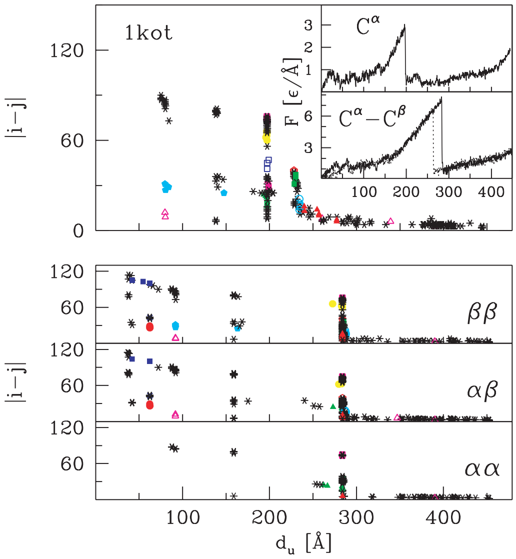

Figure 33. Similar to figure 30 but for structure 1kot. The dotted F–d curve corresponds to structure 1gnu corresponding to the same protein as 1kot. The A, B, C, D, E, F, G, H and I strands in 1kot correspond to the sequential segments. Blue solid square A + H, magenta open squares C + H, yellow solid circles C + G, blue open squares E + I, solid triangles magenta F + I, green solid triangles C + D, magenta solid triangles B + C, cyan open circles E + F, red solid triangles G + H, green solid squares E + G, red solid pentagons A + C.

and cannot be ruptured. When the disulfide bonds are represented by the standard Lennard-Jones potential, they do rupture and yield incorrect F–d curves.

This problem is illustrated in figure 34 for bovine ribonuclease A with the code 1rnz. The standard model leads to a maximum force peak of 2.9 ϵ Å^{-1} that occurs around $d = 250$ Å. This protein is listed at position 127 in table 3. There are four contacts corresponding to the disulfide bonds. Two of them break before reaching $F_{\max}$ and two contribute to $F_{\max}$. Disallowing for the rupture of the four contacts is likely to affect the stretching process and can be effectively accomplished by rescaling the value of the energy parameter ϵ in the contacts by a factor of 20 or more. Figure 34 shows that this enhancement modifies the F–d curve. It makes the major force peak occur earlier and the value of $\tilde{F}_{\max}$ increases to 4.0. Thus the protein advances to the fifth position in the overall ranking. The scenario diagram demonstrates that the first half of the adjusted stretching process proceeds as without any energy rescaling but then it starts to differ markedly. In particular, the list of contacts that contribute to the force clamp is modified substantially. A similar situation takes place in the case of the homologous protein 1rbj, also with four sulfide bonds, for which $\tilde{F}_{\max}$ jumps from 3.0 to 3.5. The ranking for 1rbj is just changed from top 100 to top 20.

The very presence of the disulfide bonds need not necessarily affect the value of $F_{\max}$, especially if their rupture is scheduled to occur past the major peak. In set S137 there are only 14 proteins, in addition to 1rnz and 1rbj, for which the standard model predicts that the disulfide bond rupture before or at the major peak. We have reconsidered these proteins and found essentially no change for four of them, 1ie5, 1i8k, 1kiq, 1lve, a slight increase for two of them, 1py9 (by 0.1), 43c9 (by 0.2), and a slight reduction for seven of them, 1h5b (by 0.15), 1c08 (by 0.2), 1ivl (by 0.1), 1kir (by 0.1), 1kip (by 0.2), 1n4x (by 0.1), 1vfb (by 0.1). The only case of a major effect of the presence of the disulfide bonds constitutes 1lsl ($N = 113$) with the

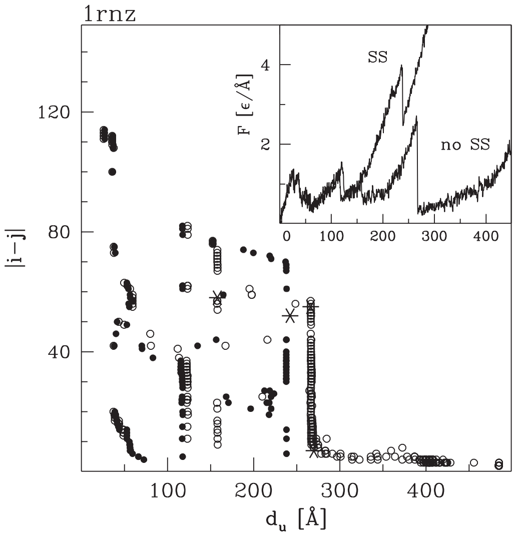

Figure 34. Unfolding scenario and the F–d curves, in the inset, for protein 1rnz in two Go-like models. One model is standard and assumes no special treatment of the disulfide bonds between the cysteines. It yields the thinner F–d curve and the contact breaking distances corresponding to open circles and stars. The stars indicate contacts between cysteines. The other model does not allow for rupture of the disulfide bonds. It yields the thicker F–d curve and the data points corresponding to black circles in the scenario diagram.

standard Go model value of $\tilde{F}_{\max}$ equal to 3.0. This protein has six disulfide bonds. The first of them is affected early on which effectively restricts the stretching process to a mere 37 amino acids which leads to no force peak.

Generally, however, the inclusion of the disulfide bonds affects the list of the strongest proteins in a minor way only: the protein 1lsl should be dropped from set S137, and 1rnz and 1rbj should be advanced in the ranking.

It should be noted that comparing the stretching process in the two Go-like models, with the energy rescaling and without, is physically meaningful since the disulfide bonds can be converted to much weaker SH bonds by application of the reducing agent dithiothreitol (DTT). Such experimental studies have been performed with cell adhesion molecules Mel-CAM [129], and the singly domained VACM-1 [130]. Thus the ordering of proteins by force as presented in table 3 may be considered as corresponding to a prior treatment by DTT whenever the disulfide bonds are involved.

5.11. Stretching of type III fibronectins

We now consider the special case of fibronectin. The natural fibronectin (FN) is a giant protein which contains more than 40 domains. This protein is well studied since it is an important component of the extracellular matrix and is involved in tissue elasticity, cell adhesion and cell migration. FN contains modules of three different structural types. Among these, there are modules of type III FN which have the topology of the immunoglobulin-fold β-sandwich consisting of seven β-strands. They contain binding sites for the cell surface receptors. FNIII, like titin, usually remains under tension in physiological conditions. ^{10}FNIII contains an RGD loop which plays an important role in binding to the extracellular matrix.

The FNIII modules which have been experimentally stretched so far are the bovine fibronectin pFN, the native fragment $^{2-14}$FNIII and its domains like ^{1}FNIII, ^{2}FNIII, ^{3}FNIII,

$^{7-10}$FN, $^{12-13}$FNIII and FNIII from the titin segments I48–I54, I54–I59. It has been found [61] (see also table 1) that unfolding of the weakest module requires a force of around 20 pN whereas 220 pN is needed to unravel the strongest module, generating a hierarchy in the dynamical behaviour. Experiments on particular domains show that most of them are weaker than the I27 domain of titin. A bigger mechanical resistance was found for ^{1}FNIII, ^{2}FNIII, and a comparable force for I48–54, I54–59. Generally, however, the resistance of a FNIII module appears to depend on what other modules or proteins it is connected to in a tandem [61].

In order to elucidate the experimental findings on FNIII, we have modelled stretching of those FNIII modules for which the structure is known individually, as for ^{1}FNIII (1oww), ^{3}FNIII (1ten) and ^{10}FNIII (three structures: 1fna, 1ttg, 1ttf), and for multiple connections such as for the natural tandem linkages of $^{7-10}$FNIII (1fnf) and $^{12-14}$FNIII (1fnh). Notice that the native coordinates of ^{10}FNIII are also represented as a part of the 1fnf tandem structure representing four domains. We now return to the C$^{\alpha}$-based model.

5.11.1. *Stretching of the ^{9}FNIII and ^{10}FNIII domains individually.*

We have found that ^{10}FNIII with the PDB codes 1ttg and 1ttf (both obtained through NMR studies) have $\tilde{F}_{\max} = 0.89 \pm 0.12$ and 1.05 ± 0.14, respectively. Both of these results agree with the experimental data if one uses the I27 domain of titin as a benchmark of force. However, using the ^{10}FNIII coordinates extracted from the tandem $^{7-10}$FNIII (1fnf, obtained through x-ray studies) leads to a much higher value of around 1.8 ± 0.18. An average over the three values is shown as an open circle in figure 3. The F–d curves and the unfolding scenarios for ^{10}FNIII corresponding to 1fnf and 1ttf are shown in figure 35. The two sets differ by 1.3 Å in their RMSD and the differences are located primarily near the N-terminus and in the A and G β-strands where there are fewer contacts for 1ttf. As a result, the events leading to big force peaks in 1fnf, like rupturing of A + B and F + G, in the case of 1ttf separate in time significantly and lead to lower peaks. Our model-based results suggest that perhaps 1ttf represents a more accurate structure as its corresponding peak force agrees with the overall trend better.

Our results on ^{9}FNIII (the native coordinates are obtained by isolating ^{9}FNIII from $^{7-9}$FNIII) suggest a larger mechanical stability than for ^{10}FNIII which disagrees with one all-atom simulation [83] but agrees with another [80] and one experimentally derived conclusion [61]. A direct comparison with experiment is not possible because the experimental linkages involved are heterogeneous. In agreement with [83], however, we observe that the scenarios of unfolding for ^{9}FNIII and ^{10}FNIII are quite distinct. ^{9}FNIII starts to unfold by unravelling the contacts that involve the F strands. When about 75% of all contacts are broken, a crucial contact between Arg6 and Asp23 unravels in conjunction with A + B.

5.11.2. *Stretching of the tandem $^{9-10}$FNIII and $^{7-10}$FNIII.*

We find that, in the tandem linkage $^{9-10}$FNIII (see figure 36), the domains unfold independently of each other. Unravelling starts from ^{9}FNIII indicating a weaker stability of this domain. Thus connecting ^{10}FNIII with ^{9}FNIII appears to influence the mechanical properties of this domain by reversing their relative stabilities. The maximal force of the first peak of ^{9}FNIII is a bit higher than expected based on the single-domain simulations. This indicates that some contacts of ^{10}FNIII contribute to it.

Comparison of the mechanical resistance between ^{10}FNIII and ^{9}FNIII, ^{8}FNIII, ^{7}FNIII stretched separately shows that the strengths of these modules could be ranked from ^{10}FNIII, through ^{9}FNIII, ^{7}FNIII to ^{8}FNIII from the weakest to the strongest. This ranking agrees with the all-atom simulations [83] under the conditions of constant force which predict that ^{10}FNIII should be mechanically weaker than the seventh, eighth and ninth domains of FNIII. Moreover, when we stretch the four-domained $^{7-10}$FNIII, ^{10}FNIII appears to be the weakest of the four, as

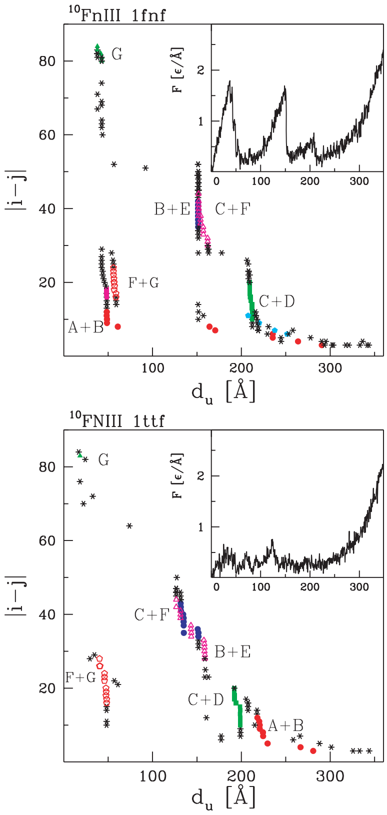

Figure 35. Unfolding F–d curves and the corresponding unfolding scenarios for the 10th domain of FNIII from the $^{7-10}$FNIII tandem (upper panel) and 1ttf (lower panel).

demonstrated in figure 37. This agrees with the experimental [61] and theoretical results [80] but contradicts our finding on the tandem $^{9-10}$FNIII. The experimental finding has relied on assigning the first force peak to ^{10}FNIII. The undecided nature of these relative rankings may indicate the existence of significant mutual interactions between various domains even though they unfold in a predominantly serial fashion. Another possibility is that the quality of the PDB coordinates is not sufficient to settle the issue in our model.

We find that when considering the C$^{\alpha}$–C$^{\beta}$ based model the minor force peaks get absorbed by the major peaks corresponding to specific domains.

5.11.3. Unfolding scenario for ^{10}FNIII and ^{9}FNIII. All-atom molecular dynamic simulations [93, 83] have predicted that ^{10}FNIII unfolds through at least two pathways. ^{10}FNIII unfolds by an intermediate state, which contains the extended A and B strands and the native RGD loop. Paci and Karplus [93] have identified still another unfolding intermediate state for ^{10}FNIII, where the A and G strands are detached from the remainder of the protein. Experimental results [49] suggest that ^{10}FNIII unfolds in two consecutive steps. The native state unfolds at 100 ± 20 pN and reaches an intermediate state, which then unfolds at 50± pN. The experiments also indicate the existence of a few cases when ^{10}FNIII unfolds directly from the native to an unfolded conformation. This is observed as a disappearance of the intermediate state (one of the peaks in the F–d curve). The substitution of several amino acids [49] has led

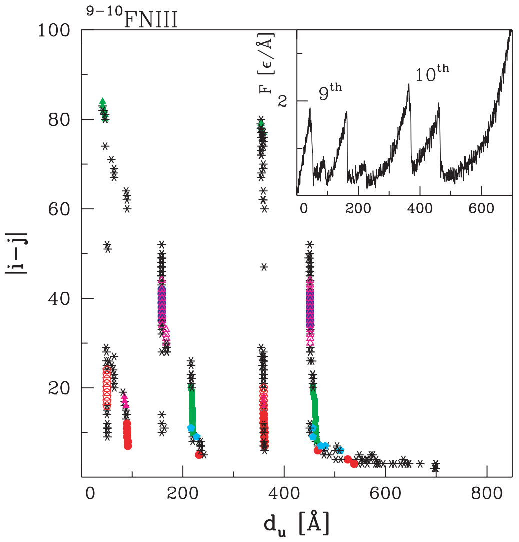

Figure 36. Unfolding scenario for the the tandem $^{9-10}$FNIII domains. Symbols indicate breaking of particular contacts; notation is the same as in figure 3. The inset shows unfolding F–d curves for the same tandem.

to a conclusion that, in one of the pathways, the A and B β-strands detach and give rise to an intermediate state. Along the second pathway, the G β-strand dissociates itself from the folded module in the first unfolding step.

Our model pathways shown in figure 35 are found to be typical when multiple unfolding simulations are considered. The F–d pattern obtained for 1ttf (and 1ttg) appears to correspond to the experimentally rarely observed unfolding from native to unfolded conformation whereas the 1fnf-based data point to the existence of the intermediate state and hence to two well articulated peaks.

5.11.4. Stretching of the 12th, 13th and 14th domains individually and in a tandem $^{12-13}$FNIII. The 12th to the 14th domains have been studied only experimentally so far. Oberhauser *et al* [61] have shown that the ^{12}FNIII domain is mechanically more stable than the ^{13}FNIII domain (120 and 90 pN respectively) and all of them are less stable than the I27 domain of titin. Our model results are consistent with the latter finding but suggest a different ranking: ^{12}FNIII, ^{14}FNIII and ^{13}FNIII when listing from the weakest to the strongest. This ranking also arises in the tandem arrangement $^{12-14}$FNIII as shown in figure 38, though the variations between the peaks are not too large. It is interesting to note that the model predicts ^{13}FNIII to unravel at two different stages with the unravelling of ^{14}FNIII taking place between the stages.

It should be noted that the experimental results [61] on the mechanical stability of ^{12}FNIII and ^{13}FNIII have been obtained by considering tandem arrangements with the I27 domain of titin which is likely to affect the pulling geometry. We find that when considering the C^{α}–C^{β}-based model the ranking changes: ^{12}FNIII,^{13}FNIII, ^{14}FNIII when listing from the weakest to the strongest.

The summary of our studies of fibronectin is that the Go-like model provides a reasonable account of the wide ranging hierarchy of forces that has been demonstrated experimentally for this system. The weakest domain in the model is ^{10}FNIII, as represented by 1ttf, and the strongest is FNIII as represented by 1nct—the I54 domain of titin. Our simulations for 1nct yield $F_{\max}$ just exceeding the one obtained for 1tit. However, several details are not well

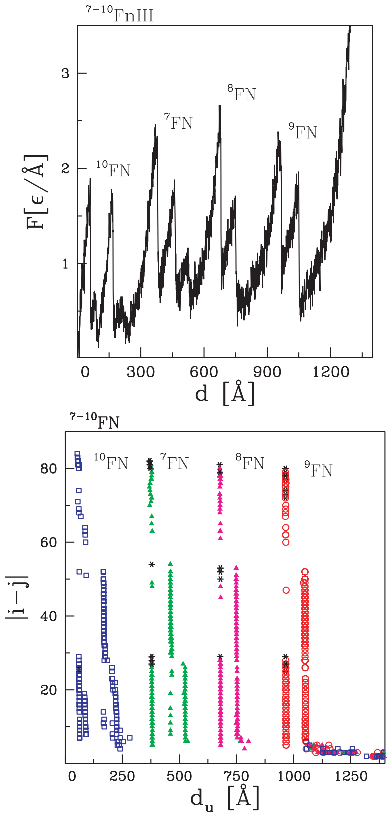

Figure 37. Similar to figure 36 but for the tandem of four domains, $^{7-10}$FNIII. Upper panel: unfolding F–d curves. Lower panel: unfolding scenario. Symbols indicate the breaking of particular contacts.

represented. For instance, theoretical stretching of the linkage $^{7-10}$FNIII suggests a ranking of forces associated with the individual domains that agrees with the experiment but yields an excessive magnitudes of the forces. The ranking corresponding to the tandem $^{12-13}$FNIII is opposite to that obtained in the experiment.

6. Conclusions

We conclude that a simple coarse grained Go-like model yields insights into the mechanical properties of proteins. A good correlation of the experimental results on stretching with those obtained within the simple model allows one to use the model to make comparisons between proteins. We provide a list of proteins that are predicted to be especially strong when undergoing stretching. These proteins belong to a short list of topologies and their strength arises from a mechanical clamp which, in most cases, consist of long and parallel β-strands. Taking into account refinements in the model, such as the presence of side groups and of the disulfide bridges, reduces the set of short strong proteins S137 by one entry but otherwise merely reshuffles the ranking of the proteins. Significantly, the reshuffling does not affect the very top of the list. This suggests that the selection principle of strong proteins based on the Go model has a good chance of being confirmed by all-atom simulations. The real test, however, should be provided by experiments. One interesting experiment could involve stretching of the

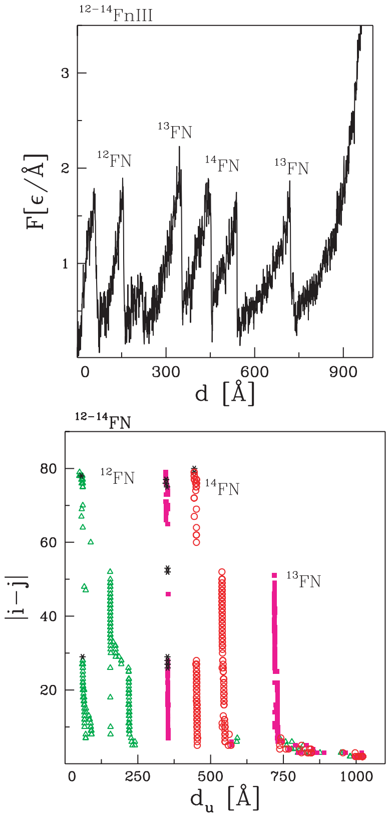

Figure 38. The same as figure 37 but for the tandem of three domains, $^{12-14}$FNIII. Particular symbols indicate the stretching of different domains.

ribonuclease 1rnz and comparing it to a system in which the disulfide bonds are replaced by weaker contacts. We anticipate that in both cases the protein should be sturdy mechanically but that the *F*–*d* patterns should be distinct. Another experiment could involve studies of the non-standard mechanical clamps. In order to facilitate the experimental studies, we plan to set up a website that will provide Go model-based findings on stretching for the protein structures deposited in the PDB.

Acknowledgments

This project was started by an inspiring conversation with Julio M Fernandez at the IOP conference in Warwick, England in April 2005. Julio M Fernandez also provided many interesting suggestions regarding the content of this review such as undertaking studies of fibronectin. This project was also helped by plenty of advice and many good thoughts from Jayanth R Banavar. We appreciate the 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 N202 021 31/0739) and by the European programme IP NaPa through Warsaw University of Technology.

References

[1] Ritort F 2006 Single-molecule experiments in biological physics: methods and applications *J. Phys.: Condens. Matter* **18** R531–83

[2] Grubmuller H, Heymann B and Tavan P 1996 Ligand binding: molecular mechanics calculation of the straptavidin biotin rupture force *Science* **271** 997–9

[3] Bockelmann U, Essevaz-Roulet B and Heslot F 1997 Molecular stick–slip motion revealed by opening DNA with piconewton forces *Phys. Rev. Lett.* **79** 4489–92

[4] Erickson H P 1997 Stretching single protein molecules: titin is a weird spring *Science* **276** 1090–2

[5] Linke W A, Ivemeyer M, Olivieri N, Kolmerer B, Ruegg J C and Labeit S 1996 Towards a molecular understanding of the elasticity of titin *J. Mol. Biol.* **261** 62–71

[6] Tskhovrebova L, Trinick K, Sleep J A and Simmons M 1997 Elasticity and unfolding of single molecules of the giant muscle protein titin *Nature* **387** 308–12

[7] Kellermayer M S Z, Smith S B, Granzier H L and Bustamante C 1997 Folding–unfolding in single titin molecules characterized with laser tweezers *Science* **276** 1112–6

[8] Rief M, Gautel M, Oesterhelt F, Fernandez J M and Gaub H E 1997 Reversible unfolding of individual titin immunoglobulin domains by AFM *Science* **276** 1109–12

[9] Carrion-Vasquez M, Oberhauser A F, Fowler S B, Marszalek P E, Broedel S E, Clarke J and Fernandez J M 1999 Mechanical and chemical unfolding of a single protein: a comparison *Proc. Natl Acad. Sci. USA* **96** 3694–9

[10] Dietz H, Berkemeier F, Bertz M and Rief M 2006 Anisotropic deformation response of single protein molecules *Proc. Natl Acad. Sci. USA* **103** 12724–8

[11] Oberhauser A F, Hansma P K, Carrion-Vazquez M and Fernandez J M 2001 Stepwise unfolding of titin under force-clamp atomic force microscopy *Proc. Natl Acad. Sci. USA* **98** 468–72

[12] Oberhauser A F, Marszalek P E, Carrion-Vazquez M and Fernandez J M 1999 Single protein misfolding events captured by atomic force microscopy *Nat. Struct. Biol.* **6** 1025–8

[13] Smith B L, Schaeffer T E, Viani M, Thompson J B, Frederic N A, Kindt J, Belcher A, Stucky G D, Morse D E and Hansma P K 1999 Molecular mechanistic origin of the toughness of natural adhesives, fibres and composites *Nature* **399** 761–3

[14] Fowler S B, Best R B, Toca Herrera J L, Rutherford T J, Steward A, Paci E, Karplus M and Clarke J 2002 Mechanical unfolding of a titin Ig domain: structure of unfolding intermediate revealed by combining AFM, molecular dynamics simulations, NMR and protein engineering *J. Mol. Biol.* **322** 841–9

[15] Lu H, Isralewitz B, Krammer A, Vogel V and Schulten K 1998 Unfolding of titin immunoglobulin domains by steered molecular dynamics simulation *Biophys. J.* **75** 662–71

[16] Carrion-Vazquez M, Li H, Lu H, Marszalek P E, Oberhauser A F and Fernandez J M 2003 The mechanical stability of ubiquitin is linkage dependent *Nat. Struct. Biol.* **10** 738–43

[17] Chyan C-L, Lin F-C, Peng H, Yuan J-M, Chang C-H, Lin S-H and Yang G 2004 Reversible mechanical unfolding of single ubiquitin molecules *Biophys. J.* **87** 3995–4006

[18] Carrion-Vazquez M, Oberhauser A F, Fisher T E, Marszalek P E, Li H and Fernandez J M 2000 Mechanical design of proteins studied by single-molecule force spectroscopy and protein engineering *Prog. Biophys. Mol. Biol.* **74** 63–91

[19] Lee G, Abdi K, Jiang Y, Michaely P, Bennett V and Marszalek P E 2006 Nanospring behavior of ankyrin repeats *Nature* **440** 246–9

[20] Li L W, Wetzel S, Pluckthun A and Fernandez J M 2006 Stepwise unfolding of ankyrin repeats in a single protein revealed by atomic force microscopy *Biophys. J.* **90** L30–2

[21] Alam M T, Yamada T, Carlsson U and Ikai A 2002 The importance of being knotted: effects of the C-terminal knot structure on enzymatic and mechanical properties of bovine carbonic anhydrase *FEBS Lett.* **519** 35–40

[22] Huang S, Ratliff K S and Matouschek A 2002 Protein unfolding by the mitochondrial membrane potential *Nat. Struct. Biol.* **9** 301–7

[23] Matouschek A 2003 Protein unfolding—and important process *in vivo*? *Curr. Opin. Struct. Biol.* **13** 98–109

[24] Prakash S and Matouschek A 2004 Protein unfolding in the cell *Trends Biochem. Sci.* **29** 593–600

[25] Deshaies R J, Sanders S L, Feldheim D A and Schekman R 1991 Assembly of yeast Sec proteins involved in translocation into the endoplasmic reticulum into a membrane-bound multisubunit complex *Nature* **349** 806–8

[26] Schatz G and Dobberstein B 1996 Common principles of protein translocation across membranes *Science* **271** 1519–26

[27] Dalbey R E and Chen M 2004 Sec-translocase mediated membrane protein biogenesis *Biochim. Biophys. Acta* **1694** 37–53

[28] Vale R D 1996 Switches, latches and amplifiers: common themes of G proteins and molecular proteins *J. Cell. Biol.* **135** 291–302

[29] Reddy A S 2001 Molecular motors and their functions in plants *Int. Rev. Cytol.* **204** 97–178

[30] Schliwa M and Woehlke G 2003 Molecular motors *Nature* **422** 759–65

[31] Bustamante C, Chemla Y R, Forde N R and Izhaky D 2004 Mechanical processes in biochemistry *Annu. Rev. Biochem.* **73** 705–48

[32] Brockwell D J, Godfrey S, Beddard S, Paci E, West Dan K, Olmsted P D, Smith D A and Radford S E 2005 Mechanically unfolding small topologically simple protein L *Biophys. J.* **89** 506–19

[33] Best R B, Li B, Steward A, Daggett V and Clarke J 2001 Can non-mechanical proteins withstand force? Stretching barnase by atomic force microscopy and molecular dynamics simulation *Biophys. J.* **81** 2344–56

[34] Brockwell D J, Paci E, Zinober R C, Beddard G, Olmsted P D, Smith D A, Perham R N and Radford S E 2003 Pulling geometry defines mechanical resistance of β-sheet protein *Nat. Struct. Biol.* **10** 731–7

[35] Bockelmann U, Essevaz-Roulet B and Heslot F 1997 Molecular stick-slip motion revealed by opening DNA with piconewton forces *Phys. Rev. Lett.* **79** 4489–92

[36] Baumann C G, Bloomfield V A, Smith S B, Bustamante C, Wang M D and Block S M 2000 Stretching of single collapsed DNA molecules *Biophys. J.* **78** 1965–78

[37] Brockwell D J, Beddard G S, Clarkson J, Zinober R C, Blake A W, Trinick J, Olmsted P D, Smith D A and Radford S E 2002 The effect of core destabilization on the mechanical resistance of I27 *Biophys. J.* **83** 458–72

[38] Carrion-Vazquez M, Marszalek P E, Oberhauser A F and Fernandez J M 1999 Atomic force microscopy captures length pheno-types in single proteins *Proc. Natl Acad. Sci. USA* **96** 11288–92

[39] Cecconi C, Shank E A, Bustamante C and Marqusee S 2005 Direct observation of the three-state folding of a single protein molecule *Science* **309** 2057–60

[40] Dietz H and Rief M 2004 Exploring the energy landscape of the GFP by single-molecule mechanical experiments *Proc. Natl Acad. Sci.* **101** 16192–7

[41] Dietz H and Rief M 2006 Protein structure by mechanical triangulation *Proc. Natl Acad. Sci.* **103** 1244–7

[42] Erickson H P 1994 Reversible unfolding of fibronectin type III and immunoglobulin domains provides the structural basis for stretch and elasticity of titin and fibronectin *Proc. Natl Acad. Sci. USA* **91** 10114–8

[43] Furuike S, Ito T and Yamazaki M 2001 Mechanical unfolding of single filamin A (ABP-280) molecules detected by atomic force microscopy *FEBS Lett.* **498** 72–5

[44] Janovjak H, Kessler M, Oesterhelt D, Gaub H and Muller D J 2003 Unfolding pathways of native bacteriorhodopsin depend on temperature *EMBO J.* **22** 5220–9

[45] Kellermayer M S Z, Grama L, Karsai A, Nagy A, Kahn A, Datki Z L and Penke B 2005 Reversible mechanical unzipping of amyloid β-fibrils *J. Biol. Chem.* **280** 8464–70

[46] Li H, Oberhauser A F, Fowler S B, Clarke J and Fernandez J M 2000 Atomic force microscopy reveals the mechanical design of a modular protein *Proc. Natl Acad. Sci. USA* **97** 6527–31

[47] Li H, Linke W A, Oberhauser A F, Carrion-Vazquez M, Kerkviliet J G, Lu H, Marszalek P E and Fernandez J M 2002 Reverse engineering of the giant muscle protein titin *Nature* **418** 998–1002

[48] Li H, Oberhauser A F, Redick S D, Carrion-Vazquez M, Erickson H P and Fernandez J M 2001 Multiple conformations of PEVK proteins detected by single-molecule techniques *Proc. Natl Acad. Sci. USA* **98** 10682–6

[49] Li L, Huang H H-L, Badilla C L and Fernandez J M 2005 Mechanical unfolding intermediates observed by single-molecule force spectroscopy in fibronectin type III module *J. Mol. Biol.* **345** 817–26

[50] Li H B and Fernandez J M 2003 Mechanical design of the first proximal Ig domain of human cardiac titin revealed by single molecule force spectroscopy *J. Mol. Biol.* **334** 75–86

[51] Lenne P F, Raae A J, Altmann S M, Saraste M and Horber J K H 2000 States and transition during unfolding of a single spectrin repeat *FEBS Lett.* **476** 124–8

[52] Law R, Carl P, Harper S, Dalhaimer P, Speicher D W and Discher D E 2003 Cooperativity in force unfolding of tandem spectrin repeats *Biophys. J.* **84** 533–44

[53] Law R, Liao G, Harper S, Yang G, Speicher D and Discher D E 2003 Pathway shifts and thermal softening in temerature-coupled forced unfolding of spectrin domains *Biophys. J.* **85** 3286–93

[54] Law R, Carl P, Harper S, Speicher D W and Discher D E 2004 Infulence of lateral association on forced unfolding of antiparallel spectrin heterodimers *J. Biol. Chem.* **279** 16410–6

[55] Liphardt J, Onoa B, Smith SB, Tinoco I and Bustamante C 2001 Reversible unfolding of single RNA molecules by mechanical force *Science* **292** 733–7

[56] Ma K, Kan L-s and Wang K 2001 Polyproline II helix is a key structural motif of the elastic PEVK segment of titin *Biochemistry* **40** 3427–38

[57] Marszalek P E, Lu H, Li H B, Carrion-Vazquez M, Oberhauser A F, Schulten K and Fernandez J M 1999 Mechanical unfolding intermediates in titin modules *Nature* **402** 100–3

[58] Muller D J, Baumeister W and Engel A 1999 Controlled un-zipping of a bacterial surface layer with atomic force microscopy *Proc. Natl Acad. Sci. USA* **96** 13170–4

[59] Mitsui K, Hara M and Ikai A 1996 Mechanical unfolding of alpha 2-macroglobulin molecules with atomic force microscope *FEBS Lett.* **385** 29–33

[60] Oberhauser A F, Marszalek P E, Erickson H P and Fernandez J M 1998 The molecular elasticity of the extracellular matrix protein tenascin *Nature* **14** 181–5

[61] Oberhauser A F, Badilla-Fernandez C, Carrion-Vazquez M and Fernandez J M 2002 The mechanical hierarchies of fibronectin observed with single-molecule AFM *J. Mol. Biol.* **31** 433–47

[62] Oesterhelt F, Oesterhelt D, Pfeiffer M, Engel A, Gaub H E and Muller D J 2000 Unfolding pathways of individual bacteriorhodopsins *Science* **288** 143–6

[63] Oberdorfer Y, Fuchs H and Janshoff A 2000 Conformational analysis of native fibronectin by means of force spectroscopy *Langmuir* **16** 9955–8

[64] Ohta S, Alam M T, Arakawa H and Ikai A 2004 Origin of mechanical strength of bovine carbonic anhydrase studied by molecular dynamics simulation *Biophys. J.* **87** 4007–20

[65] Oroudjev E, Soares J, Arcidiacono S, Thompson J B, Fossey S A and Hansma H G 2002 Segmented nanofibers of sider dragline silk: atomic force microscopy and single-molecule force spectroscopy *Proc. Natl Acad. Sci. USA* **30** 6460–5

[66] Oberdoerfer Y, Fusch H and Janshoff A 2000 Conformational analysis of native fibronectin by means of force spectorscopy *Langmuir* **16** 9955–8

[67] Rief M, Gautel M, Schemmel M and Gaub H G 1998 The mechanical stability of immunoglobulin and fibronectin III domains in the muscle protein titin measured by atomic force microscopy *Biophys. J.* **75** 3008–14

[68] Rief M, Pascual J, Saraste M and Gaub H G 1999 Single molecule force spectroscopy of spectrin repeats: low unfolding forces in helix bundles *J. Mol. Biol.* **286** 553–61

[69] Rief M, Gautel M and Gaub H E 2000 Unfolding forces of titin and fibronectin domains directly measured by AFM *Adv. Exp. Med. Biol.* **481** 129–36

[70] Rief M, Oesterhelt F, Heymann B and Gaub H E 1997 Single molecule force spectroscopy on polysaccharides by atomic force microscopy *Science* **275** 1295–7

[71] Schwaiger I, Kardinal A, Schleicher M, Noegel A A and Rief M 2004 A mechanical unfolding intermediate in an actin-crosslinking protein *Nat. Struct. Mol. Biol.* **11** 81–5

[72] Schlierf M and Rief M 2005 Temperature softening of a protein in single-molecule experiments *J. Mol. Biol.* **354** 497–503

[73] Schlierf M and Rief M 2006 Single-molecule unfolding force distribution reveal a funnel-shape energy landscape *Biophys. J.* **90** L33–5

[74] Sarkar A, Caamano S and Fernandez J M 2005 The elasticity of individual titin PEVK exons measured by single molecule atomic force microscopy *J. Biol. Chem.* **280** 6261–4

[75] Tskhovrebova L and Trinick J 2004 Properties of titin immunoglobulin and fibronectin-3 domains *J. Biol. Chem.* **279** 46351–4

[76] Watanabe K, Muhle-Goll C, Kellermayer M S Z, Labeit S and Granzier H L 2002 Different molecular mechanics displayed by titin's constitutively and differentially expressed tandem Ig segments *Struct. Biol. J.* **137** 248–58

[77] Williams P M, Fowler S B, Best R B, Toca-Herrera J L, Scott K A, Steward A and Clarke J 2003 Hidden complexity in the mechanical properties of titin *Nature* **422** 446–9

[78] Watanabe K, Nair P, Labeit D, Kellermayer M S Z, Greaser M, Labeit S and Granzier H L 2002 Molecular mechanics of cardiac titins PEVK and N2B spring elements *J. Biol. Chem.* **277** 11549–58

[79] Yang G, Cecconi C, Baase W A, Vetter I R, Breyer W A, Haack J A, Matthews B W, Dahlquist F W and Bustamante C 2000 Solid-state synthesis and mechanical unfolding of polymers of T4 lysozyme *Proc. Natl Acad. Sci. USA* **97** 139–44

[80] Craig D, Krammer A, Schulten K and Vogel V 2001 Comparison of the early stages of forced unfolding for fibronectin type III modules *Proc. Natl Acad. Sci. USA* **98** 5590–5

[81] Gao M, Craig D, Lequin O, Campbell I D, Vogel V and Schulten K 2003 Structure and functional significance of mechanically unfolded fibronectin type III1 intermediates *Proc. Natl Acad. Sci. USA* **100** 14784–9

[82] Gao M, Wilmanns M and Schulten A K 2002 Steered molecular dynamics studies of titin I1 domain unfolding *Biophys. J.* **83** 3435–45

[83] Gao M, Craig D, Vogel V and Schulten K 2002 Identifying unfolding intermediates of FN-III$_{10}$ by steered molecular dynamics *J. Mol. Biol.* **323** 939–50

[84] Krammer A, Lu H, Isralewitz B, Schulten K and Vogel V 1999 Forced unfolding of the fibronectin type III module reveals a tensile molecular recognition switch *Proc. Natl Acad. Sci. USA* **96** 1351–6

[85] Klimov D K and Thirumalai D 2000 Native topology determines force-induced unfolding pathways in globular proteins *Proc. Natl Acad. Sci. USA* **97** 7254–9

[86] Lu H and Schulten K 1999 Streed molecular dynamics simulations of conformational changes of immunoglobulin domain I27 interpret atomic force microscopy observation *Chem. Phys.* **247** 141–53

[87] Lu H and Schulten K 2000 The key event in force-induced unfolding of titin's immunoglobulin domains *Biophys. J.* **79** 51–65

[88] Lu H and Schulten K 1999 Steered molecular dynamics simulation of force-induced protein domain unfolding *Proteins* **35** 453–63

[89] Pabon G and Amzel L M 2006 Mechanism of titin unfolding by force: insight from quasi-equilibrium molecular dynamics calculations *Biophys. J.* **91** 467–72

[90] Li P-C and Makarov D E 2004 Simulation of the mechanical unfolding of ubiquitin: probing different unfolding reaction coordinates by changing the pulling geometry *J. Chem. Phys.* **121** 4826–32

[91] Makarov D E, Hansma P K and Metiu H 2001 Kinetic Monte Carlo simulation of titin unfolding *J. Chem. Phys.* **114** 9663–73

[92] Paci E and Karplus M 2000 Unfolding proteins by external forces and temperature: the importance of topology and energetics *Proc. Natl Acad. Sci. USA* **97** 6521–6

[93] Paci E and Karplus M 1999 Forced unfolding of fibronectin type 3 modules: an analysis by biased molecular dynamics simulations *J. Mol. Biol.* **288** 441–59

[94] Abe H and Go N 1981 Noninteracting local-structure model of folding and unfolding transition in globular proteins. II. Application to two-dimensional lattice proteins *Biopolymers* **20** 1013–31

[95] Takada S 1999 Go-ing for the prediction of protein folding mechanism *Proc. Natl Acad. Sci. USA* **96** 11698–700

[96] Veitshans T, Klimov D and Thirumalai D 1997 Protein folding kinetics: timescales, pathways and energy landscapes in terms of sequence-dependent properties *Fold. Des.* **2** 1–22

[97] Hoang T X and Cieplak M 2000 Molecular dynamics of folding of secondary structures in Go-like models of proteins *J. Chem. Phys.* **112** 6851–62

[98] Karanicolas J and Brooks C L III 2002 The origins of asymmetry in the folding transition states of protein L and protein G *Protein Sci.* **11** 2351–61

[99] Berman H M, Westbrook J, Feng Z, Gilliland G, Bhat T N, Weissig H, Shindyalov I N and Bourne P E 2000 The Protein Data Bank *Nucleic Acids Res.* **28** 235–42

[100] Cieplak M and Hoang T X 2000 Scaling of folding properties in Go models of proteins *J. Biol. Phys.* **26** 273–94

[101] Chang I, Cieplak M, Dima R I, Maritan A and Banavar J R 2001 Protein threading by learning *Proc. Natl Acad. Sci. USA* **98** 14351–5

[102] Orengo C A, Micchie A D, Jones S, Jones D T, Swindels M B and Thronton J M 1997 CATH—a hierarchical classification of protein domain structures *Structure* **5** 1093–108

[103] Gront D and Kolinski A 2006 BioShell—a package of tools for structural biology computations *Bioinformatics* **22** 621–2

[104] Hoang T X and Cieplak M 2001 Sequencing of folding events in Go-like proteins *J. Chem. Phys.* **113** 8319–28

[105] Cieplak M and Hoang T X 2001 Kinetic non-optimality and vibrational stability of proteins *Proteins Struct. Funct. Genet.* **44** 20–5

[106] Cieplak M and Hoang T X 2003 Universality classes in folding times of proteins *Biophys. J.* **84** 475–88

[107] Cieplak M, Hoang T X and Robbins M O 2004 Thermal effects in stretching of Go-like models of titin and secondary structures *Proteins Struct. Funct. Biol.* **56** 285–97

[108] Tsai J, Taylor R, Chothia C and Gerstein M 1999 The packing density in proteins: Standard radii and volumes *J. Mol. Biol.* **290** 253–66

[109] Cieplak M, Pastore A and Hoang T X 2005 Mechanical properties of the domains of titin in a Go-like model *J. Chem. Phys.* **122** 054906

[110] Cieplak M and Marszalek P E 2005 Mechanical unfolding of ubiquitin molecules *J. Chem. Phys.* **123** 194903

[111] Hyeon C B and Thirumalai D 2003 Can energy landscape roughness of proteins and RNA be measured by using mechanical unfolding experiments? *Proc. Natl Acad. Sci. USA* **100** 10249–53

[112] Cieplak M, Hoang T X and Robbins M O 2004 Stretching of proteins in the entropic limit *Phys. Rev.* E **69** 011912

[113] Kwiecinska J I and Cieplak M 2005 Chirality and protein folding *J. Phys.: Condens. Matter* **17** S1565–80

[114] Settanni G, Hoang T X, Micheletti C and Maritan A 2002 Folding pathways of prion and doppel *Biophys. J.* **83** 3533–41

[115] Cieplak M and Hoang T X 2003 Folding of proteins in Go modles with angular interactions *Physica* A **330** 195–205

[116] Clementi C, Nymeyer H and Onuchic J N 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

[117] Klimov D K and Thirumalai D 1997 Viscosity dependence of the folding rates of proteins *Phys. Rev. Lett.* **79** 317–20

[118] de Gennes P G 1979 *Polymer Concepts in Polymer Physics* (New York: Cornell University Press)

[119] Szymczak P and Cieplak M 2006 Stretching of proteins in a uniform flow *J. Chem. Phys.* **125** 164903

[120] Gear W C 1971 *Numerical Initial Value Problems in Ordinary Differential Equations* (Englewood Cliffs, NJ: Prentice-Hall)

[121] Szymczak P and Cieplak M 2006 Stretching of proteins in a force-clamp *J. Phys.: Condens. Matter* **18** L21–8

[122] Cieplak M, Filipek S, Janavjak H and Krzysko K A 2006 Pulling single bacteriorhodopsin out of a membrane: Comparison of simulation and experiment *Biochim. Biophys. Acta* **1758** 537–44

[123] Berman H M, Goodsell D S and Bourne P E 2002 Protein structures: from famine to feast *Am. Sci.* **90** 350–9

[124] Cieplak M 2005 Mechanical stretching of proteins: calmodulin and titin *Physica* A **352** 28–42

[125] Ye Y Z and Godzik A 2004 FATCAT: a web server for flexible structure comparison and structure similarity searching *Nucleic Acids Res.* **32** (Suppl. 2) W582–5

[126] West D K, Brockwell D J, Olmsted P D, Radford S E and Paci E 2006 Mechanical resistance of proteins explained using simple molecular models *Biophys. J.* **90** 287–97

[127] Covell D G and Jernigan R L 1990 Conformation of folded proteins in restricted spaces *Biochemistry* **29** 3287–94

[128] Micheletti C, Seno F, Banavar J R and Maritan A 2001 Learning effective amino acid interactions through iterative stochastic techniques *Proteins Struct. Funct. Genet.* **42** 422–31

[129] Carl P, Kwok C H, Manderson G, Speicher D W and Discher D 2001 Force unfolding modulated by disulphide bonds in the Ig domains of a cell adhesion molecule *Proc. Natl Acad. Sci. USA* **98** 1565–70

[130] Bhasin N, Carl P, Harper S, Feng G, Lu H, Speicher D W and Discher D E 2004 Chemistry on a single protein, vascular cell adhesion molecule-1, during forced unfolding *J. Biol. Chem.* **279** 45865–74