GREGORY V. NIKIFOROVICH

Washington University
Department of Biochemistry and Molecular Biophysics
St. Louis, MO 63110, USA
e-mail: gregory@ccb.wustl.edu

Elements of non-random structure in the unfolded states of proteins: location and possible implications for protein folding mechanisms

Summary — This study introduces a simple computational procedure to search protein sequences for the segments with above average propensity to adopt non-random structures (which includes the native-like structure) in the unfolded state. The procedure consists of systematical conformational analysis of all overlapping hexapeptide segments in the protein sequence. The main aim of the computational approach is to determine the 3D structure most preferable for a given residue in the protein sequence, as determined by local interactions within the set of hexapeptides featuring the particular residue under consideration. Specifically, this study focuses on four types of "template" 3D structures that may be adopted by a hexapeptide, namely β-strand, α-helix, β-turn and the native-like structure of the folded state (assumed to be known). The study discusses also the possible importance of such segments for the different molecular mechanism of folding of the two prototypical proteins, namely the 65-residue barley chymotrypsin inhibitor 2 (CI2) and the 110-residue ribonuclease from *Bacillus amyloliquefaciens* (barnase). The computational results suggest that dynamic equilibrium in the unfolded state for the continuous fragment 6—27 in CI2 will likely prefer the native-like structure that may be preserved during folding. For barnase, on the contrary, dynamic equilibrium preferring the native-like structure most likely will occur in the unfolded state only at several small separate fragments, so the large non-native non-random segments of the unfolded state have to be restructured during folding.

Key words: protein folding, local nucleation centers, chymotrypsin inhibitor 2, barnase, conformational analysis, computer modelling.

Significant progress has been made in recent years in understanding possible molecular mechanisms of protein folding. Numerous experimental and theoretical studies led to two main models describing the way in which proteins fold. The first model assumes a hierarchical mechanism of folding that postulates formation of several local nucleation centers corresponding to small segments of native 3D structure at the very first stages of folding (*e.g.*, [1, 2]). These local nucleation centers are thought to be α-helical fragments, β-strands or β-turns [3, 4]. β-Turns may lead to formation of β-hairpins that, in turn, form β-sheets, α-helical fragments may expand to form α-helical bundles, *etc*. This mechanism predicts an accumulation of intermediates of the "molten globule" type that has been shown experimentally for many proteins (see, *e.g.*, [5]). The second model is so-called nucleation-condensation mechanism in which a transition from an unfolded state to the folded one occurs directly in one step without intermediates (the "two-state" model, see, *e.g.*, [6]). According to this mechanism, the rate of folding depends mainly on the general protein topology that, in turn, is determined by long-range interactions. Recently, Plaxco *et al.* estimated the general protein topology by the "contact order" parameter; the values of this parameter correlated with the rates of folding for 22 two-state proteins [7].

Both models do not contradict the commonly accepted assumption of polymer statistics that the general unfolded state of any polymer is basically the random coil, *i.e.*, the state of the polymer chain where only local, but not long-range interactions are important. However, both models agree that in proteins the unfolded state should not be regarded as a uniform random coil structure, but rather as a dynamic ensemble of many local 3D

structures including native-like ones [8]. The difference focuses on the possible role of these nucleation centers in the process of folding. To assess this, one needs to know how the segments of the non-random structure are distributed along the protein sequence in the unfolded state, and what kind of structure they adopt. Direct experimental studies of the unfolded/denatured states for several proteins have been performed by NMR spectroscopy using samples uniformly labeled by ^{13}C and/or ^{15}N nuclei [9—20]. Mostly, these studies locate non-random regions in the protein sequences based on differences between the chemical shift values observed for various residues and the "standard" values for the random coil structures [9, 10, 12, 13, 17—19]. Sometimes, it is possible also to observe the sequential and even long-range NOE's (nuclear Overhauser effects) in denatured proteins [9, 16, 17, 19]. Other important NMR characteristics are various types of relaxation rates whose values may be interpreted in terms of higher or lower mobility of segments in the protein structure [10—18]. Other experimental techniques, such as ESR spectroscopy or small angle X-ray scattering (SAXS) [10] have been applied to locate ordered segments in the denatured states of proteins as well.

These experimental techniques face, however, the very difficult problem of structural interpretation. Indeed, in most cases, experimental methods can distinguish between disordered and ordered local segments in the protein sequence, but identifying the possible 3D structures of the ordered segments is difficult. Due to the dynamical nature of equilibrium between the various structures co-existing within the denatured state, experimental identification of the elements of regular structures, as α-helices or β-strands or more complicated structural motifs is rather difficult. Also, since the loosely packed denatured state behaves in solution differently from the much more compact native state, there is, as a rule, insufficient long-range NOE's observed to restore the possible 3D structures by standard procedures.

A non-direct experimental insight into possible nucleation centers that are important for folding has been pioneered by using protein engineering that allow to estimate which position in the protein may be involved in the initial nucleation centers (see [6] and references therein). The advantage of this approach is that it may identify not only the local but also the long-range nucleation centers. This approach has been applied to many proteins; two of them, the 65-residue barley chymotrypsin inhibitor 2 (CI2) and the 110-residue ribonuclease from *Bacillus amyloliquefaciens* (barnase) have been studied especially extensively for a number of years (for the earlier and the most recent reports see [21] and [12] for CI2, and [22] and [18] for barnase). These two proteins are often regarded as typical examples of two different mechanisms of folding, namely the two-state CI2 and the multi-state (involving intermediates) barnase.

In view of the problems of structural interpretation of experimental data on the denatured state, it seems quite reasonable to complement experimental techniques with theoretical approaches for elucidating possible conformational states present in the unfolded/denatured state. In fact, molecular dynamics simulations to model the denatured states of both CI2 [12] and barnase [18] have been performed recently. However, the simulations were limited to only a few nanosecond trajectories, obviously not enough for exploring millisecond dynamic equilibrium in the denatured state.

This study suggests an independent computational approach to determine the segments in the protein sequence, which may by regarded as the local nucleation centers existing in dynamic equilibrium in the unfolded/denatured state. The study discusses also the possible importance of such segments for the molecular mechanism of folding of the two prototypical proteins, namely CI2 and barnase.

METHODS

Computational approach to find local nucleation centers in the protein sequence

The approach itself consists of systematical conformational analysis of all overlapping hexapeptide segments in the protein sequence (*i.e.*, a six-residue window is employed). The hexapeptide fragments were selected as the basic units for conformational studies for two reasons. First, the length of six residues is sufficient to distinguish between 3D structures of most of the expected nucleation centers (as β-strand, α-helix or β-turn) at the chosen criterion of geometrical similarity of an root mean square (rms) value equal/less than 2 Å (C_α atoms only). Second, complete conformational sampling of hexapeptides is readily attainable in terms of the available computational resources.

The main aim of the computational approach is to determine the 3D structure most preferable for a given residue in the protein sequence, as determined by local interactions within the set of hexapeptides featuring the particular residue under consideration. Specifically, this study focuses on four types of "template" 3D structures that may be adopted by a hexapeptide, namely β-strand, α-helix or β-turn and the native-like structure of the folded state (assumed to be known). It is logical to define an inherent propensity of hexapeptide to adopt the template structure as a ratio of the number of low-energy conformers geometrically similar to each template structure (say, n_i) and the total number of low-energy conformers (say, N_i), as $S_i = n_i/N_i$ (here the index i relates to the number of a given hexapeptide in sequence). Generally, such interpretation tacitly assumes that the computational procedure always finds the correct sets of low-energy conformers. In reality, this is not the case, mainly due to the systematic errors of modelling very diverse

interatomic interactions in peptides by the uniformly parameterized force fields (see, *e.g.*, [23]). To moderate this problem, some preconditions are introduced. First, too narrow sets of low-energy conformers (those with less than 20 conformers) are not considered as they may represent possible artifacts of the computational procedure. Second, the ratios normalized by the average ratio for all hexapeptides comprising the sequence in question are regarded as measures of relative propensities to adopt the template structure for each hexapeptide rather than the absolute ratios outlined above. In other words, instead of using the S_i values, relative propensities should be measured by the values of $s_i = S_i / (\Sigma s_i/M)$, where M is the number of hexapeptides in the protein sequence. Accordingly, if for any hexapeptide the relative propensity value s_i is larger than 1, it indicates that a given hexapeptide shows an above average propensity to adopt the template structure. Third, to evaluate the relative propensity σ_k for each residue (the k-th residue) to adopt low-energy conformations compatible to the template conformation, standard averaging of the propensity values within the six-residue window should be performed as $\sigma_k = \Sigma s_i/6$, where the index i denotes all six hexapeptides featuring the k-th residue. The above average propensity to adopt the low-energy conformation compatible with the native-like one, α-helix or β-strand for each residue is defined as $\sigma_k > 1$. For β-turns, obviously, the described uniform averaging cannot be performed; therefore, the above average propensity to adopt a β-turn conformation for an individual hexapeptide is defined as $s_i > 2$.

Conformational energy calculations

Energy calculations for all linear hexapeptides were performed employing the ECEPP/2 potential field [24, 25] assuming rigid valence geometry with planar *trans*-peptide bonds (including those in proline residues; there are no *cis*-prolines in CI2 and barnase). In all cases, hexapeptides were acetylated at the *N*-terminal and *N*-methylamidated at the *C*-terminal. Aliphatic and aromatic hydrogens were generally included in the unified atomic centers of CH_n type; H^{α}-atoms and amide hydrogens, as well as H^{δ}-atoms in prolines were described explicitly. All calculations were performed with the value of the dielectric constant $\varepsilon = 80$ (the macroscopic ε value for water) to mimic, to some extent, the effect of water. The starting points for energy calculations were all possible combinations of local energetic minima in the Ramachandran map selected to cover all significant conformational possibilities of peptide backbone (see also [23]). Specifically, we have selected the local energy minima with the dihedral angle φ, ψ values of -140°, 140°; -75°, 140°; -75°, 80°; -60°, -60° and 60°, 60° for all non-glycine and non-proline residues with an addition of the φ, ψ values of 140°, -140°; 75°, -140° and 75°, -80° for glycines; for prolines, the φ, ψ values of -75°, 140°; -75°, 80° and -75°, -60° were selected. The side chain dihedral angle values were optimized before energy minimization to achieve their most favorable spatial arrangements, employing an algorithm described previously [26]. In total, the numbers of different conformations of the peptide backbone under consideration for each hexapeptide were between 15 000 and 40 000. Low-energy conformers of the peptide backbone (those with relative energies $\Delta E = E - E_{min} \leq 6$ kcal/mol; see ref. [23] for justifying this criterion) were selected after energy minimization for each hexapeptide.

RESULTS

The computational procedure determined the set of low-energy conformers of the peptide backbones for all hexapeptide fragments that comprised the entire sequences of CI2 (60 hexapeptides) and barnase (105 hexapeptides). The numbers of low-energy conformers per hexapeptide fragment, N_i, varied from 3 (fragment 96—101 in barnase) to 665 (fragment 49—54 in barnase). Fragments with $N_i < 20$ were excluded from further consideration (9 fragments in CI2 and 18 fragments in barnase). For each hexapeptide, the number of low-energy conformations, n_i, geometrically similar to the 3D structures of the four template hexapeptide fragment has been determined. The template fragments used were the standard backbone conformations of β-strand, α-helix, β-turn (centered on the residues 3 and 4 of hexapeptide) and the corresponding hexapeptides in the X-ray structure of CI2 (the PDB entry 2CI2) or the NMR derived structure of barnase (the PDB entry 1BNR, model 1).

The resulting values of the relative propensities, σ_k, to be compatible with the native-like structure, β-strand and α-helix for the residues of CI2 and barnase are depicted in Figs. 1 and 2, respectively. Figs. 1 and 2 display also, which hexapeptide fragments in both proteins show the above average relative propensity to adopt the

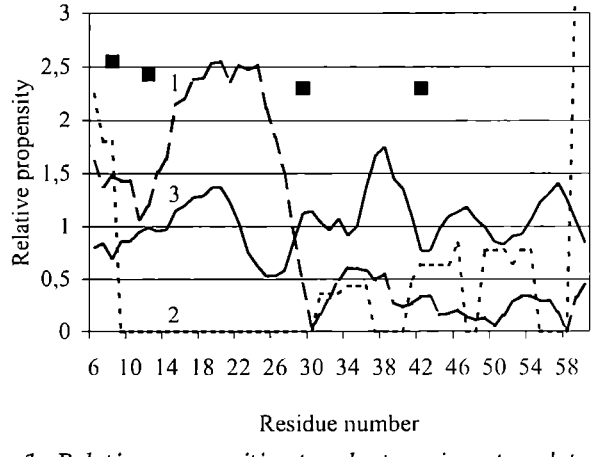

Fig. 1. Relative propensities to adopt various template 3D structures for residues of CI2; 1 — native-like, 2 — β-strand, 3 — α-helix; points denote β-turns

β-turn conformation, $s_i > 2$. For CI2 (Figure 1), there is only one continuous region 6—27 showing relative native-like propensities (curve 1) above average (see also Fig. 3, the left view; the numbering of residues in CI2 corresponds to the "short" version of CI2 observed by X-ray, where the eighteen *N*-terminal residues are absent). Two small regions, 6—8 and 59—60, show above average relative propensities to adopt the β-strand conformation (curve 2), and several fragments, 15—22, 29—31, 33, 41, 36—41, 45—48 and 54—59, display above average relative propensities to adopt the α-helix conformation (curve 3). There are also four possible regions of β-turns (points) centered at the positions 8—9, 12—13, 29—30 and 42—43. The above computational results are in good agreement with the direct NMR measurements of the chemical shift values for ^{1}H, ^{13}C and ^{15}N nuclei in the denatured state of CI2, which showed marked deviations from the random coil values for residues 33—34 (^{13}C); 19—21 and 30—32 (^{15}N); and 36 (^{1}H) [12]. One can see that almost all these residues are located in the regions corresponding to the non-random structure (the native-like or α-helical) according to the results in Fig. 1 (see also Fig. 3, the right view). The results displayed in Fig. 1 agree also with the high-temperature molecular dynamics simulations performed to model CI2 folding, which suggested that the only native-like region in the denatured state of CI2 is the α-helical fragment 17—21 [12]. However, it remains unclear whether these simulations were extensive enough (up to 20 ns [12]) to achieve the completely unfolded state of CI2.

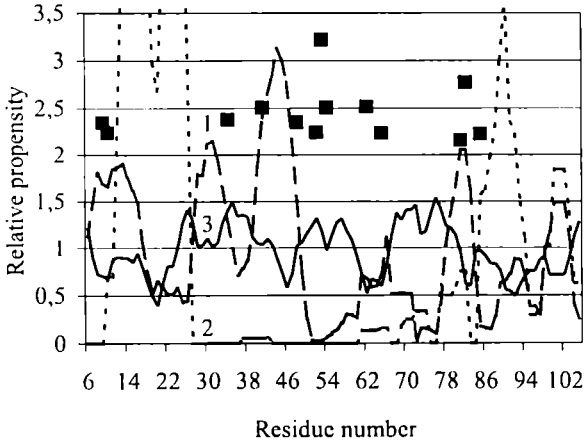

Fig 2. Relative propensities to adopt various template 3D structures for residues of barnase; 1 — native-like, 2 — β-strand, 3 — α-helix; points denote β-turns

For barnase (Figure 2), the regions with above average propensity to adopt the native-like structure (curve 1) include segments 6—17, 27—34, 39—48, 78—83, 99—103 and residue 66 (see also Fig. 4, the left view). The regions that show above average propensity to adopt the β-strand conformation (curve 2) are 12—26, 85—94 and 98—103, and those to adopt α-helical structure (curve 3) are 25—43, 48—53, 55—59, 67—80, 104—105 and residue 6. There are also a number of positions with above average propensity to center β-turns (points), namely, residues 9—11, 34—35, 41—42, 48—49, 52—55, 62—63, 65—66, 81—83 and 85—86. Direct NMR measurements of the denatured state of barnase have been performed several times under various conditions and by various techniques. The experimental NMR data obtained by measuring deviations of the chemical shift values from the random coil values for backbone and side chains, as well as the results of measurements of various types of relaxation times have been summarized in the recent publication [18]. For some residues, the data obtained by different techniques contradict each other; there are, however, 17 residues that seem to show somewhat ordered structure according to measurements by at least two independent techniques, namely residues 13—16, 24, 27, 30, 34, 78, 85, 88—89, 92, 94—96, and 101 (according to the data summarized in Fig. 15 of [18]). Almost all of them, except residues 95 and 96, are located in the regions predicted as showing the above average propensities to adopt either the native-like structure (residues 13—16, 27, 30, 34, 78 and 101) or the α-helical structure (residues 27, 30, 34 and 78) or the β-strand (residues 13—16, 24, 85, 88—89, 92, 94 and 101); see also Fig. 4, the right view. There is significant overlapping of the fragments predicted to adopt different types of template structures (see discussion below), but all of them are ordered, so whatever template prevails in dynamical equilibrium that exists in the denatured state of barnase; experimentally measured NMR parameters of the above residues will differ from those of random coil. Therefore, one can conclude that the independent computational results obtained for barnase in this study are also in good agreement with the experimental data. The high-temperature molecular dynamics simulations performed for barnase for 4 ns suggested that the residual native-like structure encompasses the α-helical fragments 9—17 and 28—43, as well as few separate residues in the third α-helix and in all the β-turns of barnase (for more details see Fig. 7 in [18]). Despite some differences, these calculations do not contradict the results of this study, since the lengths of molecular dynamics trajectories were too short to reach the fully unfolded state of barnase.

DISCUSSION

The analysis of local nucleation centers in the unfolded state of CI2 and barnase clearly showed that the starting state of folding is not random for each protein studied despite the difference in their folding behavior. The main difference between the unfolded states of the two proteins is the distribution of elements of the non-random structure along the protein sequences. Residues that adopt specific template conformations in the un-

folded state are in dynamic equilibrium with random coil structure and with other templates. The equilibrium itself is not reflected in computational estimations, so when regions of the higher than average relative propensities to adopt different template structures overlap in the protein sequence, the computational results do not point out which template is the most "preferred" one for each particular residue in the overlapping fragments. It is reasonable, however, to assume that if the higher than average relative propensities to adopt the native-like structure at the particular segment do not contradict the higher than average propensities estimated for other non-random templates, *i.e.*, β-strand, α-helix or β-turn at the same segment, dynamic equilibrium at this segment will be much more likely shifted to the native-like conformation in the unfolded state and may remain basically the same during folding. On the other hand, if the relative propensities to adopt the native-like structure and some of non-native-like templates contradict each other at the same segment, one can expect that dynamic equilibrium existing for this segment in the unfolded state will be dramatically changed during folding.

To illustrate this point, Figures 3 and 4 display two views of 3D structure of CI2 and barnase, respectively, which are the native structure (left) and representation of the unfolded structure (right). The views are color-coded according to the above average propensities for various residues to adopt the four template structures. In both views, green color marks segments considered in this study as native-like; residues within the native-like segments that may exist in dynamic equilibrium with other non-random structures contradictive to the native-like structure are showed in orange in the right view. The non-native regular structures are colored in the right view as blue for β-strands, red for α-helices, and magenta for β-turns.

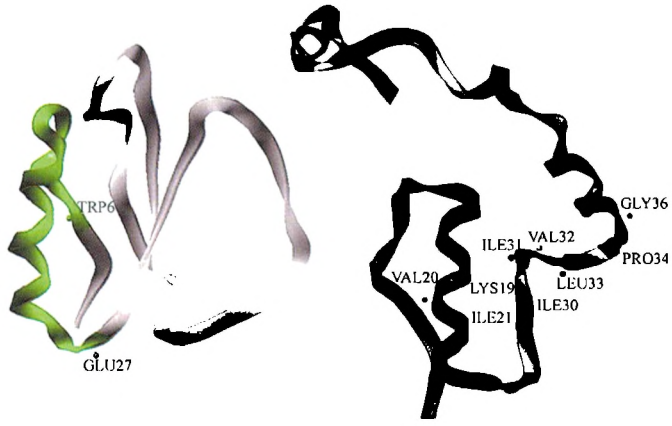

Fig. 3. Two 3D views of CI2. The left view corresponds to the native structure, the right view represents the unfolded state. Color codes are explained in the text. The end residues of the native-like segments are labeled at the left view. Residues with non-random local structure according to NMR data are labeled at the right view

Figure 3 clearly shows that computational estimates of high propensities to adopt the native-like structure and other non-random templates are distributed along the protein sequence in the unfolded state of CI2 without noticeable contradiction. Indeed, segment 15—22 is predicted as the segment with higher than average propensity to adopt both native-like structure and α-helix. The β-turns expected at residues 8—9 and 12—13 also do not contradict the expected native-like structure. Therefore, dynamic equilibrium in the unfolded state of the con-

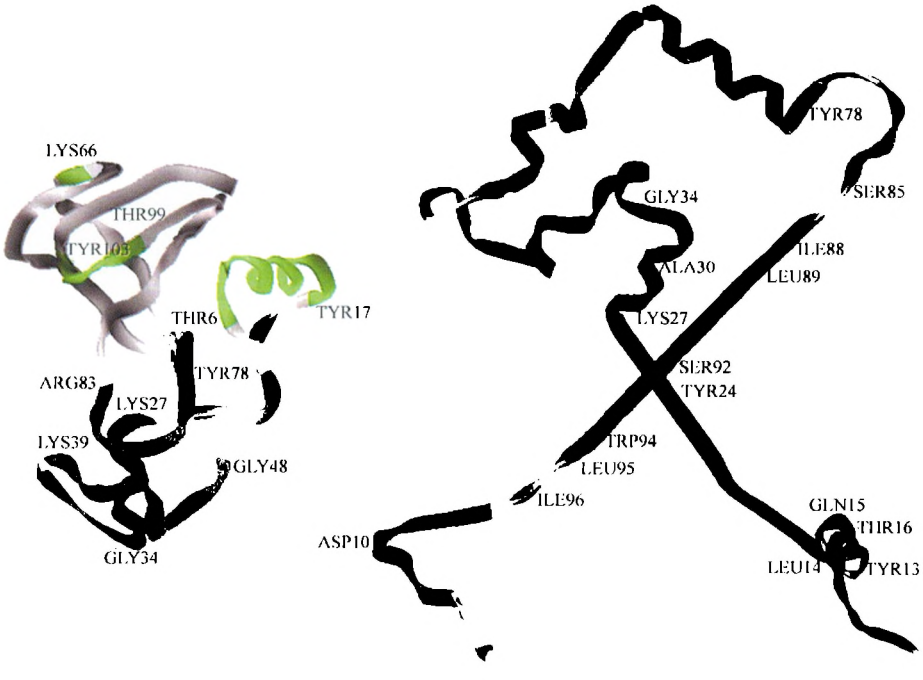

Fig. 4. Two 3D views of barnase. The left view corresponds to the native structure, the right view represents the unfolded state. Color codes are explained in the text. The end residues of the native-like segments are labeled at the left view. Residues with non-random local structure according to NMR data are labeled at the right view

tinuous fragment 6—27 in CI2 will likely prefer the native-like structure that may be preserved during folding.

For barnase (Figure 4), on the contrary, distributions of computational estimates of high propensities to adopt the native-like structure or the other non-random templates along the sequence contradict each other in several cases. Most obviously, the native-like structure at fragment 6—17, which is basically α-helix, may be distorted by dynamic equilibrium with the overlapping β-strand 12—17 and β-turns 9—10 and/or 10—11. Also, the native-like structures at fragments 78—83 and 99—103 compete with the overlapping α-helical structure at fragment 78—80 and with the β-strand structure at fragment 99—103, respectively. Therefore, dynamic equilibrium preferring the native-like structure most likely will occur in the unfolded state of barnase only at several small separate fragments, as 7—11, 27—34, 39—48, 81—83 and residue 66 (Fig. 4, the left view).

In other words, the unfolded state of CI2 may show a single continuous 22-membered fragment of the native-like structure, whereas the unfolded state of barnase may only show several separate small fragments of the native-like structure. One can hypothesize that this difference will, to some extent, account for different molecular mechanisms of folding of the two-state CI2 and the multi-state barnase. It seems natural that the sizable segment of the native-like structure in the unfolded state of CI2 may act as a scaffold stabilizing the entire set of long-range interactions that emerge during folding. As may be suggested from Fig. 3 (the right view), further structural rearrangements during folding of CI2 include transitions from α-helix to β-strand, from α-helix to the unordered structure, and some other transitions that are mostly from random coil to β-strands. There are no indications that such transitions cannot occur simultaneously in one step, since the segments of the non-random structures are small and do not overlap.

The unfolded state of barnase, on the contrary, does not show a single sizable native-like scaffold that can significantly influence stabilization of the entire globule. Also, the non-random segments of the unfolded state, which are to be restructured during folding, are rather large (see Fig. 4, the right view), as the β-strand 12—26 (transition to α-helix 12—17 and the unordered segment 18—26), or the mainly α-helical segment 48—59 (transition to β-strand 54—56 and the unordered segments). Obviously, the larger the non-native non-random segments are in the unfolded structure, the less likely is to restructure them in one direct step without intermediates.

CONCLUSIONS

This study introduces a simple computational procedure to search protein sequences for the segments with above average propensity to adopt non-random structures (which includes the native-like structure) in the unfolded state. These segments exist in dynamic equilibrium of random and non-random structures. The results of the computational procedure for two proteins, CI2 that folds the direct two-step way and barnase that folds by the multi-step mechanism, are in good agreement with direct experimental NMR measurements performed for the denatured states of CI2 and barnase. The results of this study suggest also that the difference in the mechanisms of folding of the two proteins may reflect the fact that the propensities of the same segments to adopt different types of the non-random structure do not contradict each other in the case of CI2, and are markedly contradictive in the case of barnase.

ACKNOWLEDGMENTS

The author is grateful to Dr. Garland R. Marshall for reading the manuscript and for critical comments, and to Dr. Carl Frieden for drawing his attention to the problem and for many fruitful discussions. This work has been supported by the NIH grant GM 53630.

REFERENCES

1. Ptitsyn O. B.: *Dokl. Acad. Nauk. SSR* 1973, **210**, 1213.
2. Wetlaufer D. B.: *Proc. Natl. Acad. Sci. USA* 1973, **70**, 697.
3. Baldwin R. L., Rose G. D.: *Trends Biochem. Sci.* 1999, **24**, 26.
4. Baldwin R. L., Rose G. D.: *Trends Biochem. Sci.* 1999, **24**, 77.
5. Ptitsyn O. B.: *Current Opin. Struct. Biol.* 1995, **5**, 74.
6. Fersht A. R.: *Proc. Natl. Acad. Sci. USA* 1995, **92**, 10 869.
7. Plaxco K. W., Simons K. T., Baker D.: *J. Mol. Biol.* 1998, **277**, 985.
8. Plaxco K. W., Gross M.: *Nature Struct. Biol.* 2001, **8**, 659.
9. Bai Y., Chung J., Dyson H. J., Wright P. E.: *Prot. Sci.* 2001, **10**, 1056.
10. Garcia P., Serrano L., Durand D., Rico M., Bruix M.: *Prot. Sci.* 2001, **10**, 1100.
11. Hodsdon M. E., Frieden C.: *Biochemistry* 2001, **40**, 732.
12. Kazmirski S. L., Wong K.-B., Freund S. M. V., Tan Y.-J., Fersht A. R., Daggert V.: *Proc. Natl. Acad. Sci. USA* 2001, **98**, 4349.
13. Klein-Seetharaman J., Oikawa M., Grimshaw S. B., Wirmer J., Duchardt E., Ueda T., Imoto T., Smith L. J., Dobson C. M., Schwalbe H.: *Science* 2002, **295**, 1719.
14. Kortemme T., Kelly M. J. S., Kay L. E., Forman-Kay J., Serrano L.: *J. Mol. Biol.* 2000, **297**, 1217.
15. Logan T. M., Theriault Y., Fesik S. W.: *J. Mol. Biol.* 1994, **236**, 637.
16. Neri D., Billeter M., Wider G., Wuthrich K.: *Science* 1992, **257**, 1559.
17. Sari N., Alexander P., Bryan P. N., Orban J.: *Biochemistry* 2000, **39**, 965.
18. Wong K.-B., Clarke J., Bond C. J., Neira J. L., Freund S. M. V., Fersht A. R., Daggert V.: *J. Mol. Biol.* 2000, **296**, 1257.
19. Yao J., Chung J., Eliezer D., Wright P. E., Dyson H. J.: *Biochemistry* 2001, **40**, 3561.
20. Yi Q., Scalley-Kim M. L., Alm E. J., Baker D.: *J. Mol. Biol.* 2000, **299**, 1341.
21. Itzhaki L., Otzen D. E., Fersht A. R.: *J. Mol. Biol.* 1995, **254**, 260.
22. Matouschek A., Serrano L., Fersht A. R.: *J. Mol. Biol.* 1992, **224**, 819.
23. Nikiforovich G. V.: *Int. J. Peptide Protein Res.* 1994, **44**, 513.
24. Dunfield L. G., Burgess A. W., Scheraga H. A.: *J. Phys. Chem.* 1978, **82**, 2609.
25. Nemethy G., Pottle M. S., Scheraga H. A.: *J. Phys. Chem.* 1983, **87**, 1883.
26. Nikiforovich G. V., Hruby V. J., Prakash O., Gehrig C. A.: *Biopolymers* 1991, **31**, 941.