Low-throughput model design of protein folding inhibitors

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · MAY 2007

Impact Factor: 2.63 · DOI: 10.1002/prot.21275 · Source: PubMed

5 AUTHORS, INCLUDING:

R. A. Broglia
INFN - Istituto Nazionale di Fisica Nucleare
519 PUBLICATIONS **11,513** CITATIONS
SEE PROFILE

Guido Tiana
University of Milan
82 PUBLICATIONS **1,537** CITATIONS
SEE PROFILE

Ludovico Sutto
University College London
35 PUBLICATIONS **344** CITATIONS
SEE PROFILE

Davide Provasi
Icahn School of Medicine at Mount Sinai
69 PUBLICATIONS **1,332** CITATIONS
SEE PROFILE

PROTEINS: Structure, Function, and Bioinformatics 67:469–478 (2007)

Low-Throughput Model Design of Protein Folding Inhibitors

R.A. Broglia,[1,2] G. Tiana,[1*] L. Sutto,[1] D. Provasi,[1] and V. Perelli[1]
[1]*Department of Physics, University of Milano and INFN, sez. di Milano, Milano 20133, Italy*
[2]*The Niels Bohr Institute, Blegdamsvej 17, 2100 Copenhagen, Denmark*

ABSTRACT **The stabilization energy of proteins in their native conformation is not distributed uniformly among all the amino acids, but is concentrated in few (short) fragments, fragments which play a key role in the folding process and in the stability of the protein. Peptides displaying the same sequence as these key fragments can compete with the formation of the most important native contacts, destabilizing the protein and thus inhibiting its biological activity. We present an essentially automatic method to individuate such peptidic inhibitors based on a low-throughput screening of the fragments which build the target protein. The efficiency and generality of the method is tested on proteins Src-SH3, G, CI2, and HIV-1-PR with the help of a simplified computational model. In each of the cases studied, we find few peptides displaying strong inhibitory properties, properties which are quite robust with respect to point mutations. The possibility of implementing the method through low-throughput experimental screening of the target protein is discussed.**

Key words: folding inhibitors; drug design; simplified model; Monte Carlo sampling

INTRODUCTION

Conventionally, protein inhibitors act by capping the active site (or some allosteric site), thus preventing the binding of the substrate. The search for molecules that bind tightly to the active site is the first step in the development of a lead, which eventually becomes a drug. Present powerful methods to obtain such leads include high-throughput screening,[1] QSAR methods,[2] and rational design.[3–5] However, the applicability of these methods is limited by a number of drawbacks. For example, high-throughput screening is very time-consuming, QSAR methods are weakly portable, while rational design rely on the knowledge of realistic potential functions and of computationally demanding optimizations. Moreover, the design process can become nonoperative in cases where the drug induces escape mutants with decreased affinity to the target protein, a situation encountered for example in the case of some viral proteins.

An alternative approach is to inhibit a protein by destabilizing its native state. There exists growing evidence which testify to the fact that globular proteins are mostly stabilized by a core of mutually interacting residues.[6–10] The fragments of the protein which carry these "hot" residues (called "local elementary structures," or LES, in Ref. 9) play also an important role in the folding of the protein, as they get structured in the early stages of folding process. Theoretical considerations[11,12] and experimental assays on the HIV-1-Protease[13] have shown that it is possible to destabilize the native state of proteins by means of peptides (called "p-LES"[11]) displaying the same sequence as that associated with LES. These peptides can substitute the corresponding LES in binding the rest of the protein, shifting its equilibrium toward the unfolded state and thus decreasing its biological activity. The advantage of this approach of drug design over conventional ones is that one has nothing to design: it is the protein itself which suggests its own inhibitor. Moreover, since p-LES bind to the residues, which are most important to stabilize the protein, it is unlikely that the protein can mutate them to escape the inhibitor,[13] without at the same time jeopardizing its stability.

From the above discussion it emerges that the main task to be carried out in designing a nonconventional inhibitor is that of finding the LES of the target protein. This can be done, for example, by: (a) searching for clusters of strongly interacting residues with the help of empirical potentials of mean force functions,[14] (b) evaluating the degree of conservation of residues in families of structurally similar proteins,[7] (c) carrying out simulations with the help of simplified models,[15–17] (d) a combination of (a–c). While this strategy is quite effective, it is very time consuming and little transferable. In an attempt to improve this situation we present, in what follows, a fully automated protocol for individuating the LES of a protein.

Whatever the size of a protein is, the associated LES must be short, of the order of 5–15 residues long, so as to be able to become structured in the early stages of the folding process.[18] Consequently, to individuate the LES of a protein one has to test the inhibitory properties

*Correspondence to: G. Tiana, Department of Physics, University of Milano and INFN, sez. di Milano, via Celoria 16, Milano 20133, Italy. E-mail: tiana@mi.infn.it

Received 9 May 2006; Revised 19 September 2006; Accepted 25 September 2006

Published online 12 February 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21275

of a set of peptides of length 5–15. Each of these peptides has a sequence identical to a segment of the target protein, and displays a consistent overlap (20–50%) with the neighboring peptides. The full set of peptides covers the entire protein with a consistent amount of redundancy. The (few) peptides that mostly destabilize the protein, inhibiting its activity are likely to be the p-LES. This protocol can be implemented equally well computationally, as we do in the present work, or experimentally, by means of enzymatic or structural assays. Typically, the number of peptides to be tested either computationally or experimentally is of the order of few tens. This is indeed a much smaller effort than that used in pharmaceutical high-throughput screening, where hundreds of thousand molecules are usually tested. The present approach is more similar to that used in structural biology, for example to locate amyloidogenic fragments in aggregation-prone proteins (see, e.g., Ref. 19).

Note that, aside from the potential interest concerning the design of leads of nonconventional inhibitors, the outcome of the above protocol will also shed light on the mechanism which is at the basis of the folding mechanism. In particular, it will be helpful discussing the validity of the hierarchical scenario.[20]

In principle, p-LES can be directly used as drugs, provided they meet the requirements of *in vivo* stability, specificity, good pharmacokinetics, and bioavailability. However, this will not be in general the case, in particular, because of the peptidic character of the p-LES as well as because of its length. Eventually, one should modify the peptide employing, for example, bioisosteres, modified terminals and modified amino acids, or use the peptide as template for mimetic molecules.

Folding inhibitor peptides interact specifically with those fragments of the target protein which are mostly responsible for its stability. Mutations in the associated stabilization core lead, as a rule, to protein denaturation.[14] Consequently, inhibitor peptides are less likely to induce substitutions, which eventually lead to escape mutants, than conventional drugs. Of course, this does not mean that there is no escape way for the protein to decrease its affinity to a p-LES, for example, through multiple, coordinated mutations. In fact, proteins can change their stabilization core during evolution through pathways of active mutant proteins.[21] These concerted mutations are anyway consistently less probable than single-point mutations, thus making p-LES, although not completely immune to resistance, at least definitely more robust than traditional drugs. Furthermore, it is in principle possible, with the help of models, to individuate the (few) stabilization cores associated with a given native conformation.[22] The inhibitor action provided by a cocktail of p-LES having the same sequence as representatives of each of these cores will be very difficult to evade.

In what follows we apply the protocol described above to four globular proteins (Src-SH3, Protein G, CI2, and HIV-1-Protease), performing computer simulations, which allow one to assess the range of validity and the degree of generality of the method.

METHODS

To find the fragments of the protein which are efficient inhibitors, we resort to an exhaustive search through the protein sequence. The whole protein, of length n, is divided in fragments of length L, also allowing for an overlap of L_{ov} between the fragments. Each of the $n/(L\text{-}L_{ov})$ fragments corresponds to a peptide to be tested as inhibitor. Since we expect the LES of the protein to have a length of the order of 10 amino acids, we will use peptides of comparable length. In fact, LES are structured in the transition state, and protein engineering experiments[23–25] show that typically structured regions in the transition state span of the order of 10 residues. Of course, the precise value of L to be used has some degree of arbitrariness. In the present study it is the inhibitory efficiency to justify *a posteriori* the particular choice of L. We have also chosen $L_{ov} \approx L/2$, that is a trade-off between the accuracy of sequence scan and computational heaviness. Because a typical single-domain protein has $n \approx 150$, this prescription leads to $\approx$30 peptides to be tested. Multi-domain proteins are, as a rule, composed of modules of characteristic length equal again to[26] 150, displaying the ability to fold independently of each other.[27,28] Consequently, one can repeat the above considerations for each module. The inhibitory ability of $\approx$30 peptides can be tested with the help of either computational or experimental probes. In what follows we illustrate the protocol in terms of Monte Carlo simulations making use of a simplified computational model, but the standard experimental techniques to detect enzymatic activity or structural stability can be used as well (see e.g., Ref. 13).

Operatively, we have used $L = 6$, $L_{ov} = 2$ for proteins Src-SH3, G, and CI2, while we used $L = 10$, $L_{ov} = 5$ in the case of HIV-1 protease. This choice leads to 15, 14, 16, and 21 different peptides to be tested, respectively, and thus to an equal number of simulations to be carried out. In each of these simulations in which three of the peptides are evolved in the presence of the corresponding target protein (this is a typical peptide/protein ratio found to be sufficient to give a clear signal in experimental assays,[13] and at the same time avoid peptidic aggregation. For a discussion of the dependence of the inhibitory effectiveness on the peptidic ratio see Ref. 12). The cpu time needed to study the time evolution of such a system in the case of a protein containing 60 residues (like Src-SH3) is of the order of 2 h on a 3 GHz Xeon.

The model we employ is a modified G$\bar{o}$ model,[29] where each amino acid is represented by a spherical bead interacting with the other amino acids through a contact potential of range 7.5 Å. The contact energies depend on the pair of interacting residues. The native contacts are attractive, while non-native contacts are repulsive. The native energies are either extracted from experimental[12,29] values of $\Delta\Delta G_{U-N}$ or calculated by averaging the GROMACS force field, in explicit solvent, over the atoms which build each amino acid.[15] The model is described in the Appendix (see also Ref. 12). The parameters con-

trolling the interaction between the protein and a peptide are taken to be identical to those controlling the interaction between the fragment of the protein corresponding to the peptide in question, and the rest of the protein.

The equilibrium probabilities p_N^0 and $p_N^{i\text{-}j}$ with which the target protein by itself or in presence of three peptides (corresponding to fragment *i-j* of the protein) respectively, populate the native state N in a cubic box of size 100 Å is calculated making use of Monte Carlo simulations[30] (see Appendix for the operative definition of the state N). It is found that the systematic study of the ratio $r^{i\text{-}j} = p_N^{i\text{-}j}/p_N^0$ makes it possible to fulfill the goal of the protocol: to individuate, for each target protein, the peptides displaying the largest inhibitor properties, in other words, the peptides displaying a small value of $r^{i\text{-}j}$. Because the simulations are carried out for a sufficiently long time (of the order of 10^9 Monte Carlo steps) in order to describe the equilibrium of the system, the observed inhibitor ability displayed by the peptides is independent whether the target protein starts from a denatured or from a folded conformation. As a rule, the simulations are to be carried out at temperatures close to the folding temperature so as to make possible to collect good statistics within sensible computational times. Only HIV-1-PR has been studied at a lower temperature to be consistent with the simulations of Ref. 12. An important point connected with this issue refers to the time needed by the p-LES to destabilize the protein. If the simulations or the experiments are performed at a temperature much lower than the folding temperature of the protein, the equilibrium-thermodynamic considerations discussed above still hold, but thermal fluctuations become smaller and then the time needed by the system to reach thermodynamic equilibrium become larger. In other words, in the case in which the folding temperature of the target protein is much higher than room temperature the *in vivo* applicability of the method may become problematic. On the other hand, proteins in the cell are expressed through the ribosome in an out-of-equilibrium conformation which, in principle, can be targeted by the p-LES.

RESULTS AND DISCUSSION

The above method is applied to the SH3 domain of src,[23] protein G,[24] Chymotrypsin Inhibitor 2[25] (CI2), and HIV-1 Protease[31] (HIV-1-PR). While only the inhibition of this last protein is of potential clinical interest, the other three proteins have been exhaustively characterized in protein folding studies. Within this context, CI2 is usually described as a protein which folds following a nonhierarchic nucleation-condensation scenario,[32] thus representing an important test to our approach.

The SH3 Domain

The SH3 domain displays five antiparallel β-sheets and an α-helix, and is composed of 60 residues (see Fig. 1, pdb code: 1FMK). It is an interesting benchmark for the method as it has been widely characterized through thermodynamic as well as kinetic experiments.[23]

Monte Carlo simulations of this protein were carried out making use of the G$\bar{o}$ model and of interaction parameters extracted from experimental values of $\Delta\Delta G_{U-N}$ (see Appendix). The ratio $p_N^{i\text{-}j}/p_N^0$ for each of the 15 peptides of length $L = 6$ tested are shown in Figures 2 and 3. These results clearly indicate that folding is inhibited by peptides spanning the regions 1–9 (p-LES 1–6 and 5–10), 17–29 (p-LES 17–22, 21–26, and 25–30), 33–42 (p-LES 33–38 and 37–42), and 45–50 (p-LES 45–50; see Fig. 3). These regions approximately correspond to the first beta strand, the diverging turn, the second beta strand, and the distal hairpin. A φ-value analysis[23] indicate that these structures display, in the transition state, a high degree of stability. This is particularly true for the distal hairpin.

To test the effect mutations have on the inhibitory ability of p-LES, we have repeated the calculations reported above but for the system composed of the mutated protein (see Appendix) and of three peptides of Type 37–42, chosen because of its strong inhibitory ability. The results are reported in Figure 2(b). From this figure it is seen that the only mutations able to lead to escape mutants are nonconservative mutations (n-c). In fact, in sequences carrying such mutations, the inhibitory ability of the peptide is much reduced as compared to that observed in the case of the wt sequence, but the resulting sequence is not able to fold any longer. Examples of these mutations are those at sites 18 and 48. From the figure it is also seen that mutations which do not affect the stability nor the folding ability of the protein (e.g., mutations on sites 24, 37, and 40) leave essentially unchanged the inhibitory ability of the peptide.

Protein G

The B1 domain of protein G is a component of the cell wall of streptococcals and binds to mammalian immunoglobulins. It is composed of 56 residues and its native structure is shown in Figure 1 (pdb code: 1PGB). We have tested the effect on protein G of 14 different peptides, following the same procedure used for SH3. The results are displayed in Figure 4(a) and indicate that peptides corresponding to fragments 41–46 and 49–54 have a consistent destabilization effect, while the others have essentially no effect. These fragments correspond to the second β-hairpin of protein G, which has been experimentally observed to be structured in the transition state.[24] In Figure 4(b) the effects of (n-c) mutations on the inhibitory ability of these peptides is displayed. It is seen that escape mutants, e.g. sequences with n-c mutations on Site 6 or Site 52, are in any case not able to fold.

Protein CI2

Chymotrypsin Inhibitor 2 is an inhibitor of serine proteases and is composed of 64 amino acids (cf. Fig. 1, pdb code: 3CI2). It is conventionally taken as example of the

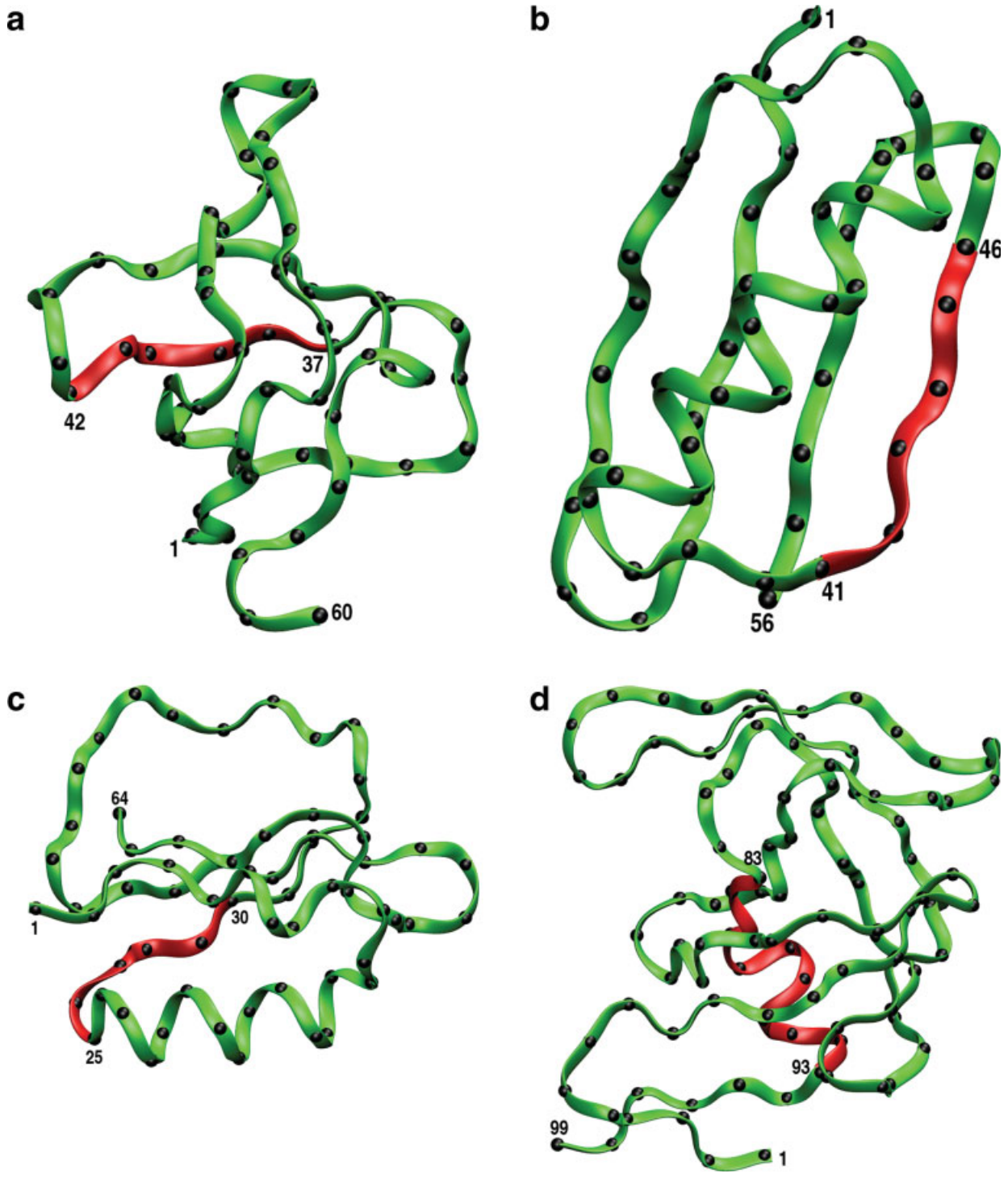

Fig. 1. A cartoon of the crystallographic structure of Src-SH3, Protein G, CI2, and HIV-1-PR. The p-LES corresponding to the fragment marked in red is the one we used in studying the effects of mutations (see text).

condensation-nucleation mechanism of folding, where secondary and tertiary interactions appear to form concurrently,[32] and thus a hierarchic mechanism is not evident. It was argued that this is because sequential events along the folding pathways are of difficult experimental detection.[20,29] We have tested the effect 16 different peptides have on CI2, following the procedure used in the case of the SH3 protein. The results are shown in Figure 5(a). From this figure it is seen that there are three peptides which completely destabilize the protein,

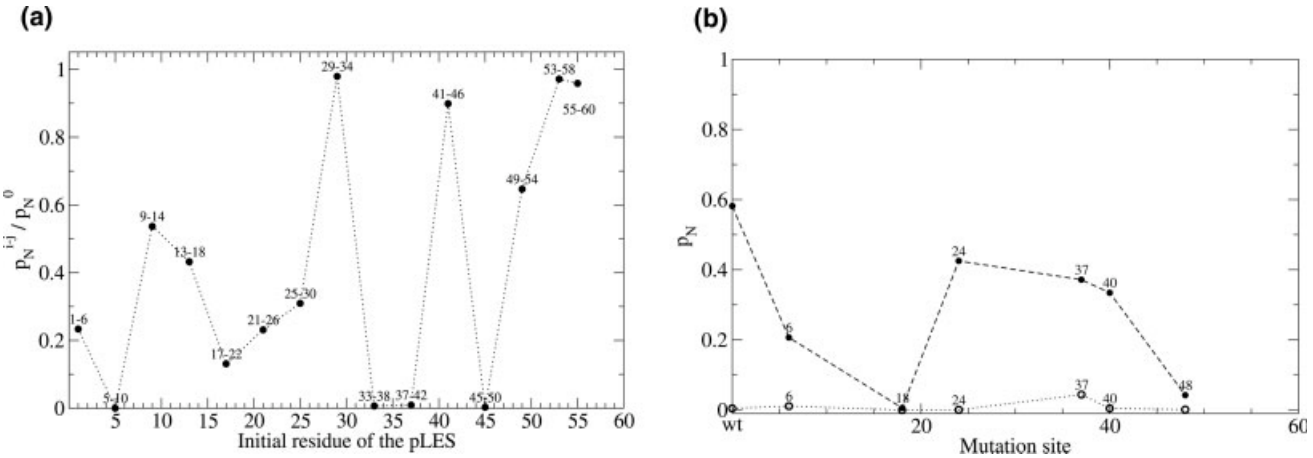

Fig. 2. (**a**) The population $p_N^{i,j}$ of the native state at $T = 1.05\ T_f$ for which $p_N^0 = 0.504$ (T_f being the folding temperature, i.e. the temperature at which the population of the native state is equal to that of the unfolded state, see Appendix) for a system composed of the Src-SH3 protein domain and three peptides whose sequence is identical to that of the fragment of the protein of length 6 starting at the site *i* indicated on the abscissa and ending at site $j = i + 5$, normalized with respect to the native state population of the protein by itself (p_N^0). The first peptide displays a sequence identical to segment 1–6 while the sequence of the last peptide coincides with that of segment 55–60. The overlap between the different peptides is 67%. (**b**) The population of the native state p_N for the (mutated) protein Src-SH3 alone (dashed curve through solid dots) and with three peptides 37–42 (dotted curve through open dots). The number associated with each of the open and solid dots indicate the site of the wt Src-SH3 sequence in which a n-c mutation has been introduced. The points at abscissa zero indicate the wild-type sequence. In both pictures the lines are to guide the eye.

namely those corresponding to fragments 25–30, 29–34, and 45–50, and one peptide which decreases the stability of the protein by about 50%, namely peptide 57–62. These results are consistent with the high value of $\Delta\Delta G_{U-N}$ experimentally observed[25] in correspondence with regions 24–34, 47–52, and 57–60 of the protein, thus suggesting that amino acids belonging to these regions are essential for the correct folding of the protein. The effect of (n-c) point mutations on the inhibitory role of peptide 25–30 is shown in Figure 5(b). It is again observed, as in the case of the SH3 and Protein G, that (n-c) mutations leading to escape mutants completely destabilize the protein.

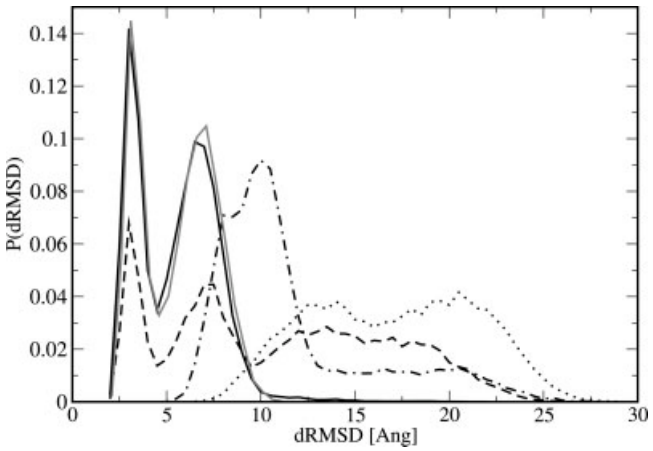

Fig. 3. The equilibrium distribution probability of the distance root mean square deviation dRMSD (see Appendix) calculated at $T = 1.05\ T_f$ for the system composed of the Src-SH3 protein domain alone (continuous black curve) and with three peptides characterized by the sequence of the protein fragments: 5–10 (dotted black curve), 45–50 (dot-dashed black curve), 13–18 (dashed black curve), and 53–58 (solid grey curve). The first two peptides strongly destabilize the protein toward the unfolded state (dRMSD >6), the effect of the third being less pronounced, while the last one leaves essentially unchanged the stability of the protein.

HIV-1 Protease

The protease of the HIV-1 virus is a dimer composed of two identical monomers of 99 amino acids each (see Fig. 1, pdb code: 1BVG). It is essential for the replication of the virus because it cleaves the polyprotein encoded by viral RNA, eventually producing each of the proteins needed in the assemblage of new virions. For this reason, it is one of the major targets of HIV-therapies. At neutral pH the dimer is at equilibrium with the monomer.[31] Consequently, we focus our attention on the destabilization of the monomer. We have tested 22 different peptides corresponding to fragments of the sequence of HIV-1-PR. The associated values of $r^{i,j} = p_N^{i,j}/p_N^0$ displayed in Figure 6(a) indicate that the peptides corresponding to fragment 9–19, 26–35, 41–50, 71–80, and 81–93 (i.e., 81–90 and 83–93) of the protein reduce by 60% or more the native state population (i.e., $r^{i,j} \leq 0.4$). In other words, these segments are efficient inhibitors of the folding of the enzyme.

We have analyzed the role (n-c) mutations play on the inhibitory properties of peptide 83–93 [see Figs. 6(b) and 7 and Table I]. In particular those leading to escape mutants in patients treated with FDA-approved antiviral drugs,[33] like for example K20MTIV and E35G (Ritonavir (RTN)), L63PSTCQH (RTN, Nelfinavir (NLF), Indinavir (IND), Sanquinavir (SQV), Amprenavir (APR)), A71T (RT, NLF, IND, SQV), and so forth. The calculations indicate that these mutations affect neither the inhibitory ability of the (83–93) peptide, nor the stability of

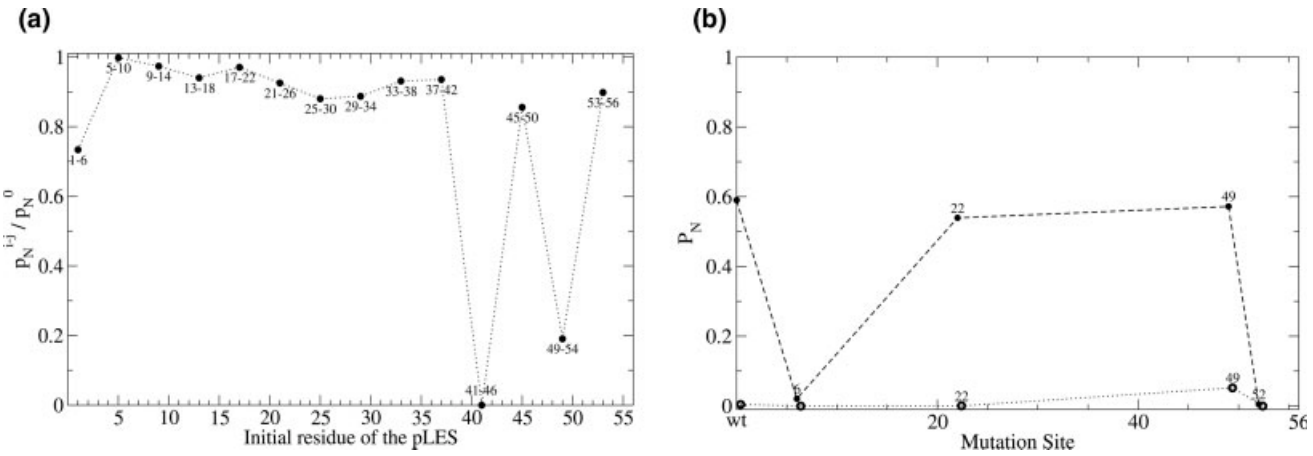

Fig. 4. (**a**) The population of the native state at $T = 0.94\ T_f$ for a system composed of the protein G and three peptides whose sequence is identical to that of the fragment of the protein of length 6 which starts at the site indicated on the abscissa, normalized with respect to the population of the native state of the protein by itself (p_N^0). The peptides start with segment 1–6 and end with segment 53–56, the overlaps between the different peptides being 67%. (**b**) The population of the native state p_N for the protein G alone (dashed curve through solid dots) and with three peptides 41–46 (dotted curve through open dots) with different mutation on the sequence. The points at abscissa zero indicate the wild-type sequence. In both pictures the lines are to guide the eye.

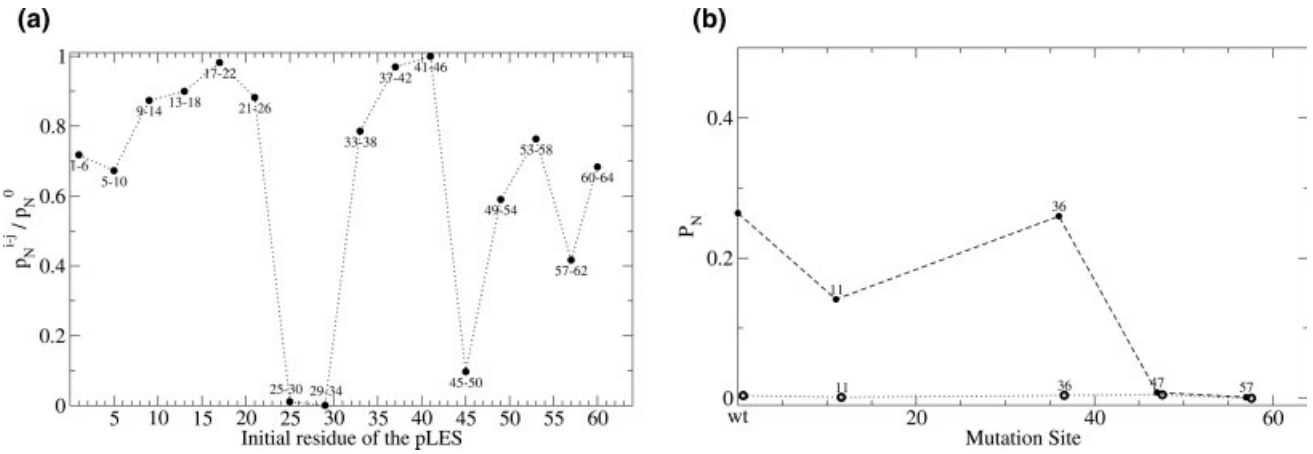

Fig. 5. (**a**) The population of the native state at $T = 1.1\,T_f$ for a system composed of the CI2 and three peptides whose sequence is identical to that of the fragment of the protein of length 6 which starts at the site indicated on the abscissa, normalized with respect to the population of the native state of the protein by itself (p_N^0). The peptides start with segment 1–6 and end with segment 60–64, the overlaps between the different peptides being 67%. (**b**) The population of the native state p_N for the protein CI2 alone (dashed curve through solid dots) and with three peptides 25–30 (dotted curve through open dots) with different mutation on the sequence. The points at abscissa zero indicate the wild-type sequence. In both pictures the lines are to guide the eye.

the protease. Similar results are found in the case of other (n-c) mutations on sites 31, 37, and 40 (randomly selected among the "cold" sites of the protease, that is, sites which do not play any special role neither in the stability nor in the folding of the protein[14]).

On the other hand, (n-c) mutations which quench the inhibitory ability of the 83–93 peptide (e.g., (n-c) mutations on sites 33 and 85), destabilize the protein in an important way, consistent with the fact that 33 and 85 are "hot" sites of the HIV-1-PR.[12] This result can be better understood in terms of the folding inhibitor factors $F_I(\mathrm{i})$ (see Fig. 7 and Table I) associated with the 83–93 peptide acting on HIV-1-PR monomers which carry a (n-c) mutation on site i of the wild type sequence. Noting that $F_I(\mathrm{i}) = 1$ implies that the peptide has no inhibitory effect on the mutated sequence, while $F_I(\mathrm{i}) = 0$ reflects the fact that the peptide is extremely efficient in inhibiting the folding of the HIV-1-PR. The values $F_I(33) = 0.95$ and $F_I(85) = 0.65$ (as compared to an average value $\langle F_I \rangle = 0.36$ and a standard deviation $\sigma = 0.21$) are illustrative of the (nonlinear) sum of two different phenomena: (a) the important role played the hot amino acids 33 and 85 on the folding and stability of the protease; (b) the docking mechanism of the non-conventional inhibitor (83–93 peptide) to the protease. In fact, n-c mutations on site 33 prevents not only the protein to fold, but at the same time the peptide to attach to it, as testified by the values $p_N^0(33) = 0.37$ and $F_I(33) = 0.95$ (values to be compared with those associated with wt sequence $p_N^0(\mathrm{wt}) = 0.95$ $F_I(\mathrm{wt}) = 0.45$). On the other

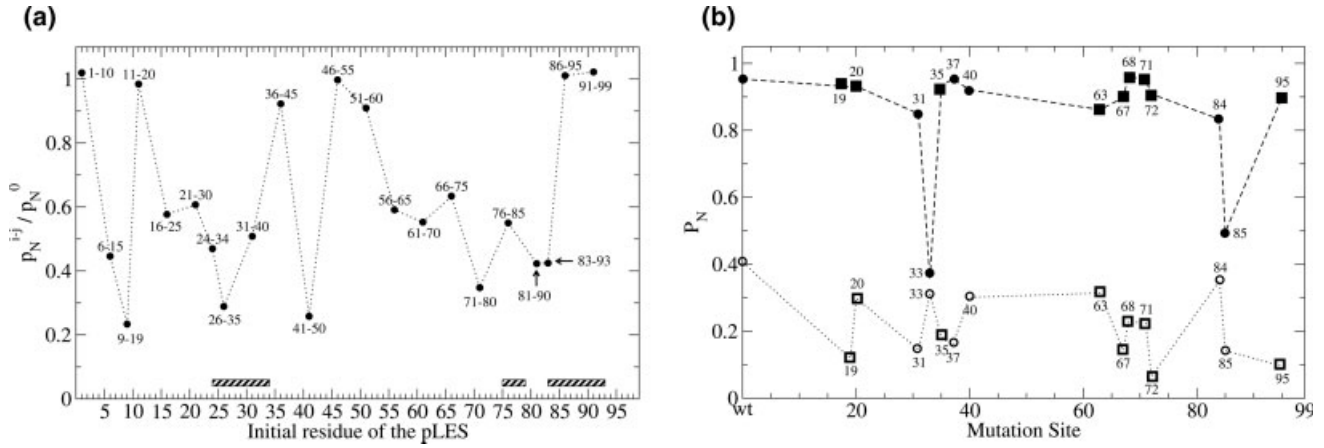

Fig. 6. (**a**) The population of the native state at $T = 0.67T_f$ for a system composed of the HIV-1-PR and three peptides whose sequence is identical to that of the fragment of the protein of length 10 which starts at the site indicated on the abscissa, normalized with respect to the population of the native state of the protein by itself (p_N^0). The peptides start with segment 1–10 and end with segment 91–99, the overlaps between the different peptides being 50%. The three highlighted sequence 24–34, 75–78, and 83–93 correspond to the HIV-1-PR LES (see Ref. 12). (**b**) The population of the native state p_N for the protein HIV-1-PR alone (dashed curve through solid symbols) and with three peptides 83–93 (dotted curve through open symbols) with different mutation on the sequence. (n-c) mutations induced by drug pressure are represented with a square, while spontaneous (n-c) mutations are shown in terms of a circle. The points at abscissa zero indicate the wild-type sequence. In both pictures the lines are to guide the eye.

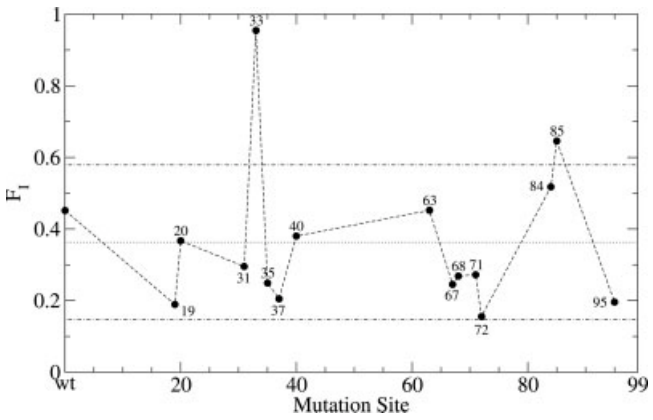

Fig. 7. The inhibitory folding factor F_I(i) associated with the peptide 83–93 acting on the HIV-1-PR wild type sequence carrying a (n-c) mutation on site *i*. The average value $\langle F_I \rangle = 0.33$ is shown with a horizontal dotted line, while the values $\langle F_I \rangle \pm \sigma$ ($\sigma = 0.21$) are displayed with dashed-dotted lines (see caption of Table I).

TABLE I. Inhibitory Folding Factor $F_I(i) = 1 - \Delta p(i)$ Plotted as a Function of the HIV-1-PR Site Number That Has Undergone a (Nonconservative) Mutation Associated With Peptide 83–93

Mutation site	$\Delta p(i) = p_N^0(i) - p_N^{83-93}(i)$	$F_I(i) = 1 - \Delta p(i)$
wt	0.55	0.45
19	0.81	0.19
20	0.63	0.37
31	0.71	0.29
33	0.05	0.95
35	0.75	0.25
37	0.80	0.20
40	0.62	0.38
63	0.55	0.45
67	0.76	0.24
68	0.73	0.27
71	0.73	0.27
72	0.85	0.15
84	0.48	0.52
85	0.35	0.65
95	0.81	0.19
Average		0.36
Standard deviation		0.21

The quantity Δp is given by the difference $\Delta p(i) = p_N^0(i) - p_N^{83-93}(i)$, where $p_N^{83-93}(i)$ is the probability that the HIV-1-PR carrying a (n-c) mutation populates, in the presence of three peptides (83–93), the *N* state, while $p_N^0(i)$ is the corresponding probability of the protein by itself. Also given are the average value of $F_I(i)$ as well as the standard deviation.

hand, mutations on site 85 affects essentially the folding and stability of the protease ($p_N^0(85) = 0.5$) but not the docking of the peptide and thus, in principle, neither its inhibitor ability. In fact, $F_I(85) = 0.65$ ($\approx F_I(\text{wt}) + \sigma$, see Fig. 7). This result is of course to be qualified by the fact that effects a) and b) stated above do not act independently of each other but are somewhat coupled.

Note that all of the observed mutations (spontaneous or FDA-drug induced) on sites 33 and 85 are conservative mutations (L33F,V,I and I85V). This in keeping with the fact that only structural mutations that prevent the protein from folding to its native state are not detected experimentally, as misfolded proteins are rapidly degraded in the cell. Moreover, only conservative, neutral mutations occur in the strands 23–33, 75–80, and 83–93 of the protease's monomer.

Extensive experimental assays to assess the inhibitory efficiency *in vitro* of peptide 83–93 have been performed.[13] Standard spectrophotometric measurements (at pH6, peptide concentration from 3 μM to 20 μM) indicate that this peptide displays an inhibition constant $k_I = 2.58 \pm 0.78$ μM. Moreover, circular dichroism performed under the same conditions shows that the peptide induces

a loss of beta structure in the protein from 30% (i.e., its native content) to 14%, compatibly with the inhibition mechanism by destabilization indicated by the simulations.

While these results nicely confirm model predictions, it is also important to emphasize some of the limitations of the model used to carry out the simulations. In particular, the fact that the internal native contacts of the isolated peptides are treated on pair to those of the corresponding segment of the target protein in its native state. Such approximation leads to isolated peptides which are likely to be, in general, too structured. This fact, together with the lack of sidechains and of the fact that the presence of the solvent is essentially ignored, does not prevent the identification of the target protein LES [26–35, 71–80, and 83–93, see Fig. 6(a)]. It may, however, lead to few nongenuine assignments (e.g., sequences 9–19 and 41–50). This fact will result in some amount of waste of time in carrying out *in vitro* control of the model predictions (e.g., we know that peptide 9–19 does not inhibit the folding of the protease, see Ref. 13), but otherwise immaterial for the purposes of the protocol.

CONCLUSIONS AND CAVEATS

It is possible, with the help of a modest number of numerical simulations and/or experimental assays, to individuate peptides (p-LES) that inhibit the folding (as well as the stability of the native conformation) of proteins and thus their biological function. This can be done in a fully automated fashion.

Of course this is not a ready-to-go solution to the complex problem of drug design. A peptidic lead may suffer from many limitations, like lack of solubility, internalization as well as being prone to hydrolyzation. Moreover, if there exist human homologs of the protein one wishes to inhibit, it is likely that they display similar LES,[22] and consequently p-LES may also interfere with them. On the other hand, the eventual availability of a consistent number of leads is likely to facilitate the selection of those peptides, or of their mimetic molecules, which display the most favorable properties in terms of interaction with human proteins, bioavailability, and so forth.

Furthermore, p-LES are expected to have an important advantage with respect to conventional drugs, that is to be perdurably effective. This is because they interact specifically with those regions of the protein containing those residues which play a central role in the folding and in the stability of the protein. This does not mean that the protein has no way out to evade the inhibitory action of the associated p-LES, since it can anyway produce coordinated multiple mutations able to substitute the stabilization core without destabilizing the protein. However, the time needed for the virus, or the organism expressing the protein to produce escape mutants to p-LES is expected to be much longer than that associated with the appearance of resistance in the case of conventional drugs, where single point mutations can do the job. Furthermore, one may use a cocktail of p-LES corresponding to representative LES of the different stabilization (folding) cores.

Last but not least, the results discussed above shed also light on the mechanism, which is at the basis of the folding of globular proteins, strongly supporting a hierarchical scenario.

REFERENCES

1. Lazo GS, Wipf P. Combinatorial chemistry and contemporary pharmacology. J Pharmcol Exp Ther 2000;293:705–709.
2. Hansch CA. Quantitative approach to biochemical structure-activity relationships. Acc Chem Res 1969;2:232–239.
3. Joseph-McCarthy D. Computational approaches to structure-based ligand design. Pharmcol Ther 1999;84:179–191.
4. Grzybowski BA, Ishchenko AV, Shimada J, Shakhnovich EI. From knowledge-based potentials to combinatorial design in silico. Acc Chem Res 2002;35:261–269.
5. Klebe G. Recent developments in structure-based drug design. J Mol Med 2000;78:269–281.
6. Lesk AM, Rose GD. Folding units in globular proteins. Proc Natl Acad Sci USA 1981;78:4304–4308.
7. Mirny L, Shakhnovich EI. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. J Mol Biol 1999;291:177–184.
8. Panchenko AR, Luthey-Schulten Z, Wolynes PG. Foldons, protein structural modules, and exons. Proc Natl Acad Sci USA 1995;93:2008–2013.
9. Broglia RA, Tiana G. Hierarchy of events in the folding of model proteins. J Chem Phys 2001;114:7267–7273.
10. Broglia RA, Tiana G, Provasi D. Simple model of protein folding and of non-conventional drug design. J Phys Condens Matter 2004;16:111–144.
11. Broglia RA, Tiana G, Berera R. Resistance proof, folding-inhibitor drugs. J Chem Phys 2003;118:4754–4758.
12. Broglia RA, Tiana G, Sutto L, Provasi D, Simona F. Design of HIV-1-PR inhibitors which do not create resistance: blocking the folding of single monomers. Protein Sci 2005;14:2668–2679.
13. Broglia RA, Provasi D, Vasile F, Ottolina G, Longhi R, Tiana G. A folding inhibitor of the HIV-1 protease. Proteins 2006;62:928–933.
14. Tiana G, Broglia RA, Roman HE, Vigezzi E, Shakhnovich EI. Folding and misfolding of designed protein-like chains with mutations. J Chem Phys 1998;108:757–761.
15. Tiana G, Simona F, De Mori GMS, Broglia RA, Colombo G. Understanding the determinants of stability and folding of small globular proteins from their energetics. Protein Sci 2004; 13:113–117.
16. Wallqvist A, Smythers GW, Covell DG. A cooperative folding unit in HIV-1 protease. Implications for protein stability and occurrence of drug-induced mutation. Prot Eng 1998;11:999–1005.
17. Cecconi F, Micheletti C, Carloni P, Maritan A. Molecular dynamics studies on HIV-1 protease drug resistance and folding pathways. Proteins 2001;43:365–372.
18. Tiana G, Broglia RA. Statistical analysis of native contact formation in the folding of designed model proteins. J Chem Phys 2001;114:2503–2507.
19. Khare SD, Wilcox KC, Gong P, Dokholyan NV. Sequence and structural determinants of Cu, Zn superoxide dismutase aggregation. Proteins 2005;61:617–632.
20. Baldwin RL, Rose GD. Is protein folding hierarchic? TIBS 1990;24:26–83.
21. Tiana G, Dokholyan NV, Broglia RA, Shakhnovich EI. The evolution dynamics of model proteins. J Chem Phys 2004;121: 2381–2386.
22. Tiana G, Broglia RA, Shakhnovich EI. Hiking in the energy landscape in sequence space: a bumpy road to good folders. Proteins 2000;39:244–250.
23. Riddle D, Grantcharova SVP, Santiago JV, Alm E, Ruczinski I, Baker D. Experiment and theory highlight role of native state topology in SH3 folding. Nat Struct Biol 1999;6:1016–1024.
24. McCallister EL, Alm E, Baker D. Critical role of β-hairpin formation in protein G folding. Nat Struct Biol 2000;7:669–763.

25. Itzhaki LS, Otzen DE, Fersht AR. The structure of the transition state of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation–condensation mechanism for protein folding. J Mol Biol 1995;254:260–288.
26. Goodsell DS, Olson AJ. Soluble proteins—size, shape and function. Trends Biochem Sci 1983;18:65–68.
27. Richardson JS. The anatomy and taxonomy of protein structure. Adv Protein Chem 1981;34:167–339.
28. Janin J, Wodak SJ. Structural domains in proteins and their role in the dynamics of protein function. Prog Biophys Mol Biol 1983;42:21–78.
29. Sutto L, Tiana G, Broglia RA. Sequence of events in folding mechanism: beyond the Gō model. Protein Sci 2006;15:1638–1652.
30. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. J Chem Phys 1953;21:1987–1996.
31. Xie D, Gulnik S, Gustchina E, Yu B, Shao W, Qoroneh W, Nathan A, Erickson JW. Drug resistance mutations can affect dimer stability of HIV-1 protease ar neutral pH. Protein Sci 1999;8: 1702–1713.
32. Karplus M, Weaver DL. Protein folding dynamics. Nature 1976;260:404–406.
33. Shafer RW, Hu P, Patick AK, Craig C, Brendel V. Identification of biased amino acid substitution patterns in human immunodeficiency virus type 1 isolates from patients treated with protease inhibitors. J Virol 1999;73:6197–6202.
34. Clementi C, Nymeyer H, Onuchic JN. 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 2000;298:937–942.

APPENDIX

The calculations have been carried out making use of a simplified model of proteins, as the computational cost associated with the calculation of the equilibrium properties of a protein do not allow the use of realistic models. In the model, each amino acid is described as a spherical bead centered at the position of its C_α atom and linked by an inextensible bond into a flexible chain. Each bead interacts with the other amino acids through a two-body potential built in such a way that the crystallographic native conformation is the state of minimum energy (these type of models are referred to as Gō models). More precisely, residues i and j interact through a square-well potential of range 7.5 Å and depth B_{ij} if they are within this range also in the crystallographic native conformation. Otherwise, they repel each other on the same length range. The energies B_{ij} are obtained from the experimental changes in free energy $\Delta\Delta G_{U-N}(\mathrm{i})$ upon mutation of the ith residue, solving the system of equations $\Delta\Delta G_{U\text{-}N}(\mathrm{i}) = \Sigma_j\ B_{ij}$ (for more details, see Ref. 24). In the case of HIV-1-Protease, where experimental $\Delta\Delta G_{U\text{-}N}$ values are not available, use was made of B_{ij}-values obtained from the average of the nonbonded terms of the Gromacs force field, following the strategy discussed in Ref. 15. Of course the present model describes in an approximated way both the geometry of the protein and the interaction between the amino acids. In particular, the matrix elements B_{ij} are calculated in the native state and consequently are expected to describe this state suitably, getting worse as the system departs from it. On the other hand, the unfolded state is stabilized entropically, and one expects that the details of the interaction are less relevant in determining its properties than in the case of the native conformation. Less justified is the use of the same matrix elements B_{ij} to describe the interaction between a LES and a peptide displaying the same sequence as the complementary LES (see the section on inhibition of HIV-1-PR). On the other hand, although this kind of interaction takes place in an environment different from that in which B_{ij} has been determined, the hydrogen bonds, the electrostatic attraction, and the Van der Waals interaction (in the nonpolarizable approximation used in Gromacs) depend mainly on the kind of amino acids, and less on the environment. In contrast, the hydrophobic interaction is much more dependent on the environment. In fact, this represents the crudest of all the approximations used in the simulations.

In spite of these limitations as well as of their simplicity, Gō models have proven useful in describing important features of small proteins, such as their two-state phase transition, folding kinetics, features of the transition state, and effect of mutations (see e.g. Refs. 17, 25, 34).

In particular, these models can investigate the distribution of dRMSD (i.e., the root mean square deviation of the distances between each pair of monomers from the corresponding quantities in the native conformation) and of the similarity parameter q (i.e., the fraction of native contacts). At biological temperatures the p(q,dRMSD) distribution shows two separated peaks, corresponding to the native (N) and to the unfolded (U) state, respectively. The transition state is the saddle point between the native and the unfolded state. The folding temperature T_f is defined as the temperature for which the volumes of the two peaks are equal.

Studying the distribution of the (q,dRMSD) values, one can identify the transition state and consequently the states N and U. Operatively, N is defined as the set of conformations characterized by $q > 0.65$ and dRMSD $<$ 6.0 Å for Src-SH3, protein G and CI2, whereas the threshold values in the case of the HIV-1-Protease are $q > 0.70$ and dRMSD $<$10 Å.

The definition of the ratio $r^{i\text{-}j} = p_N^{i\text{-}j}/p_N^0$ is quite stable with respect to the threshold values introduced above. This is shown in Figure A1, where $r^{i\text{-}j}$ associated with the SH3 domain in presence of three p-LES 45–50 (solid curve) and in presence of three peptides 29–34 (dashed curve) is displayed as a function of the q-value threshold q_{th}. It is seen that, provided $0.6 < q_{th} < 0.8$, an equally accurate measure of the inhibitory ability of the peptides is obtained from the simulations. Similar results have been obtained by varying the threshold value of the dRMSD (data not shown).

In what follows we discuss the rationale which is at the basis of which sites of the target protein are to undergo mutation and of how to implement these mutations. Let us start with the second issue.

In the present model, a mutation is made operative by substituting the (attractive) interaction parameters of the i-th mutated residue with the (repulsive) parameters typical of non-native contacts. Of course, this prescription does not apply to neutral, conservative mutations, that is mutations which do not alter in any important way the physico-chemical nature of the residue, like in

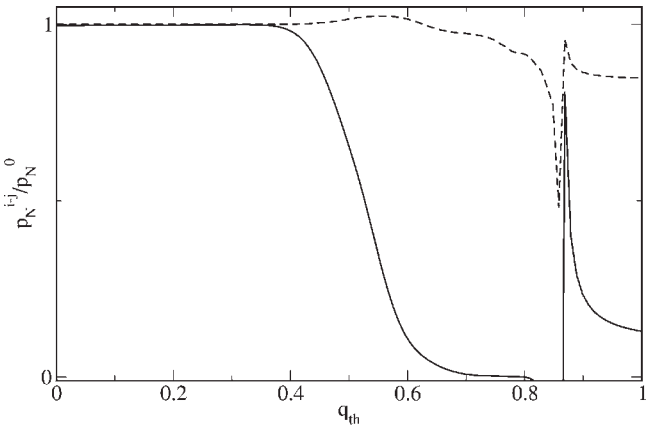

Fig. A1. The stability of the definition of inhibitory efficiency p_N^{i-j}/p_N^0 with respect to the threshold q_{th} used in the calculations (solid curve for the p-LES 45–50 and dashed curve for peptides 29–34). The plot shows that any value of q_{th} between 0.6 and 0.8 accounts properly for the inhibitory character of such peptides. For too large values of q_{th} numerical instabilities affect the definition.

the case of hydrophobic-hydrophobic mutations (e.g., in the case in which amino acid L is substituted by any of the amino acids F, V or I, see subsection on HIV-1-PR, Site no. 33). The way mutations are modeled is a crude approximation that has, nevertheless, proven successful in predicting ϕ-values and $\Delta\Delta G_{U-N}$ of proteins SH3, G, and CI2.[29] The mutated residues were chosen picking among the set of highly destabilizing ones (residues showing high $\Delta\Delta G_{U\text{-}N}$) and among the set of non destabilizing one (low or average $\Delta\Delta G_{U\text{-}N}$).

Let us now discuss a question which is tantamount to asking which mutations lead to escape mutants.

From the results of the simulations shown in Figures 2(a), 4(a), 5(a), and 6(a) we choose the peptides displaying the same sequence as stretches 37–42 (SH3), 41–46 (G), 25–30 (CI2), and 83–93 (HIV-1-PR) as good candidates to be a nonconventional inhibitor of the protein shown in parenthesis.

Under the pressure of the inhibitor (whose sequence we keep fixed throughout), the target protein may mutate. To which extent one would observe escape mutants? The answer is given in Figures 2(b), 4(b), 5(b), and 6(b), respectively. Let us concentrate on the HIV-1-PR. As seen from Fig. 6(a) and Table I, the inhibition factor associated with the peptide 83–93 is 0.45. Escape mutants would correspond to mutations leading an inhibition factor much larger than 0.45. From Table I it is seen that mutation on Site 33 fulfills this requirement, the associated inhibition factor being $F(33) = 0.95$. On the other hand the folding probability of the HIV-1-PR carrying a n-c mutation on Site 33 is very small, as small as the wt sequence in the presence of the peptide 83–93. Consequently, within the model, escape mutants are unlikely to be expressed. Summing up the role of mutations which destabilize the formation of LES complementary to p-LES (e.g. the LES (=23–33) of the HIV-1-PR complementary to the p-LES (=83–93)) is twofold: (1) avoid inhibition, (2) denaturate the target protein. In other words, nonconventional inhibitors are unlikely to create resistance, at least resistance induced by few uncoordinated point mutations.