Role of Native-State Topology in the Stabilization of Intracellular Antibodies

Giovanni Settanni,*† Antonino Cattaneo,*† and Amos Maritan*†‡

*International School for Advanced Studies, 34014 Trieste; †Istituto Nazionale Fisica della Materia, 34014 Trieste; ‡International Center for Theoretical Physics, 34100 Trieste, Italy

ABSTRACT The role played by the geometric position of each amino acid in the folding process of the immunoglobulin (Ig) variable domain is identified and measured through molecular dynamics simulations of models based on the topology of its native state. This measure allows identifying the parts of the protein that, for geometrical reasons, when mutated, would result in relevant protein stability changes. Simulations were performed without considering the covalent disulfide bond present in most of the Ig domains. The results are in good agreement with site-directed mutagenesis experiments on the folding of intracellular antibodies in which the disulfide bond does not form. We also found agreement with data on amino acid conservation in the Ig variable domain sequences. This indicates a new way for a rational approach to the design of intracellular antibodies more resistant to the suppression of the disulfide bond that occurs in the cytoplasm.

INTRODUCTION

Antibodies are secreted by plasma cells and have evolved to act in a variety of compartments of the mammalian body outside of cells. The demands on stability have kept a selection pressure on immunoglobulin domains to retain disulfide bonds in all germ-line immunoglobulin genes. In fact, all antibody variable domains, the antigen-binding domains, contain one conserved disulfide bond linking two pairs of conserved cysteine residues. The disulfide bonds form during the process of secretion from the cell in the endoplasmic reticulum. Recent advances in the field of recombinant antibodies have allowed engineering a wide variety of recombinant forms of antibody domains, e.g., the single-chain Fv (scFv) fragments (Bird et al., 1988; Huston et al., 1988), which consist only of the variable domains connected by a linker. These simpler antibody domains can be expressed in a variety of hosts, ranging from bacteria to mammalian cells, but still exploiting disulfide formation in the host secretory pathway.

However, it has been shown that, by suitable engineering, it is also possible to redirect the expression of antibodies or antibody domains away from the secretory pathway, to other intracellular compartments (Biocca et al., 1990, 1995; Richardson and Marasco, 1995). This exploits the use of intracellular targeting signals that normally dictate the intracellular fate and traffic of proteins. Antibody domains, equipped with targeting sequences borrowed from other proteins, have indeed been successfully targeted to novel intracellular compartments (Biocca et al., 1995), where antibodies are not normally found, such as the cytoplasm, the nucleus, and the mitochondria. Because from the initial studies it was clear that if a sufficient functional expression, i.e., a correct folding, of such intracellular antibodies (intrabodies) could be achieved, this would enable them to bind to their target protein in the appropriate intracellular compartment and evoke specific biological effects. In fact, intracellular antibodies have been shown to act as protein knock-out reagents by inactivating the recognized protein. Thus, proteins involved in signal transduction (p21 ras (Biocca et al., 1993;Cardinale et al., 1998)), plant viral pathogenesis (Tavladoraki et al., 1993), and human virus replication (Duan et al., 1994; Mhashilkar et al., 1995; Gargano and Cattaneo, 1997) have been successfully blocked by this emerging technology. This has an immediate application for research, where tools to inactivate proteins are fundamental in the process of elucidating protein function. In perspective, the intrabody technology has been proposed to have broad therapeutic applications, possibly in a gene therapy setting (Marasco, 1995).

More recently, the intrabody technology has been proposed as a potential tool of choice in functional genomics and functional proteomics (Cattaneo and Biocca, 1997). The completion of the sequencing of the human genome calls for methods to assess the function of genes, coming from genome sequencing projects or from proteomic screens. Protein knock-out by the intrabody technology promises to become an essential tool in the endeavor to facilitate the discovery of gene and protein function, provided its present bottlenecks are solved.

As discussed above, in principle, antibody domains can be directed to all intracellular compartments by encoding the corresponding targeting sequence attached to that encoding for the antibody. Among the different intracellular locations, the cytoplasm is a cross-road, because although proteins targeted to the secretory pathway are co-translationally targeted to the endoplasmic reticulum, proteins targeted to other intracellular compartments are first synthesized in the cytoplasm, as are also the proteins resident of the cytoplasm itself. For antibodies, expression in the cytoplasm is the most difficult task, because of its reducing

Received for publication 19 March 2001 and in final form 20 June 2001.

Address reprint requests to Dr. Giovanni Settanni, SISSA, via Beirut 2, 34014 Trieste, Italy. Tel.: 39-040-2240-460; Fax: 39-040-3787-485; E-mail: settanni@sissa.it.

0006-3495/01/11/2935/11 $2.00

environment (Gilbert, 1990; Hwang et al., 1992). This reducing potential prevents the formation of disulfide bonds, including the conserved intradomain disulfides in antibody domains (Johnson and Wu, 2000; Williams and Barclay, 1988). Indeed, it was demonstrated that scFv fragments expressed in the cytoplasm do not form the disulfide bonds (Biocca et al., 1995; Martineau et al., 1998; Visintin et al., 1999). The intradomain disulfide contributes 4–5 kCal/mol to the stability of antibody domains (Goto and Hamaguchi, 1979; Frisch et al., 1994). Therefore, antibody fragments expressed in a reducing environment are strongly destabilized, compared with the same molecules containing disulfides, and a smaller fraction of these antibody domains is likely to fold to the correct native structure. This fact is believed to be responsible for the fact that not all antibodies perform equally well, when expressed in the cell cytoplasm (Cattaneo and Biocca, 1999). Indeed, although a number of cytoplasmically expressed antibody fragments were reported to show specific and well controlled biological effects (see Cattaneo and Biocca, 1997, for a review of representative examples), the average scFv fragment isolated from the corresponding monoclonal antibody or from a phage library will fold inefficiently in the cell cytoplasm. Therefore, for many applications, an antibody of interest may require further ad hoc optimization by protein engineering to make it more stable (Jung and Pluckthun, 1997).

Moreover, if intrabody technology is to hold its promises as the technology of choice for high-throughput functional genomics, a reliable and/or a priori predictable access to these molecules is needed.

Inspired by the study of a small number of naturally occurring antibodies that, due to somatic mutations, lack the intradomain disulfide bond and yet appear to fold correctly (Rudikoff and Pumphrey, 1986), tolerating the absence of this stabilizing bond, it is now believed that the overall stability of antibody domains is contributed in a distributed way by many residues, with the disulfide bond being one of the different concurring stabilizing factors (Proba et al., 1997). Thus, the overall folding stability of different antibody domains covers a wide range, and those antibodies naturally falling near the upper stability range will tolerate the absence of the disulfide bond, whereas those near to the lower stability limit will not. Consistent with this prediction, it was found that antibodies of the latter class could be engineered to tolerate the absence of the disulfide bond by mutations elsewhere in the framework region (Proba et al., 1998; Worn and Pluckthun, 1998, 1999). According to this view, it should be possible to design suitable selection procedures to isolate from natural repertoires of antibodies those that are able to bind antigen under conditions of intracellular expression. Such a selection procedure has been recently described and shown to allow the isolation of functional intrabodies (intrabody trap technology, or ITT) (Visintin et al., 1999; Visintin, 2000).

The alternative approach is to investigate the limiting factors for antibody folding stability. An understanding of the principles underlying, and the interplay between, antibody stability, affinity, and their performance as cytoplasmically expressed intrabodies would greatly facilitate the rational engineering of superior antibody frameworks as hosts for complementarity determining region (CDR) grafting (Jung and Pluckthun, 1997) or the rational improvement of individual antibodies (Nieba et al., 1997).

As a first step in this direction, we have exploited native-state topological constraints on the folding of the immunoglobulin domain to understand the relative importance of the positions of the amino acids for the stability of the protein. We have applied a refinement of a recently introduced technique (Micheletti et al., 1999; Cecconi et al., 2001; Maritan, 2001; Settanni, 2000) that pinpoints the important events during the folding process of a protein (Micheletti et al., 1999) using the geometrical features of its native state (Maritan et al., 2000a,b; Cecconi et al., 2001; Banavar and Maritan, 2001; Alm and Baker, 1999; Munoz et al., 1998; Galzitskaya and Finkelstein, 1999; Clementi et al., 2000; Martinez et al., 1998; Chiti et al., 1999). The model we will introduce in the next section is based on the following observations. There are, in principle, a huge number of sequences that can be formed using a repertoire of 20 amino acids even for relatively short proteins. Nevertheless, a negligible percentage of this astronomically large number of total possibilities has been observed. It is believed that a similar situation occurs in structure space too because there are likely to be only several thousand distinct folds. Another well-known fact is that several sequences may fold into similar native state structures. These facts suggest that the topology of the native state of a protein should determine some of the main properties of protein folding and stability. Once a viable native state structure is chosen, the sequence selection mechanism essentially satisfies the principle of minimal frustration (Wolynes et al., 1995; Go, 1983); i.e., sequences are designed in such a way to be maximally compatible and able to fold into the native state. Because of the rich repertory of amino acids, this can be done in many ways, and recent experimental studies have shown that carefully designed sequences can indeed be more stable than the wild type (Martinez et al., 1998). Any of the viable protein-like sequences is then characterized not only by thermodynamic stability in the native state but also by an efficient folding toward the correct native state.

For a given native structure and for well designed (or evolved) sequences, a caricature of the folding process, which is essentially sequence independent and is controlled by the native-state structure topology, is provided by considering a Go-like model (Go, 1983) where attractive interactions exist only between the native-like contacts. Such a model is indeed minimally frustrated (Wolynes et al., 1995) and exhibits many of the characteristics that one might expect of an ideally designed sequence on a given native-

state structure. It is important to recognize that such a model can only address questions pertaining to the role of the native-state topology in the folding process or in the thermodynamics stability of the protein.

Immunoglobulin Fv domains represent an excellent testing table for the hypothesis mentioned above and can greatly help in clarifying the range and limits of applicability of a description based on the topological assumption. They present a high degree of sequence diversity and high structure conservation so that in principle we can assume that the detailed amino acid chemistry has been incorporated in the native-state topology through an averaging process. Moreover, the importance of the folding stability of this particular protein family in the reducing environment of the cytoplasm, the role of the disulfide bond, and the behavior of the domain after its removal represent aspects highly worthy of study. For this reason, we performed our calculations without considering the disulfide bridge as a covalent bond. As a comparison, we have also performed calculations on titin, whose native-state topology is similar to the one of Ig variable domain.

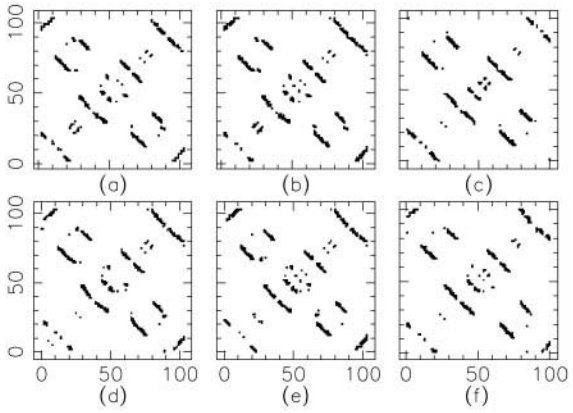

FIGURE 1 Contact maps of some immunoglobulin variable domains from the heavy chain and from the light chain after being structurally aligned to our model map (*a*). The PDB IDs of the structures from which maps have been extracted are: 2fb4 VL (b), VH (c); 1Fig VL (d), VH (f); 1mfc VL (e). See Methods for the details about the structural alignments. The high structural homology is evident.

METHODS

The model

To elucidate the importance of the topological features of the Ig variable domain we make use of a simplistic representation of the three-dimensional structure of the protein. We extracted from the crystallographic structure of an immunoglobulin variable domain (deposited at Protein Data Bank (PDB) as 1mco) the 109 C_α atom positions, which will be used to represent the amino acid positions in our simplified model. We then determined the contacts between amino acids by simply checking the relative distance of the C_α. If the distance was less than a cutoff of 6.5 Å (usually this can be set from 6.5 and 8.0 (Vendruscolo et al., 2000)), then the amino acids were considered in contact, but otherwise as no contact. For the reasons explained above the disulfide bridge is treated at the same extent of the other contacts.

The energy function for our system is a standard in biopolymer molecular dynamics (MD) simulations (Pearlman et al., 1995; Brooks et al., 1983). It is composed of a bonded and a nonbonded term and written as:

$$E = V_B + V_{NB}, \tag{1}$$

where V_B represents a potential between close-in-sequence C_α and is supposed to mimic the peptide bond and the geometrical constraints of close-in-sequence amino acids (i.e., Ramachandran angle propensities).

Its detailed form is:

$$\begin{aligned} V_B &= gV_p + hV_a + kV_d \\ V_p &= \sigma \sum_{i=1}^{N-1} (r_{i,i+1} - r^{(n)}_{i,i+1})^2 \\ V_a &= \sigma \sum_{i=1}^{N-2} (\theta_{i,i+1,i+2} - \theta^{(n)}_{i,i+1,i+2})^2 \\ V_d &= \sigma \sum_{i=1}^{N-3} 1 - \cos(\tau_{i,i+1,i+2,i+3} - \tau^{(n)}_{i,i+1,i+2,i+3}), \end{aligned} \tag{2}$$

where $r_{i,j}$ is the distance between residues i and j, $\theta_{i,j,k}$ is the angle with the jth amino acid as vertex and the ith and the kth as edges and $\tau_{i,j,k,l}$ is the dihedral generated by the ith, the jth, the kth, and the lth amino acid. The n as superscript means the native-state value. σ represents a suitable scale factor to fix the temperature scale. The value of σ was numerically set to 10. As formulas show, the minimum of the potential energy is set in the native-state geometry, independently of the coupling constants g, h, and k. We tested different sets of coupling regimes to have a robust estimate of our measures.

The nonbonded term represents, in a very simplified way, the interactions that take place between far-in-sequence residues and that keep the protein in its native state(Go, 1983). The interactions are present for each pair of amino acids (but the sequence neighbors) and are accounted for by a van der Waals-like potential of the type:

$$V_{NB} = \sigma \sum_{i<j-3}^{N} V_{i,j} = \sigma \sum_{i<j-3}^{N} \left(5\left(\frac{r^{(n)}_{i,j}}{r_{i,j}}\right)^{12} - 6\left(\frac{r^{(n)}_{i,j}}{r_{i,j}}\right)^{10} \Delta_{i,j} \right), \tag{3}$$

where $\Delta_{i,j}$, the contact map, is 1 or 0 if the ith and jth C_α are in contact or not in the native state, respectively. This means that every pair of contacting C_α in the native state interacts through an attractive potential whereas non-contacting C_α feel just hard-core repulsion. Low values of V_{NB} correspond to the protein in the folded conformation whereas high values correspond to the unfolded state.

The choice of a particular Ig domain is irrelevant because, with the definition of contacts here adopted, many different domains both from VL and from VH region present essentially the same contact map $\Delta_{i,j}$, as shown in Fig. 1, apart from insertions and deletions in loop regions.

We analyzed three coupling regimes: in the first one, which will be denoted by A, we set $g = 50$, $h = 0$, and $k = 0$, focusing on non-local interactions only; in the second regime (B) we set $g = 50$, $h = 9$, and $k = 0$, partially turning on local (angular) interactions; in the third case (C) we set $g = 50$, $h = 9$, and $k = 9$, to have structure-determining local interactions (that was not the case in the previous regime). The ratio between coupling parameters has been chosen to roughly match vibration frequencies of the corresponding terms in all-atom MD simulations (data not shown).

The masses of the amino acids have been set to 100 arbitrary units.

Heating and cooling simulations

We carried out MD runs with the leapfrog algorithm(Allen and Tildesley, 1989) and a time-step of 0.005. The temperatures during the runs were kept constant using the Berendsen algorithm (Allen and Tildesley, 1989) with a coupling constant set to 0.01. The motion of the center of mass was removed every 1000 steps. The temperature of each successive run was increased by a fixed amount. We scanned the whole range of temperatures: from a temperature where the native state is fully stable to temperatures well above the complete unfolding transition temperature. For A and B, kT was increased from 0.01σ to 0.72σ with steps of 0.01σ. For C, kT was increased from 0.7σ to 1.06σ with steps of the same size as above. A cooling process was also carried out using a reverse procedure and was used to assess the reliability of the configurational sampling. Each run consists of 2×10^6 time steps that allowed for a good exploration of the whole phase space, as will be illustrated below.

Analysis of trajectories

The Swendsen-Ferrenberg (Ferrenberg and Swendsen, 1989) method has been applied to van der Waals energy data (V_{NB}) from simulations. The 200,000 equilibration steps at the beginning of each run have been discarded. The snapshots for the analysis are collected every 500 steps to allow proper decorrelation. That method helps in interpolating data from runs at different temperatures and in building a temperature-independent density of states. With that, we can, in principle, access all thermodynamic quantities, such as the mean potential energy, specific heat, etc. In particular, we used the peaks in the specific heat to pinpoint the folding transition temperatures.

A more detailed analysis has been carried out at single-contact level, starting from data collected in the previous step. We measured the temperature dependence of the energy associated with each contact near transitions, through the single-contact specific heat:

$$\frac{d\langle V_{ij}\rangle}{dT} = (\langle V_{ij}V_{NB}\rangle - \langle V_{ij}\rangle\langle V_{NB}\rangle)/kT^2 \tag{4}$$

This quantity also measures the energy change due to a mutation in the strength of a contact.

Indeed if V_{NB} is modified multiplying each V_{ij} by ϵ_{ij}, then:

$$\left.\frac{d\langle V_{NB}\rangle}{d\varepsilon_{hk}}\right|_{\varepsilon_{ij}=1} = \left.\frac{d\left\langle \sigma\sum_{i<j-2}\varepsilon_{ij}V_{ij}\right\rangle}{d\varepsilon_{hk}}\right|_{\varepsilon_{ij}=1} = \sigma(-(\langle V_{hk}V_{NB}\rangle - \langle V_{hk}\rangle\langle V_{NB}\rangle)/(kT) + \langle V_{hk}\rangle) = \sigma\left(-T\frac{d\langle V_{hk}\rangle}{dT} + \langle V_{hk}\rangle\right), \tag{5}$$

where ϵ_{ij} are dummy coefficients that represent the change of strength in the contact between amino acids i and j and are eventually set to 1 to get the original energy function. The variation of the average total potential energy of the protein model is directly related to (minus) the specific heat contribution of the modified contact. This means that, as far as the modifications are not very large, the stabilization of a contact with high specific heat would increase the stability of the model more than the same modification on a contact with small specific heat. The formula above, thus, represents the susceptibility of the system to a change in the strength of a contact. It can be directly related to experimental data on the folding capabilities of mutated protein sequences.

A measure of the role played by a given amino acid at ith position along the chain during the (un)folding is obtained by:

$$c_i = \sum_j \frac{d\langle V_{ij}\rangle}{dT}\Delta_{ij}, \tag{6}$$

which can be considered as the contribution to the specific heat related to the ith amino acid position.

The single-residue specific heats (SRSHs) have been used to identify folding and unfolding phases and to pinpoint residues that play a relevant role in the process.

Analysis of antibody sequences

Aligned VL and VH immunoglobulin sequences have been obtained from the Kabat database (Johnson and Wu, 2000).

We collected data about the frequency to find a specific amino acid in a specific position of the aligned sequences. We estimated the degree of variability of the amino acids for each position of the alignment as the site-specific entropy:

$$S_i = \sum_{j=1}^{N} p(i, A_{ij})\log(p(i, A_{ij})), \tag{7}$$

where i is the site index (that runs from the N-terminus to the C-terminus of the Ig domain), N is the total number of sequences in the database, A_{ij} is the amino acid type in the ith position of the jth sequence, and $p(i, A)$ is the frequency to find the amino acid A at ith position of the sequences in the database.

Structural alignment

We developed a simple algorithm to find an alignment between the structure of a VL domain and a VH domain that mostly takes into account their topology. The algorithm consists of a Monte Carlo exploration of the possible alignments in which the quantity to be maximized is the overlap between the two aligned contact maps (i.e., that take into account insertions and deletions). It is defined as:

$$\mathbb{O} = \sum_{i<j-2}\Delta^{(1)}_{i,j}\Delta^{(2)}_{i,j}, \tag{8}$$

where $\Delta^{(x)}_{i,j}$ are the aligned contact maps of the two proteins.

Comparison with titin

SRSHs have been computed for titin (the Ig-like domain PDB 1tit) (Improta et al., 1996) following the same procedure adopted for the Ig variable domain (model B has been used for the MD). The results have been compared after alignment of titin with Ig variable domain using the same algorithm described above.

RESULTS

We analyze the topological features of the immunoglobulin variable domain through MD simulations of topology-based coarse-grained models of the protein. The simulations are performed in the absence of the disulfide bond. The results demonstrate that the topology of the native state of the proteins contains information about relevant events of the folding pathways and their bottlenecks. The role of the geometrical position of each amino acid in the folding

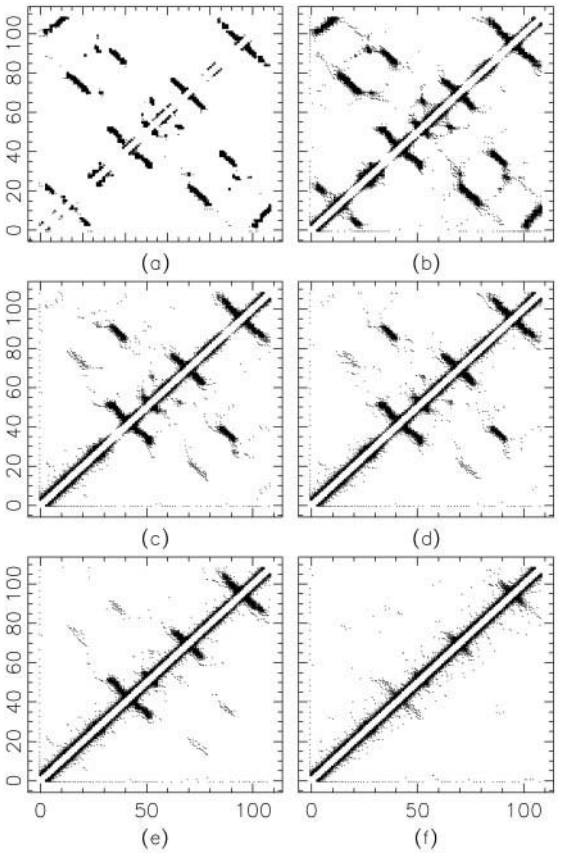

FIGURE 2 Time-averaged contact maps of the protein, within regime A. The averages are computed at $kT = 0.01\sigma$ (*a*), $kT = 0.33\sigma$ (*b*), $kT = 0.42\sigma$ (*c*), $kT = 0.46\sigma$ (*d*), $kT = 0.52\sigma$ (*e*), and $kT = 0.72\sigma$ (*f*). The partial unfolding of the peptide is evidenced by the residual structures present at temperatures in between the two folding temperatures (*c* and *d*) as obtained by the two peaks of the specific heat for regime A (see Fig. 3). In regimes B and C average contact maps were as in *a* below folding transition temperature and as in *f* above it.

process is evaluated by monitoring its contribution to the specific heat at the folding transition. This measure is, thus, basically derived from the average behavior of the system during several folding and unfolding events. Then we relate topologically relevant amino acid positions with the possible stabilizing mutations.

General temperature-dependent behavior of the system

The model system has shown different heating behaviors (Figs. 2 and 3) according to the coupling regime governing the bonded interactions (but nevertheless the role of different positions along the chain doesn't depend on that).

In the case in which only the elastic bonds were present and angle and dihedral terms were set to 0 (A), the system has experienced a multiple-phase folding process. The specific heat shows two evident peaks, as is shown in Fig. 3, revealing two main events in the process. Mean contact

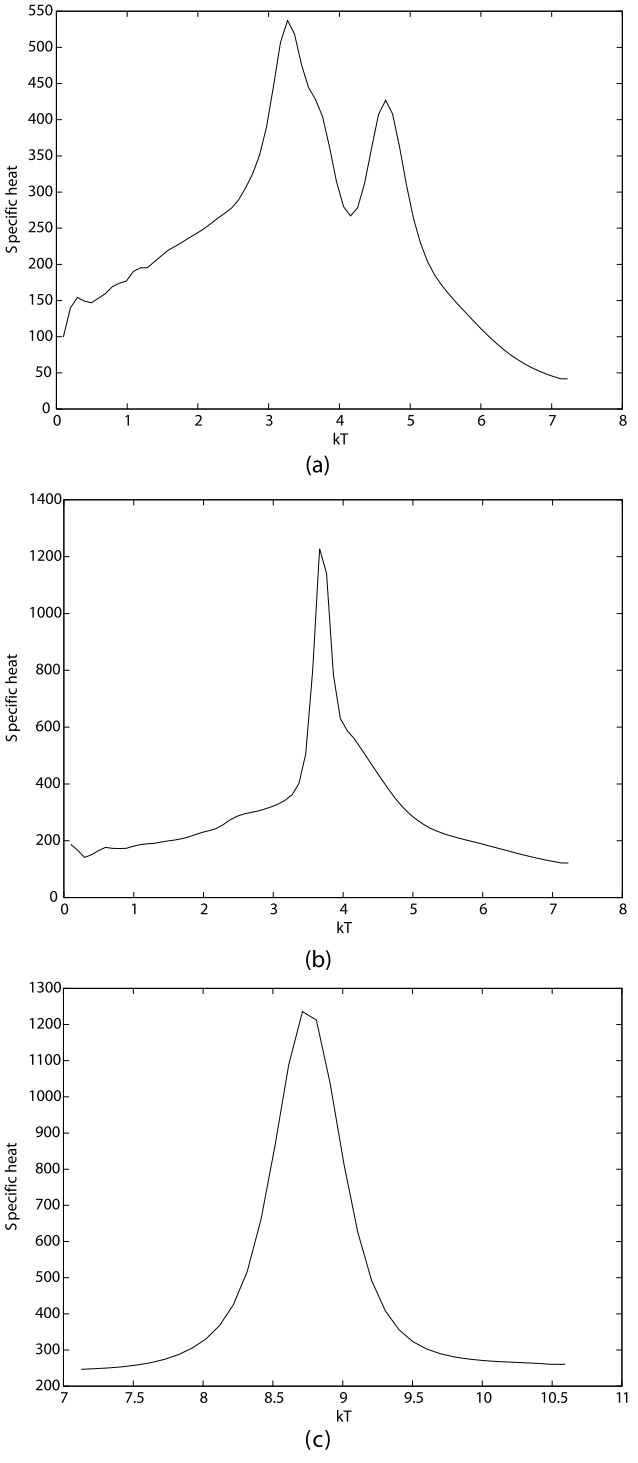

FIGURE 3 Specific heat of the protein computed with the Swendsen-Ferrenberg method for regime A (*a*), regime B (*b*), and regime C (*c*). The peaks indicate folding transitions. The cooperativity of the folding transition is much higher in *b* and *c* than in *a* where residual structure is present at temperatures between the two peaks (see Fig. 2).

maps, drawn at different temperatures (Fig. 2), clearly represent a partially structured configuration in between the completely folded and completely unfolded configurations.

In the partially structured configurations, three of the seven native β-strands are broken apart. The contacts that break apart in correspondence with the low-temperature peak are between strands 3–8 and 20–25 and between 3–8 and 98–109 (see the contact maps in Fig. 2). Strands 15–25 detach from 68–80, gradually increasing the temperature from the low-temperature peak to the high-temperature peak. The strands between 85–93 and 35–40 break apart in correspondence with the high-temperature peak.

The coupling regime with elastic bonds and angle interactions (B) shows a sharper behavior. Only a single peak is present in the specific heat, indicating that the transition from the folded to the unfolded conformation occurs in a single step, even though a smooth tail makes its appearance. In that case, all the strands of the domain break apart at the same temperature and only folded or unfolded conformations are seen (Fig. 2, *a* and *f*). The extent to which the residues cooperatively contribute to the transition (Freire et al., 1992) is, thus, higher in this case with respect to the previous one, as the transition occurs in a single step instead of two.

Finally, when dihedral interactions are also turned on (C), the behavior is even sharper. The specific heat shows again a single peak but now with no tail, cooperativity being maximal. The transition temperature is much higher than in the previous cases, because the angular and torsional energy terms together make the system more stable. Indeed, they are sufficient to enforce the native topology even without the van der Waals term.

The energy density of states computed during a heating simulation is almost identical to that computed during a cooling run. This indicates that the trajectories we considered are long enough for a good sampling of the configurational space of our model.

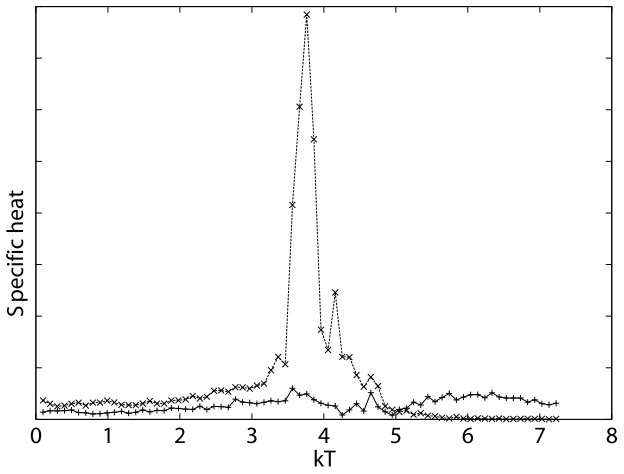

FIGURE 4 Typical temperature-dependent contributions to total specific heat of a relevant amino acid (— — —) and a scarcely relevant one (——) within regime B. The relevant amino acid is strongly involved in the folding transition as showed by the height of the peak in correspondence of the total specific heat peak (Fig. 3). For regime A and C the behavior is practically the same.

Behavior of residues and contacts at transition

We measured the specific heat of every native contact and the SRSH (c_i, see Methods) for every amino acid position. Two typical temperature-dependent behaviors of SRSH are presented in Fig. 4. As shown by the definition (Eq. 4), the contact specific heat also measures the correlation between the formation of the contact (represented by the associated van der Waals energy V_{ij}) and the folding of the whole protein (represented by the total van der Waals energy V_{NB}). The formation of contacts linking residues with high SRSH peak takes place, on an average over several folding and unfolding processes, in correspondence with the folding of the whole protein. Thus, we expect that the more the residue is involved in the folding transition the higher is the peak of the SRSH (Fig. 4). We performed the measurement for the trajectories from the different coupling regimes. In Fig. 5 we show the profiles of the SRSH peaks for every amino acid position as computed in the three cases. The figure tells us that their mutual correlation is quite high, 0.87 on average. Another very simple measure of the topological importance of an amino acid is given by its degree of burial defined as the number of contacts that the residue forms in the native-state conformation or, alternatively, its exposed surface area. However, this measure does not contain important topological information such as chain connectivity or contact order (the maximal distance along the sequence of contacting amino acids). Indeed, the correlation of our results with the degree of burial (i.e., the number of native contacts of each amino acid) is lower than the mutual correlation shown above (0.79 on average). If the degree of

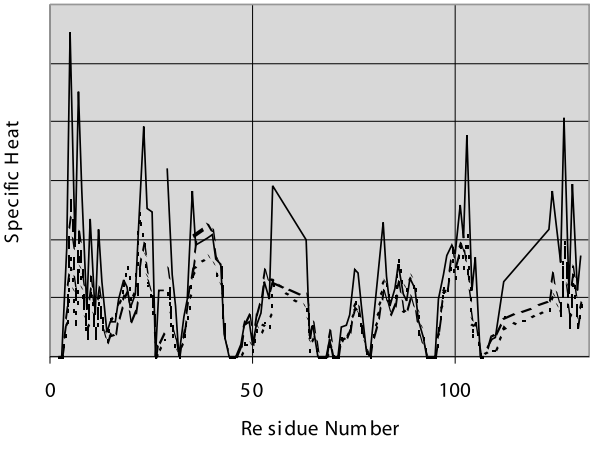

FIGURE 5 Single-residue contributions to specific heat peak for each amino acid (c_i). Regime A (——), regime B (— — —), and regime C (· · ·). The high correlation in the profiles (0.87) supports the idea that topology acts similarly on the folding transition notwithstanding the way it is enforced in the model.

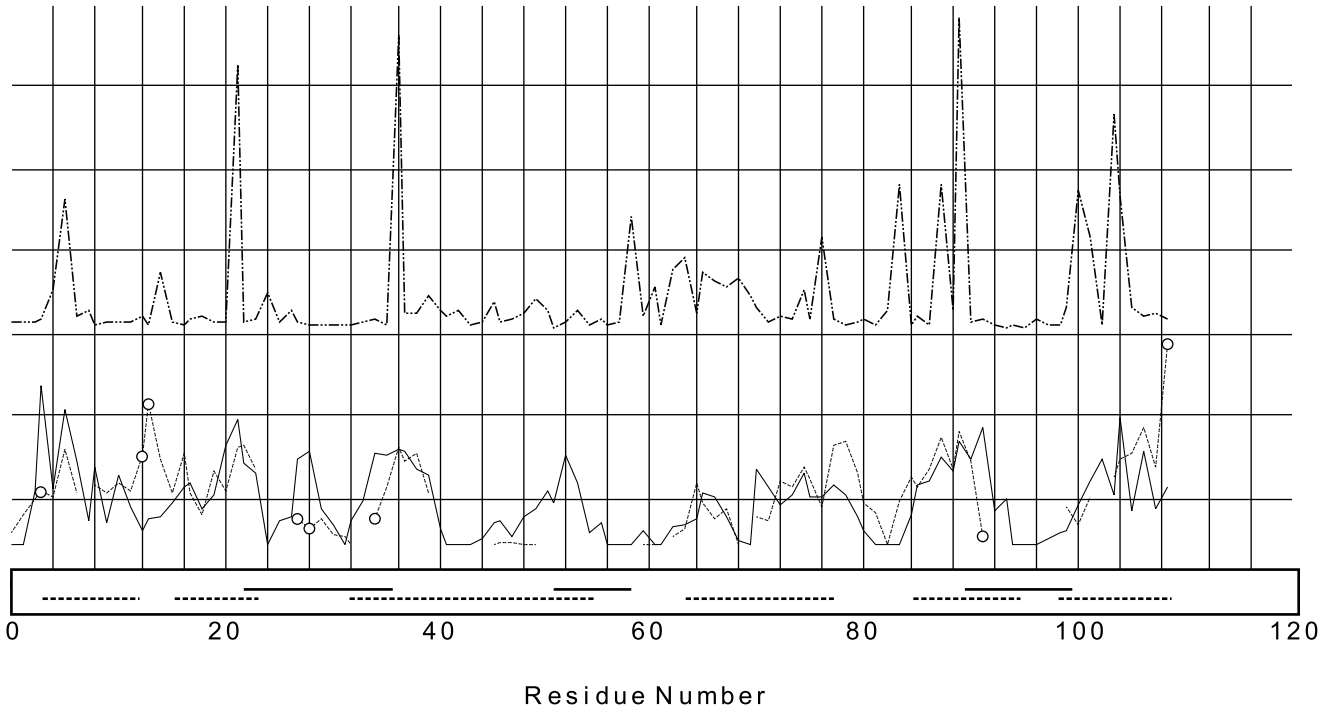

FIGURE 6 Inverse site entropy (— · —, upper part of the graph) as computed from the VL domain database (see Methods), model average SRSHs for Ig variable domain (——), and SRSHs for titin (— — —) from our calculations. Large values of inverse site entropy correspond to highly conserved amino acids. Large values of SRSH correspond to residues geometrically important for the folding process according to our calculations. The coincidence of the highest peaks of the different profiles (six peaks of inverse entropy out of nine coincide with the highest peaks in SRSH of Ig variable domain and titin) enlightens the very important role the corresponding amino acids play in the stabilization of the native state, according to their topological location. The same or a better match by random guess of the sites with highest conservation would occur with a probability of 0.016%. The graph also shows a good agreement in the SRSH profiles computed on titin and Ig variable domain. The disagreement is localized on eight sites only (○). Removing them from the comparison, the correlation coefficient between the two sets is 0.79. On the bottom of the graph the black solid lines indicate the CDRs of Ig variable domain, whereas the dashed lines mark residues in strand conformation.

burial is computed as the fraction of buried surface of each amino acid, then the correlation becomes even worse.

Comparison with data from multiple sequence analysis

A large amount of immunoglobulin variable domain sequences is available from standard databases (see Methods). They are easily alignable thanks to their high sequence identity. An aligned database has been created by Kabat and co-workers (Johnson and Wu, 2000). Because these sequences are evolved under the natural selective pressure for stability and functionality, they are supposed to show conserved amino acids in positions that are relevant for their stability (i.e., the cysteine residues are very well conserved). However, these sequences are not tested for intracellular expression. We compared our per-amino-acid specific heat with the variability associated with each position of the alignment (see Methods) and we found a striking correlation between our most relevant sites and the most conserved positions (Fig. 6). Among the 9 most conserved positions of the alignment, 6 belong to the 15 highest SRSH peaks (averaged over models A, B, and C). Another one is nearest neighbor of an SRSH peak. Only residues 82 and 98 (Kabat VL numeration) are outside topologically relevant regions. The probability to obtain the same or a better match (6 or more matches) by chance, i.e., picking up randomly two sets of, respectively, 9 and 15 sites, would be 0.016%. If n_1 is the number of relevant mutations and n the sequence length, the probability of guessing at least q of the n_1 by a random choice of n_2 sites is:

$$\sum_{m=q}^{\min(n_1,n_2)} \frac{\binom{n_1}{m}}{\binom{n}{n_2}} \binom{n-n_1}{n_2-m}. \tag{9}$$

Thus, highly conserved amino acids are located in the topologically relevant positions.

Comparison with data from site-directed mutagenesis

We compared our results with data from the literature in which site-directed mutagenesis of the Ig domain was cor-

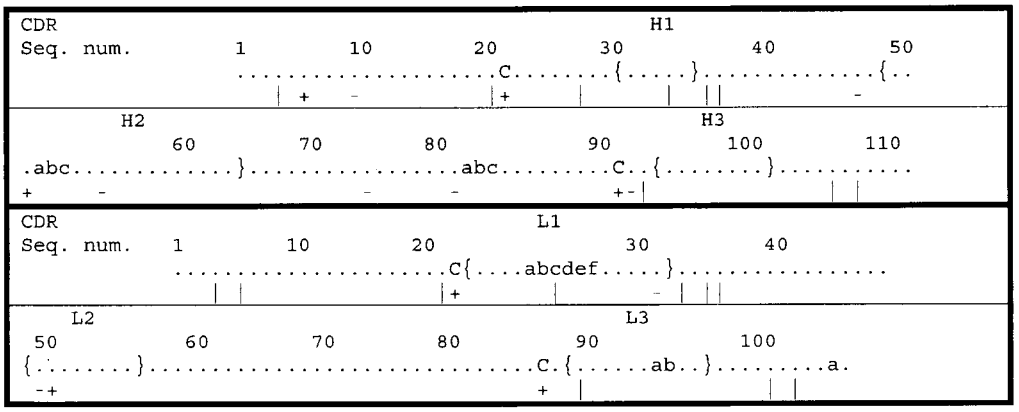

FIGURE 7 The immunoglobulin variable domain sequence VH (*top half*) and VL (*bottom half*). In the first row of each section, the CDR positions are indicated. In the second and third rows is the standard Kabat sequence numeration. In the fourth row, the experimental relevant mutations are indicated by a dash (-) and the calculated topologically relevant sites with a vertical line (|). When they overlap, a plus (+) is used. The positions of the cysteines are indicated with a C in the third row.

related experimentally with the ability of the corresponding Ig domain to fold under reducing conditions (Martineau et al., 1998; Kolmar et al., 1995; Langedijk et al., 1998; Proba et al., 1998). These data identify sites that have positive or negative effects on the overall stability of the domain in the absence of the disulfide bond. We included the two cysteines in these sites because of their obvious relevance. We compared these 13 sites of mutations with the 13 sites showing the highest SRSH peaks (averaged over models A, B, and C). Fig. 7 presents the results of the comparison. Sites are indicated according to Kabat standard numeration and aligned to our sequence. The match between our key sites (i.e., the one with the highest SRSH peak) and the relevant mutations is very good. An estimate of that can be given by measuring the probability to obtain the same match by chance. We found that the match with the experiments (5 matches of 13 hot sites versus 13 experimental point mutations on a sequence of 109 amino acids) would be equal or larger than our results show in only 3.5% of the possible random choices of the key sites. Furthermore, the use of the number of contacts for guessing the relevant mutations (the higher the number of contacts, the more relevant the mutation) gives a poor match with experimental data (38% of the random picks match better).

Comparison with data from titin

Titin is a protein involved in muscular elasticity; one of its domains shares the generic Ig topology. However, it differs from the Ig variable domain topology for several aspects; it is 89 residue long, so 20% shorter than the Ig variable domain. It also lacks an anti-parallel β-strand present in the Ig variable domain between residue 48 and 58 (standard Kabat VL numeration). This means that the quality of the alignment is worse than for VH-VL alignments. However, it represents a useful comparison to test the degree of variability of our measures on proteins belonging to a similar folding family. The SRSH profiles for the aligned residues of titin have been reported in Fig. 6. Because the alignment of the two structures is not very good, to get a more sensible comparison between the two SRSH profiles, we have smoothed the data, averaging the value c_i at position i with the half-weighted neighboring c. This lead to a correlation of 0.56 instead of the 0.47 as one would obtain without the smoothing procedure. However, the profiles clearly disagree in only eight localized positions. These can be identified as the sites where the absolute value of the difference in SRSH overcomes a fixed threshold. By removing them, the correlation coefficient between the two profiles becomes 0.79. The eight sites showing low agreement are localized in the top and bottom loops of the structures and are organized in three subgroups. One subgroup contains three residues and is localized on the non-CDR loops (see Fig. 8 as a reference); on titin, these residues interact with three C-terminal residues that are not present in the Ig variable domain; we argue that this perturbation is the cause of this disagreement. The second subgroup is composed of three contacting residues, two of which belong to the CDR1. This region is poorly aligned with very high root mean square deviation (6.0 Å versus 2.8 Å of the whole 83 aligned residues); that might be the reason for the discrepancy in SRSH. The last subgroup is composed of two contacting residues. In Ig variable domain, one of the residues is also in contact with the strand that is missing in titin, and the other one is located at the beginning (N-ter) of CDR3, which is eight residues long in Ig variable domain, whereas its titin counterpart is only two residues long.

The two SRSH profiles from titin and Ig variable domain agree in many regions. In particular, they agree where the conservation in VL sequences is the highest. Setting the same threshold used to identify the 15 highest SRSH peaks in the comparison with VL sequence entropy (see above),

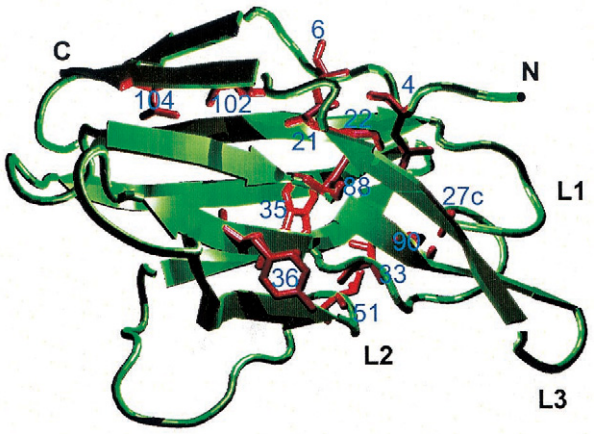

FIGURE 8 A cartoon representing the structure of the Ig variable domain. The residues with large SRSH are shown in red licorice and numbered according to standard Kabat numeration for VL domain. The N and C termini are indicated with the corresponding letter. CDRs are indicated as L1, L2, and L3.

we identified 16 SRSH peaks on titin structure. Seven of them coincide with Ig variable domain SRSH peaks, and six of these correspond to VL sequence entropy peaks, too. A comparison with titin sequence conservation has less statistical significance than with VL domain sequence conservation because we retrieved only tenths of similar sequences using standard alignment algorithms and sequence databases, which are very few with respect to the thousands of sequences in the Kabat database used to compile the VL inverse entropy profiles.

DISCUSSION

The comparison between the phase diagrams obtained in the different coupling regimes shows that activation of angular and dihedral energy terms brings an amplification to the cooperative behavior of the system. Dihedral and angular interactions are local multi-body interactions. In our model, their presence determines the disappearance of intermediate states between the folded and unfolded conformation and the enhancement of an all-or-none behavior in which all the residues cooperatively promote one of the two states. Indeed, the increase in the stiffness of the chain due to the angular and dihedral terms determines the narrowing of the allowed conformational space and the exclusion of the region occupied by the intermediate state in model A.

Experimentally measured thermal denaturation of ScFv domains (Martineau et al., 1998) is in accordance with results from regimes B and C showing a two-state transition from the folded to the unfolded conformation.

Furthermore, in our model, stability increases only if both dihedral and angular interactions are present. This directly descends from the fact that both torsional and angular degrees of freedom have to be fixed to uniquely determine the conformation of a polymer chain. That is in agreement with the increase in denaturation resistance due to secondary structure optimization experimentally found (Niggemann and Steipe, 2000).

Despite the diverse thermodynamic behavior of the system to heating and cooling in the three cases, the agreement in SRSH profiles (Fig. 5) is in accordance with the idea that the topology of the native state plays the same role in the three regimes. Furthermore, the weak correlation with the number of contacts per amino acid indicates that the summation on the native contacts to obtain the SRSHs does not yield trivial results. The per-amino-acid number of contacts represents, of course, an important part of the topological features of the proteins. Nevertheless, other topological features, such as chain connectivity, play a determinant role. That indication is confirmed by the fact that the number of contacts per amino acid is not enough to guess the relevance of the amino acids in the folding process, as shown in the previous section. On the contrary, our topology-based estimate is far more reliable and allows a non-casual discrimination of important residues.

Finally, the surprising match of the single-residue specific heat profiles with the degree of variability of the amino acids on generic Ig variable domain sequences is a strong indication that the topology of the native state has a very important role in the folding process of this type of proteins, too. The non-sequence-specific nature of this evidence and the extent of its validity to slightly different topologies are confirmed by the match between the conserved sites on VL sequences and the high SRSH sites on titin. The role of topology on the degree of conservation of amino acids has been addressed by different methods (reviewed in Gunasekaran et al., 2001).

The amino acids with high SRSH are supposed to play a relevant role in the folding process due to geometrical reasons; mutations occurring at those sites might affect the protein stability and/or its folding mechanism depending on the strength of the perturbation. The method that we used allows identifying them in a general and sequence-independent way. Our results are valid for proteins sharing the same Ig variable domain topology (without the covalent disulfide bond). Moreover, the agreement of SRSH profiles of titin and Ig variable domain underscores the robustness of our results with respect to topology changes.

However, the relevance of a mutation on the folding behavior of a given sequence would generally depend both on the geometric position in which it occurs and on the detailed chemical changes it produces. The approach used here addresses only the first issue and takes into account the local detailed chemistry underlying the positive and negative interactions only in an average way. Indeed, some mutations can be very effective for a particular Ig sequence and irrelevant for a different one because of the different chemical context in which they occur; thus, we do not

expect that all relevant mutations lie on high SRSH sites and vice versa.

The sites that we identified represent good targets for the design of optimized intracellular antibodies. Both a semi-rational method and a selection-based approach (see Introduction) would take advantage of a reduction in the number of sites on which the optimization has to be performed. In the first case, the search for stabilizing interactions would be focused on these sites only. In the second case, libraries could be built up with antibody sequences in which interactions involving these sites have been optimized, allowing a higher yield of intrabody domains than standard libraries do.

REFERENCES

Allen, M. P., and D. J. Tildesley. 1989. Computer Simulation of Liquids. Clarendon Press, Oxford, UK.

Alm, E., and D. Baker. 1999. Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures [see comments]. *Proc. Natl. Acad. Sci. U.S.A.* 96:11305–11310.

Banavar, J. R., and A. Maritan. 2001. Computational approach to the protein-folding problem. *Proteins.* 42:433–435.

Biocca, S., M. S. Neuberger, and A. Cattaneo. 1990. Expression and targeting of intracellular antibodies in mammalian cells. *EMBO J.* 9:101–108.

Biocca, S., P. Pierandrei-Amaldi, and A. Cattaneo. 1993. Intracellular expression of anti-p21ras single chain Fv fragments inhibits meiotic maturation of *Xenopus* oocytes. *Biochem. Biophys. Res. Commun.* 197: 422–427.

Biocca, S., F. Ruberti, M. Tafani, P. Pierandrei-Amaldi, and A. Cattaneo. 1995. Redox state of single chain Fv fragments targeted to the endoplasmic reticulum, cytosol and mitochondria. *Biotechnology.* 13: 1110–1115.

Bird, R. E., K. D. Hardman, J. W. Jacobson, S. Johnson, B. M. Kaufman, S. M. Lee, T. Lee, S. H. Pope, G. S. Riordan, and M. Whitlow. 1988. Single-chain antigen-binding proteins. [Published erratum appears in *Science* 244:409.] *Science.* 242:423–426.

Brooks, B. R., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. 1983. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* 4:187–217.

Cardinale, A., M. Lener, S. Messina, A. Cattaneo, and S. Biocca. 1998. The mode of action of Y13–259 scFv fragment intracellularly expressed in mammalian cells. *FEBS Lett.* 439:197–202.

Cattaneo, A., and S. Biocca. 1997. Intracellular Antibodies: Development and Applications. Springer, New York.

Cattaneo, A., and S. Biocca. 1999. The selection of intracellular antibodies. *Trends Biotechnol.* 17:115–121.

Cecconi, F., C. Micheletti, P. Carloni, and A. Maritan. 2001. Molecular dynamics studies on HIV-1 protease: drug resistance and folding pathways. *Proteins.* 43:365–372.

Chiti, F., N. Taddei, P. M. White, M. Bucciantini, F. Magherini, M. Stefani, and C. M. Dobson. 1999. Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nat. Struct. Biol.* 6:1005–1009.

Clementi, C., H. Nymeyer, and J. N. Onuchic. 2000. Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* 298:937–953.

Duan, L., H. Zhang, J. W. Oakes, O. Bagasra, and R. J. Pomerantz. 1994. Molecular and virological effects of intracellular anti-Rev single-chain variable fragments on the expression of various human immunodeficiency virus-1 strains. [Published errata appear in *Hum. Gene Ther.* 8:510 and 9:765.] *Hum. Gene Ther.* 5:1315–1324.

Ferrenberg, A. M., and R. H. Swendsen. 1989. Optimized Monte Carlo data analysis. *Phys. Rev. Lett.* 63:1195–1198.

Freire, E., K. P. Murphy, J. M. Sanchez-Ruiz, M. L. Galisteo, and P. L. Privalov. 1992. The molecular basis of cooperativity in protein folding. Thermodynamic dissection of interdomain interactions in phosphoglycerate kinase. *Biochemistry.* 31:250–256.

Frisch, C., H. Kolmar, and H. J. Fritz. 1994. A soluble immunoglobulin variable domain without a disulfide bridge: construction, accumulation in the cytoplasm of *Escherichia coli*, purification and physicochemical characterization. *Biol. Chem. Hoppe-Seyler.* 375:353–356.

Galzitskaya, O. V., and A. V. Finkelstein. 1999. A theoretical search for folding/unfolding nuclei in three-dimensional protein structures [see comments]. *Proc. Natl. Acad. Sci. U.S.A.* 96:11299–11304.

Gargano, N., and A. Cattaneo. 1997. Inhibition of murine leukaemia virus retrotranscription by the intracellular expression of a phage-derived anti-reverse transcriptase antibody fragment. *J. Gen. Virol.* 78: 2591–2599.

Gilbert, H. F. 1990. Molecular and cellular aspects of thiol-disulfide exchange. *Adv. Enzymol. Relat. Areas Mol. Biol.* 63:69–172.

Go, N. 1983. Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.* 12:183–210.

Goto, Y., and K. Hamaguchi. 1979. The role of the intrachain disulfide bond in the conformation and stability of the constant fragment of the immunoglobulin light chain. *J. Biochem. (Tokyo).* 86:1433–1441.

Gunasekaran, K., S. J. Eyles, A. T. Hagler, and L. M. Gierasch. 2001. Keeping it in the family: folding studies of related proteins. *Curr. Opin. Struct. Biol.* 11:83–93.

Huston, J. S., D. Levinson, M. Mudgett-Hunter, M. S. Tai, J. Novotny, M. N. Margolies, R. J. Ridge, R. E. Bruccoleri, E. Haber, and R. Crea. 1988. Protein engineering of antibody binding sites: recovery of specific activity in an anti-digoxin single-chain Fv analogue produced in *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.* 85:5879–5883.

Hwang, C., A. J. Sinskey, and H. F. Lodish. 1992. Oxidized redox state of glutathione in the endoplasmic reticulum. *Science.* 257:1496–1502.

Improta, S., A. S. Politou, and A. Pastore. 1996. Immunoglobulin-like modules from titin I-band: extensible components of muscle elasticity. *Structure.* 4:323–337.

Johnson, G., and T. T. Wu. 2000. Kabat database and its applications: 30 years after the first variability plot. *Nucleic Acids Res.* 28:214–218.

Jung, S. and A. Pluckthun. 1997. Improving in vivo folding and stability of a single-chain Fv antibody fragment by loop grafting. *Protein Eng.* 10:959–966.

Kolmar, H., C. Frisch, K. Gotze, and H. J. Fritz. 1995. Immunoglobulin mutant library genetically screened for folding stability exploiting bacterial signal transduction. *J. Mol. Biol.* 251:471–476.

Langedijk, A. C., A. Honegger, J. Maat, R. J. Planta, R. C. van Schaik, and A. Pluckthun. 1998. The nature of antibody heavy chain residue H6 strongly influences the stability of a VH domain lacking the disulfide bridge. *J. Mol. Biol.* 283:95–110.

Marasco, W. A. 1995. Intracellular antibodies (intrabodies) as research reagents and therapeutic molecules for gene therapy. *Immunotechnology.* 1:1–19.

Maritan, A. 2001. Geometrical aspects of protein folding. *In* Proceedings of the International School of Physics "Enrico Fermi" course CXLV. Societa' italiana di Fisica, Varenna, Italy. in press.

Maritan, A., C. Micheletti, and J. R. Banavar. 2000a. Role of secondary motifs in fast folding polymers: a dynamical variational principle. *Phys. Rev. Lett.* 84:3009–3012.

Maritan, A., C. Micheletti, A. Trovato, and J. R. Banavar. 2000b. Optimal shapes of compact strings. *Nature.* 406:287–290.

Martineau, P., P. Jones, and G. Winter. 1998. Expression of an antibody fragment at high levels in the bacterial cytoplasm. *J. Mol. Biol.* 280: 117–127.

Martinez, J. C., M. T. Pisabarro, and L. Serrano. 1998. Obligatory steps in protein folding and the conformational diversity of the transition state [see comments]. *Nat. Struct. Biol.* 5:721–729.

Mhashilkar, A. M., J. Bagley, S. Y. Chen, A. M. Szilvay, D. G. Helland, and W. A. Marasco. 1995. Inhibition of HIV-1 Tat-mediated LTR transactivation and HIV-1 infection by anti-Tat single chain intrabodies. *EMBO J.* 14:1542–1551.

Micheletti, C., J. R. Banavar, A. Maritan, and F. Seno. 1999. Protein structures and optimal folding from a geometrical variational principle. *Phys. Rev. Lett.* 82:3372–3375.

Munoz, V., E. R. Henry, J. Hofrichter, and W. A. Eaton. 1998. A statistical mechanical model for beta-hairpin kinetics. *Proc. Natl. Acad. Sci. U.S.A.* 95:5872–5879.

Nieba, L., A. Honegger, C. Krebber, and A. Pluckthun. 1997. Disrupting the hydrophobic patches at the antibody variable/constant domain interface: improved in vivo folding and physical characterization of an engineered scFv fragment. *Protein Eng.* 10:435–444.

Niggemann, M., and B. Steipe. 2000. Exploring local and non-local interactions for protein stability by structural motif engineering. *J. Mol. Biol.* 296:181–195.

Pearlman, D. A., D. A. Case, J. W. Caldwell, W. S. Ross, T. E. Cheatham, S. Debolt, D. Ferguson, G. Seibel, and P. Kollman. 1995. AMBER, a package of computer programs for applying molecular mechanics, normal-mode analysis, molecular-dynamics and free-energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.* 91:1–41.

Proba, K., A. Honegger, and A. Pluckthun. 1997. A natural antibody missing a cysteine in VH: consequences for thermodynamic stability and folding. *J. Mol. Biol.* 265:161–172.

Proba, K., A. Worn, A. Honegger, and A. Pluckthun. 1998. Antibody scFv fragments without disulfide bonds made by molecular evolution. *J. Mol. Biol.* 275:245–253.

Richardson, J. H., and W. A. Marasco. 1995. Intracellular antibodies: development and therapeutic potential. *Trends Biotechnol.* 13:306–310.

Rudikoff, S. and J. G. Pumphrey. 1986. Functional antibody lacking a variable-region disulfide bridge. *Proc. Natl. Acad. Sci. U.S.A.* 83: 7875–7878.

Settanni, G. 2000. The role of topology in immunoglobulin variable domain. Talk at Protein Folding and Evolution, International School of Physics "Enrico Fermi" course CXLV. Societa' italiana di Fisica, Varenna, Italy.

Tavladoraki, P., E. Benvenuto, S. Trinca, D. De Martinis, A. Cattaneo, and P. Galeffi. 1993. Transgenic plants expressing a functional single-chain Fv antibody are specifically protected from virus attack. *Nature.* 366: 469–472.

Vendruscolo, M., R. Najmanovich, and E. Domany. 2000. Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading? *Proteins.* 38:134–148.

Visintin, M. 2000. The development of the intrabody trap technology (ITT) for functional genomics. Ph.D. thesis. International School for Advanced Studies, Trieste, Italy. 128 pp.

Visintin, M., E. Tse, H. Axelson, T. H. Rabbitts, and A. Cattaneo. 1999. Selection of antibodies for intracellular function using a two-hybrid in vivo system. *Proc. Natl. Acad. Sci. U.S.A.* 96:11723–11728.

Williams, A. F., and A. N. Barclay. 1988. The immunoglobulin superfamily: domains for cell surface recognition. *Annu. Rev. Immunol.* 6:381–405.

Wolynes, P. G., J. N. Onuchic, and D. Thirumalai. 1995. Navigating the folding routes [see comments]. *Science.* 267:1619–1620.

Worn, A., and A. Pluckthun. 1998. Mutual stabilization of VL and VH in single-chain antibody fragments, investigated with mutants engineered for stability. *Biochemistry.* 37:13120–13127.

Worn, A., and A. Pluckthun. 1999. Different equilibrium stability behavior of ScFv fragments: identification, classification, and improvement by protein engineering. *Biochemistry.* 38:8739–8750.