Identification of cooperative folding units in a set of native proteins

A. WALLQVIST, G.W. SMYTHERS, AND D.G. COVELL

Frederick Cancer Research and Development Center, National Cancer Institute, Science Applications International Corporation,
Frederick, Maryland 21702

(RECEIVED October 31, 1996; ACCEPTED April 12, 1997)

Abstract

Cooperative unfolding penalties are calculated by statistically evaluating an ensemble of denatured states derived from native structures. The ensemble of denatured states is determined by dividing the native protein into short contiguous segments and defining all possible combinations of native, i.e., interacting, and non-native, i.e., non-interacting, segments. We use a novel knowledge-based scoring function, derived from a set of non-homologous proteins in the Protein Data Bank, to describe the interactions among residues. This procedure is used for the structural identification of cooperative folding cores for four globular proteins: bovine pancreatic trypsin inhibitor, horse heart cytochrome *c*, French bean plastocyanin, and staphylococcal nuclease. The theoretical folding units are shown to correspond to regions that exhibit enhanced stability against denaturation as determined from experimental hydrogen exchange protection factors. Using a sequence similarity score for related sequences, we show that, in addition to residues necessary for enzymatic function, those amino acids comprising structurally important folding cores are also preferentially conserved during evolution. This implies that the identified folding cores may be part of an array of fundamental structural folding units.

Keywords: autonomous folding units; empirical protein folding potential; structurally conserved amino acids; threading

Determining how the three-dimensional structure of proteins is specified requires a detailed understanding of numerous complicated types of interactions. Extensive experimental and theoretical studies have revealed that a complete description of a protein structure involves many levels of complexity, from secondary structure propensities for a few connected amino acids, through domains of hundreds of amino acids, to subunit assemblies involving several different proteins (Creighton, 1992). Our purpose here is to develop a computational method useful for identifying the existence of stable structural components of a protein and rank ordering their stability. These structural elements have sometimes been referred to as folding domains, foldons, or folding cores (Jaenicke, 1991). Though lacking definition, the operational definition of these units is the three-dimensional structure of the first observable native-like structure into which a protein folds from a denatured state. Although we will focus on native-like folding pathways, nature being abundantly diverse, it is certain that not all initially folded units will correspond to native-like fragments.

Experimentalists have broadly exploited the marginal stability of proteins to explore the roles of specific interactions that stabilize folding intermediates. Protein denaturation can be induced in many different ways, e.g., through thermal degradation, cold-denaturation, pH changes, or through addition of a denaturant such as urea or guanidinium hydrochloride (Kauzmann, 1959). Although an exact description of the denatured states cannot always be specified (Shortle, 1995), any condition where the physical and chemical properties are different from the designated native state is called a non-native state. Since a folded protein conformation is thought to be governed by a thermodynamic equilibrium, there is a finite probability that even under native conditions some molecules exist in non-native states (Anfinsen, 1973). These special naturally occurring non-native states are the equilibrium unfolding states.

An accumulating body of experimental evidence identifies these folding units for a wide variety of proteins. Although we still cannot generalize from this small but important data set, it is obvious that many different context dependent strategies are involved in the folding process. Hydrophobic patches as well as hydrogen-bonding patterns are included among the interactions that define these folding units. That sequence heterogeneity may be high implies that minimal models, such as simplified two- or three-state lattice models, will not help in analyzing experimentally observed folding units in any detail. The available experi-

Reprint requests to: Anders Wallqvist or D.G. Covell, Frederick Cancer Research and Development Center, National Cancer Institute, Science Applications International Corporation, Frederick, Maryland 21702; e-mail: wallqvist@ncicrf.gov.

mental data support the notion that not all folding pathways are realized as a protein folds under native conditions. Instead, distinct, and in some cases well characterized, intermediates have been detected during the folding event. These intermediates reduce the astronomical number of possible conformations of the polypeptide chain to a manageable set, directed towards the native state. This scenario involves clearly defined structural entities folding in a hierarchical fashion. In nature this would represent an ideal folding event, as alternative mechanisms have evolved in the cell to aid the folding process (Jaenicke, 1991). Apart from pre- and post-sequences, in which extra stretches of amino acids are attached to the primary sequence in order to facilitate folding or transport, misfolded proteins can be corrected. The presence of chaperonins and disulfide and proline isomerase enzymes indicates that misfolded proteins occur frequently enough for nature to devise general strategies to overcome such dead ends in the folding path.

However, since the number of all possible amino acid combinations for a given chain length is extremely high, and the number of currently available protein structures—or even sequences—is relatively low, we cannot rule out the alternative that each protein's, or more correctly each protein family's, folding pathway may actually be unique. If so, a general code for protein folding may not exist.

The identification of cooperative folding units may not yield a unique answer for the folding problem. Still, such a classification is useful for understanding higher levels of sequence structure relationships of the genetic material. Several experimental methods are available to investigate cooperative protein folding intermediates. One of the most direct methods studies the equilibrium folding pathway by analysis of amide hydrogen exchange (Englander & Mayne, 1992). This idea, by Linderström-Lang (1955; Hvidt & Nielsen, 1966), assumes that the native state is in thermodynamic equilibrium with all the possible existing denatured states at the specified external conditions. A freely moving peptide chain without any secondary or tertiary structure exchanges its amide hydrogens with the surrounding solvent at some intrinsic rate, which is dependent only upon the adjoining amino acids, temperature, pH, and salt concentration (Molday et al., 1972; Bai et al., 1993). As the protein populates a part of phase space where some initially buried amide hydrogens come in contact with the solvent, these exposed hydrogens can now exchange. Schematically, for a buried amide hydrogen belonging to residue i ,

where $k_{ch,i}$ is the intrinsic rate and $K_{op,i} = K_{op,i}/k_{cl,i}$ describes the equilibrium denaturation residue i in the protein's structural opening and reclosing rates. If the structure is stable ($K_{op,i} < k_{cl,i}$) and the intrinsic exchange rate is slower than the rate at which the protein is reclosing ($k_{ch,i} < k_{cl,i}$), the observed rates $k_{ob,i}$ simplify to

$$k_{ob,i} = K_{op,i} k_{ch,i} \quad (2)$$

Since the intrinsic rates $k_{ch,i}$ are known, the equilibrium constant and hence the free energy associated with the opening and closing of the protein can be directly measured. Thus, protons measured as exchanging quickly are interpreted as being part of the structurally least protected regions of the protein, whereas we assume slowly

exchanging protons to be part of the molecule that is most stable with respect to unfolding. Many other experimental observables, such as CD spectra of folding intermediates, truncated proteins exhibiting native-like structures (Dyson & Wright, 1993; Wu et al., 1994), and fluorescence spectroscopy, indicate that this is the correct interpretation of many hydrogen exchange experiments. Alternative exchange mechanisms and interpretations have also been proposed (Finucane & Jardetzky, 1995; Perrett et al., 1995).

The theoretical calculation of unfolding penalties for an amino acid in a native protein must necessarily introduce several approximations. The most crucial is the description of the denatured states accessible to the protein under native conditions. In this work we use the approach of Hilser and Freire (1996, 1997), in which segments of the native proteins are considered to be in their respective native or non-native configuration. A non-native segment of the polypeptide chain is simply a stretch of the protein that does not interact with other segments. For a given division of a protein sequence into segments, an ensemble of denatured states is generated by including all possible pairwise combinations of these segments. Thus, the probability that a residue i is in its native state is calculated by summing the probabilities of all segments which contain residues i in their designated native state. The probability of a residue being in a non-native state is calculated analogously. From the ratio of these probabilities we then construct a free-energy-like quantity corresponding to the unfolding penalty from the native state to the ensemble of denatured states. This description is straightforward but requires a knowledge of the interactions between residues in the native state. We have characterized these interactions on the basis of observed distance and angle dependencies of high quality protein structures in the Protein Data Bank. This knowledge-based scoring scheme is used to assess the probabilities of interresidue contacts and conformational bias among the native states in the ensemble. These procedures are detailed further in Methods.

The initial set of studied proteins contains four small monomeric proteins ranging in size from 58 to 149 amino acids: bovine pancreatic trypsin inhibitor, cytochrome *c*, plastocyanin, and staphylococcal nuclease. The experimental data on folding and folding intermediates for these selected proteins are vast. These data help us corroborate the correct identification of the folding units with the theoretical model calculation. With this model, the structural variation of folding units is striking and thus interactions between regions with and without well-defined secondary structures characterize the most protected region of the native protein. This effect does not always correlate with surface burial nor with the energetics of single amino acid replacements. It is the ensemble of denatured states that cooperatively stabilizes the folding cores. Thus folding units will be strongly context-dependent.

Correlations between folding units and sequence conservation within a protein family may suggest that, apart from functionally conserved amino acids in a catalytic site, folding units are essential elements of proteins that have been fine-tuned and preserved by evolutionary processes.

Results and discussion

Bovine pancreatic trypsin inhibitor

The bovine pancreatic trypsin inhibitor (BPTI) is a thoroughly studied, 58-residue protein whose function is to block peptide cleavage by trypsin (Lodish et al., 1995). Inserting a conserved

lysine side chain on BPTI into the active site of trypsin effectively blocks the cleavage mechanism (Vincent & Lazdunski, 1972; Vincent et al., 1974). BPTI is a member of the protease-inhibitor enzyme family and, as such, shares its general structure with other similar inhibitors. The structure of BPTI derived from the Wlodawer et al. data (Wlodawer et al., 1984) shows that it is a monomeric protein with secondary structure elements of a 3_{10} -helix between residues 2 and 7, two twisted and antiparallel β -sheets between residues 18–24 and 29–35 (residue 45 also has $\phi - \gamma$ angles compatible with a β -sheet), and an α -helical region from residue 47 to 56. The structure has three characteristic disulfide bonds between residues 5 and 55, 14 and 38, and 31 and 50. Mutants lacking one disulfide bond $[5-55]_{SS}$ can still fold and achieve essentially the same structure as the native protein (Staley & Kim, 1992). Although the general fold itself may not be sensitive to restricted mutations, the stability of the protein (Makhatadze et al., 1993) can be drastically affected (i.e., reduced by over 75%) by selecting single point mutations at critical locations in the sequence (Goldenberg & Zhang, 1993; Kim et al., 1993b).

Unfolding penalties ΔW_i^F , calculated using the procedure outlined in Methods, are displayed in Figure 1. Segments with the largest unfolding penalties appear as the highest values plotted in Figure 1. The correlation coefficient r , for the experimental data and the calculated penalties, is 0.67, which, for the number of data points ($n_f = 31$), is significant at less than the 1% level ($p < 0.01$). The most protected residue is predicted to be F22, located in the middle of the first β -sheet, together with its close neighbors. It has long been known that the slowest exchanging amide protons are located in this region (Richarz et al., 1979; Wagner et al., 1984). The second- and third-most protected areas, as predicted from our model, are centered around residue F45 and in the adjacent second β -sheet. Residues not strongly protected against unfolding are in the N and C terminus as well as in the loop regions between residues 11–16 and 36–41. The protection afforded by cysteine

bonds $[5-55]_{SS}$ and $[30-51]_{SS}$ have similar unfolding penalties, whereas the $[14-38]_{SS}$ bond in the outer loop is considerably less stable against unfolding. This decreased stability is consistent with the experimental observation that the $[14-38]_{SS}$ bond is the last to occur in the folding pathway of BPTI (Kosen et al., 1981; Goldenberg, 1992; Weissman & Kim, 1992). The peaks in Figure 1 define a small core that is relatively protected against unfolding. That the disulfide bonds do not contribute strongly to this protection suggests that they may not be essential to define the native structure of BPTI. BPTI's native fold is instead the property of the entire sequence with special emphasis on the core unit identified by the peaks in Figure 1. Even taken individually and out of context, parts of these protected regions have been shown to be native-like. A truncated sequence comprising the last 12 residues of the native sequence shows an α -helical structure in solution (Goodman & Kim, 1989). The synthetic $P_\alpha P_\beta$ construct of Oas and Kim encompasses residues 20–33 and 43–58 of the native sequence (i.e., essentially including the regions of highest unfolding penalties shown in Fig. 1) has been shown to acquire a native-like conformation (Oas & Kim, 1988).

Figure 1 also displays the experimentally measured protection factors against unfolding as determined by Kim et al. (1993a). Following the procedures outlined by Englander and co-workers (Molday et al., 1972; Bai et al., 1993), we have converted the exchange data to free energies of unfolding. The overall shape of the calculated curve agrees well with the experimental quantities and is statistically significant. The local enhanced stabilities against unfolding has a maximum at residues L6–E7 and the decreased stabilities in the 11–16 loop region are centered around C14. The peak region at residues 21–23 is closely matched, as is the following minimum at A27 in the turn between β -sheets. The β -sheet of residues 29–35 exhibit enhanced stabilities consistent with our observations. A local minimum of protection is predicted to occur around C38, followed by a region of enhanced stability in the coil region around F45. The helix region of residues 50–55 is not well resolved in the calculations but does correspond to the experimentally observed local maximum in unfolding free energies.

A comparison of the predicted unfolding penalties with the conserved residues in the trypsin inhibitor family is plotted in Figure 2. This figure demonstrates the strong correlation between sequence conservation within this family of proteins and *structurally* important regions spanning residues 20–25 and 42–46 that contain the core folding unit of BPTI. Calculating a correlation coefficient between the sequence conservation score and the unfolding penalties, excluding the residues important for function, yields a statistically significant observation ($r = 0.65$, $n_f = 40$, $p < 0.01$). We also find to a lesser degree a correspondence between the protected and conserved residues at position 5–7 and a relative lack of protection and conservation for residues 26–28. It is noteworthy that residues of the inhibitor that form the contact surface with trypsin and contain the functionally important K15 residue are preferentially conserved but have low unfolding penalties. Thus, the BPTI molecule incorporates both features that are *structurally* as well as *functionally* critical to its stability and biological activity. The portion of BPTI responsible for function does not appear to have built-in structural features to withstand unfolding.

Mutation experiments with BPTI have shown that several single amino acid mutants fold and crystallize in essentially the same gross configurations as the native sequence (Danishefsky et al., 1993). All these engineered, non-natural mutations are in structurally important regions of the sequence. Their thermodynamic sta-

Fig. 1. Unfolding penalties of BPTI for each residue in the sequence calculated with a segment window N_w of 3. Thus a total of $10^{6.5}$ configurations were used to form the averages of ΔW_i^F . Protection factors measured from hydrogen exchange experiments at 30 °C and pH 3.5 (Kim et al., 1993a) are also indicated on this graph. The unfolding penalties were calculated from the 5pti data set of the Protein Data Bank. The correlation between the two data sets is statistically significant ($r = 0.67$, $n_f = 31$, $p < 0.01$). The unfolding penalties are scaled so as to visually compare the two data sets.

Fig. 2. Sequence similarities of trypsin inhibitor-like sequences. The cross-line indicates a relative measure of amino acid conservation. Data for residues above this line show an enhanced conservation score; data below this line are more prone to mutations among naturally occurring sequences. The functionally important residues for BPTI are indicated as boxes in the lower half of the graph. These residues contain the crucial Lys 15 residue and the two coil regions forming the contact surface between the inhibitor and the trypsin enzyme. The unfolding penalties of Figure 1 are superimposed over the graph. Note the correlation between conserved amino acid residues and the structurally and functionally important residues. The observation that structurally important amino acids (as defined by a higher unfolding penalty) and low sequence variability correlate is statistically significant ($r = 0.65$, $n_f = 40$, $p < 0.01$).

bility is sharply reduced against the denatured state (Kim et al., 1993b) as has been observed for other single mutations (Goldenberg & Zhang, 1993). Some mutations that are structurally less important, such as S47A, are less drastic in reducing the thermodynamic stability of mutated BPTI (Goldenberg & Zhang, 1993). The calculated reduction of the raw score W in Equation 5 for these mutant structures shows a decrease of 15–30% from the wild-type score compared to the experimental free-energy change of 15–75%. Calculations of unfolding penalties for these mutants show that the overall shape of ΔW_i^F exhibits no gross changes from the native protection pattern. Since all these mutated structures fold into the same three dimensional conformation, one would not expect large changes in folding patterns.

Cytochrome *c*

The cytochromes *c* (cyt *c*) family of proteins is part of the electron transport process in both eukaryotes and prokaryotes. Typical sequence length of these proteins is 104–113 amino acids. These proteins act as intermediates in shuttling electrons from a cytochrome *c* reductase complex to a cytochrome *c* oxidase complex, which ultimately reduces oxygen to water (Dickerson & Timkovich, 1975). The functional group is a prosthetic porphyrin ring-structure, which is covalently linked to the protein. As part of the electron transport chain, cyt *c*'s function is to protect the electron from being exposed to reducing agents by localizing the excess electron in the heme group, which in turn is surrounded by a hydrophobic pocket formed by the protein itself.

The structure of cyt *c* is dictated largely by its function; i.e., the protein has evolved as a convenient scaffolding to surround and

protect the heme group. In contrast to the myoglobins, not all cyt *c*'s will fold correctly in the absence of its prosthetic group. Three amphiphilic helices—one at the amino terminal, one in the middle 60's helical region, and one at the carboxy terminal—form a structural unit with a hydrophobic pocket stabilizing the heme, and a hydrophilic exterior. Less well formed helices and loops further enclose the heme group, which is covalently linked to cys 14 and 17. The changes in side-chain conformations that occur upon oxidation and reduction of the heme are well documented (Takano & Dickerson, 1981a, 1981b); backbone changes, however, are minimal.

The calculated unfolding penalties of horse heart cyt *c* in the absence of the heme group are shown in Figure 3 together with the estimated free energies derived from two hydrogen exchange experiments, one at low pH in strong saline solution (Jeng & Englander, 1991) and one at a pD of 7 (Bai et al., 1995). The heme group deletion was affected by removing the appropriate PDB entries. The correlation between the measured folding energies at pD = 7 and the calculated penalties is statistically significant, $r = 0.71$, $n_f = 36$, $p < 0.01$. The unfolding penalties are color-coded directly onto the three dimensional structure in Figure 4. The carboxy terminal helix (residues 87–104) is the most protected environment of the structure. Together with the spatially close N-terminal helix (residues 7–12), these helices are predicted to form the most stable unit against unfolding and are found experimentally to be most protected against hydrogen exchange. The identification of this unit as being crucial in the stability of cyt *c* has also been made by Fisher and Taniuchi (1992), who considered various combinations of fragment complexes from different cytochromes *c*. Kuroda (1993) showed that the N-terminal helix had no secondary structure in solution by itself, in contrast to the C-terminal peptide which exhibited helical structure. However, connecting these two peptides with a disulfide bond enhanced the overall stability of

Fig. 3. Unfolding penalties ΔW_i^F of horse heart cyt *c* for each residue in the sequence. The calculation was performed with a window size of 6, allowing for $10^{6.2}$ configurations in the denatured-state ensemble. The coordinate data set from the Protein Data Bank was 1crc. Indicated experimental protection factors were measured from hydrogen exchange studies at 30 °C and pD 7 (Bai et al., 1995) and at 20 °C with 1.5 M NaCl and pD 2.2 (Jeng & Englander, 1991). The agreement between the theoretical unfolding penalties and the pD 7 data set is characterized by $r = 0.71$, $n_f = 36$, $p < 0.01$, and the correlation between the two experimental data sets is $r = 0.65$, $n_f = 35$, $p < 0.01$.

Fig. 4. The three-dimensional structure of cyt *c* (1lrc) color coded with respect to the unfolding penalties in Figure 3. The graphics program MOLSCRIPT (Kraulis, 1991) was used to generate this and the following drawings of schematic protein structures. The color blue corresponds to the folding core, green to intermediately protected regions, and red to relatively unprotected regions. The heme group is added for visual orientation purposes only and does not enter into the calculation of the unfolding penalties. The viewer's perspective of the protein is looking down upon the hydrophobic heme pocket and the surrounding residues. In this structure, most of the top loops form the rim of the hydrophobic heme pocket and are relatively unprotected compared to the three helices. The amino acid sequence and secondary structure assignment appear at the bottom of the figure. Thickened lines below the sequence indicate helices (no arrow head) and sheets (with arrow head).

both helices. Two unconnected segments comprising residues 1–38 and 87–104, i.e., including both terminal helices, have been shown to combine and form a flexible, yet tightly bound, complex with helical structure (Wu et al., 1993). In kinetic studies, an early intermediate of transient stability has been identified that involves the two terminal helices (Elöve et al., 1992). A second important structural region in cyt *c* involves the 60's helix, i.e., the residues 60–70 in Figures 3 and 4. This region is identified as important for cyt *c* stability as determined by the hydrogen exchange experiments of Englander and coworkers (Bai et al., 1995; Sosnick et al., 1996), and as a part of the core domain delineated by Fisher and Taniuchi (1992). Together with the third region of enhanced stability and high unfolding penalty, residues 32–39, these folding units define the core domain that is protected against unfolding. At the same time these folding units construct the deep hydrophobic pocket in which the heme group resides. In the absence of the heme group, the remaining residues do not show any strong protection.

The difference in the two experimental data sets in Figure 3 is ascribed to the altered protection between the native state at

pD = 7 and the molten globule character at low pH. Although the overall stability of the different folding units in the molecule has been reduced in the molten globule state, the relative strengths of unfolding penalties for each folding unit has not. Thus, the qualitative pattern of the cooperative protection afforded the cyt *c* structure is similar in the two different states, and we can deduce that the ensemble of denatured states is of the same nature in the two cases.

In Figure 5, we have superimposed the predicted unfolding penalties for two different coordinate sets of horse heart cyt *c* (Bushnell et al., 1990) over unfolding penalties for the yeast cyt *c* structure (Louie & Brayer, 1990). Slight differences are observed in the unfolding penalties between the two horse heart structures, whereas the yeast structure exhibits some quantitative differences in protection patterns, e.g., in the region between residues 40–50. The overall agreement between the experimental data and the unfolding penalties does, however, remain. In the same figure, we also include a graph of sequence similarities within the cyt *c* family to depict the evolutionary conserved amino acids. The correlation of the structurally important residues with preserved amino acids is statistically significant at the 1% level ($r = 0.52$, $n_f = 74$, $p < 0.01$). Residues forming part of the hydrophobic pocket that surrounds the heme group are indicated by arrows in the top portion of Figure 5. The structurally most important elements, the terminal helices, are conserved over the cytochromes *c*, whereas the 60's helix is less well maintained. The importance of the third structural domain of residues 32–39 also agrees with the low sequence variability in this region. The most important theme, however, is that amino acids surrounding the prosthetic group are most strongly conserved (Dickerson & Timkovich, 1975; Hampsey et al., 1986). These residues appear to be more important than the regions that

Fig. 5. Theoretical unfolding protection patterns for different crystal structures of cytochromes *c* from horse heart and yeast superimposed on a sequence similarity plot. The sequence correlations are indicated by bars. The statistical correlation between the unfolding penalties of cytochrome *c* (1lrc) and the preservation of amino acids for structurally important residues is 0.54, which is statistically significant for $n_f = 74$ and $p < 0.01$. The segment window resolution was set to $N_w = 6$. The Protein Data Bank call names for the crystals are 1lrc, 1hrc, and 1yc. The cross-line indicates a relative measure of conservation. Data for residues that are above this line show an enhanced conservation score; data below this line are more prone to mutations among naturally occurring sequences. Residues that line the hydrophobic heme pocket are indicated by the arrows in the graph.

determine the core folding element. The absence of correlation between sequence conservation and the protection factors in the 60's helix is mainly due to the variability of the helix composition between eukaryotes and prokaryotes (Dickerson & Timkovich, 1975). The helix propensities themselves for either of these residue combinations do in fact remain high (Schulz & Schirmer, 1979).

The cooperative nature of the cyt *c* folding domains is displayed in Figure 6, where unfolding penalties are shown for different combinations of chain segments. Three hypothetical proteins are formed from the coordinates of the native structures by only including residues 1–86 (I), 39–104 (II), and two disconnected snippets, 1–38 and 87–104 (III), in the calculation. The remaining segments of cyt *c* are excluded in this analysis. In construct (II), which lacks the amino terminal end, the protection afforded to the carboxy terminal helix has now vanished; likewise in construct (I), the N-terminal helix is relatively unprotected compared to the 60's helix. The conserved pattern of N- and C-terminal protection is restored only for the third sequence, which contains the requisite components of the core domain.

In the evolution of organisms that utilize electron transport proteins the overall function of cytochrome *c* has remained the same. Structural differences do occur, e.g., the fourth domain between residues 39 and 55 appears to have been inserted rather late in the evolution (Almassy & Dickerson, 1978; Dickerson, 1980; Fisher & Taniuchi, 1992). The unfolding penalties—adjusted for gaps in sequence—of the horse heart cyt *c* and the *Pseudomonas* cytochrome *c*₅₅₁ are shown in Figure 7. The unfolding penalties calculated for these two sequences display a remarkable agreement in protection pattern; the correlation coefficient *r* is 0.88, which for the number of data points, *n_f* = 83, is significant at the 1% level. This agreement suggests that cooperative folding patterns are equally well—or in some instances even better—conserved than sequences. The protection pattern of the residues surrounding the heme underscores the importance of the heme group in the cyt *c* function and the requirements for constructing the proper pocket for binding heme.

Fig. 6. Unfolding penalties for various theoretical cytochrome *c* constructs. Three clipped peptides of various length, employing the native configuration of full-length horse heart cyt *c*, show differences in protection pattern. We used a segment window resolution of *N_w* = 6 for the calculations.

Fig. 7. Conservation of the actual protection pattern for two distantly related cyt *c* proteins, horse heart and *Pseudomonas* *c*₅₅₁. The missing residues in *Pseudomonas* form gaps in the sequence alignment and are indicated by vertical lines. The correlation between the two data sets is high, *r* = 0.88, *n_f* = 83, *p* < 0.01. The Protein Data Bank call name for the *Pseudomonas* cyt *c* data set is 351c.

Plastocyanin

Plastocyanins constitute a class of copper-binding proteins that form part of the electron transport process in photosynthesis. Electrons are shuttled from photosystem I to photosystem II by the soluble plastocyanin protein, which carries excess electrons via a non-covalently bound Cu(I) atom (Lodish et al., 1995). The structure of several plastocyanins reveals a similar three-dimensional arrangement, essentially constituting two pleated β -sheets forming an oblong protein. The copper attachment site is situated at the loop connections between the two β -sheets (Guss & Freeman, 1983). Within the cell, plastocyanin molecules are initially synthesized with different pre- and pro-sequences needed for transport into the lumen. Some of these plastocyanins cannot support an apo-form at neutral pH. NMR studies suggest that the three-dimensional structure of plastocyanin is virtually the same in the crystal form as in solution, and that the apo-form, when stable, does not differ structurally from the copper-loaded state (Chazin & Wright, 1988; Moore et al., 1988, 1991).

In Figure 8 we have plotted the unfolding penalties ΔW_i^f for the apo-form of poplar (Guss & Freeman, 1983) and French bean plastocyanin (Chazin & Wright, 1988), together with the protected regions qualitatively determined by protein exchange experiments (Chazin & Wright, 1988). The calculated values for French bean plastocyanin are also color-coded in Figure 9 to show the spatial relationships between protected amino acids. The three data sets all designate the β -strands as the important units as opposed to the relatively unprotected, small helical segment at residues 52–55. Our calculation differentiates these β -strands into at least two possible folded core units: the two strands in the first turn encompassing residues 1–15 and the two antiparallel strands of residues 40–42 and 78–84. These regions exhibit a cooperative protection, which, with additional strands, form part of the two pleated β -sheets defining the native fold.

Sequence similarity analysis indicates that the initial four β -strands are conserved within the family of plastocyanins (data

Fig. 8. The unfolding penalties ΔW_i^F calculated in Equation 31 for plastocyanin from poplar leaves and French bean. The PDB call name for the coordinate data sets are 1plc and 9pcy, respectively. A segment window of $N_w = 5$ resulted in $10^{6.7}$ configurations considered in the denatured state ensemble. The experimental data of Chazin and Wright (1988) are available only as either protected or unprotected residues, and in this graph the protected residues are simply indicated by staples. In order to estimate the correlation between the theoretical and experimental folding domains, we assigned the unprotected regions a value of zero protection. The calculated correlation is then statistically significant at the 1% level, $r = 0.56$, $n_f = 99$, $p < 0.01$.

not shown). In addition, the region in close spatial proximity to the copper attachment site, especially residues 34–38 and 90–93, is also conserved. This is the coil-like region around the copper atom in Figure 9. Two regions that do not appear to be structurally

Fig. 9. The theoretical unfolding penalties are color-coded onto the three-dimensional structure of French bean plastocyanin. The copper atom is shown for illustrative purposes only and is not included in the calculation of the unfolding penalties. The color blue indicates the strongest protection against unfolding, green intermediate, and red the weakly protected regions against denaturation. The secondary structure assignment of the sequence is given at the bottom.

important, the short helix at residues 52–55 and the turn at 62–66, also show enhanced sequence preservation. The importance of these regions in terms of plastocyanin function is not clear to us but may identify attachment or recognition sites to other proteins in the photocenter complex (Gross, 1993).

Dyson et al. (1992) have studied fragments of plastocyanin in order to characterize the folding of the native protein. Unfortunately, the most interesting sections—encompassing residues 1–15 and those including a larger region of the 78–84 β -strand—were not soluble. Segments that did show measurable secondary structure included residues 27–29 (hydrophobic character), 42–45 (hydrophilic character), and the turn at 65–68. These regions also show enhanced protection against unfolding. It is possible that these sites, in addition to the above two proposed folding cores, form a larger set of initiation sites for folding.

Koide et al. (1993) studied the cooperative folding pattern of a proline *cis/trans* isomerization intermediate in an amide-hydrogen exchange experiment. Although this intermediate was much less protected against unfolding than the native state, parts of the β -sheet strands showed enhanced protection factors, particularly regions around residue positions 70 and 80 located in the two pleated β -sheets in the native structure. The effect of the P16 isomerization may be to prevent proper packing of the two sheets. In our calculation of unfolding penalties, no simple combination of deletion mutants, e.g., a combination of residues 1–24 and 50–99, showed any enhanced protection for the strand around residue 70. Thus, intermediates in the folding path that are trapped through a structural modification, such as proline isomerization, may not be characterized by the same set of denatured states as the native protein.

Staphylococcal nuclease

Staphylococcal nuclease is an extracellular protein responsible for degrading DNA and RNA into short oligonucleotides in the presence of free calcium. It specifically cleaves phosphodiester bonds. Staphylococcal nuclease is classified as an $\alpha + \beta$ protein with five β -strands forming a barrel and three α -helices. The helices form a plane on top of which the barrel rests. The active site is formed by a shallow depression at the edge of the barrel away from the helices plus a short coil region (Cotton et al., 1979). The amino acids essential for function are R35, E43, and R87. The structure contains no disulfide bonds. Under native conditions, the protein undergoes a *cis/trans* isomerization about the K116/P117 and H46/P47 bonds, favored by 90% in the *cis* conformation for the latter pair and by 90% *trans* in the former. Although such an isomerization equilibrium changes the characterization of the denatured states, our calculation is based only on the most probable conformer.

The experimental unfolding free energy penalties of Loh et al. (1993) are exhibited in Figure 10 along with the calculated ΔW_i^F values. The qualitative agreement of the data sets is confirmed by a statistically significant correlation at the 1% level, $r = 0.53$, $n_f = 66$, $p < 0.01$. We investigated several sets of native crystal coordinates from the Protein Data Bank for sensitivity to resolution and amino acid composition, but all data sets (Cotton et al., 1979; Loll & Lattman, 1989; Hynes & Fox, 1991) gave essentially the same theoretical protection pattern (data not shown). The theoretical values tend to follow closely the secondary structure elements, i.e., the five strands and the three helices are designated as regions of enhanced protection and are observed as such in the hydrogen exchange experiments. The folding unit identified here comprises the first three strands together with the fifth strand and the second helix.

Fig. 10. Unfolding penalties for each residue in the native structure of staphylococcal nuclease. The calculation utilized a segment window size $N_w = 8$, generating $10^{6.3}$ configurations in the denatured ensemble. The secondary structure elements are indicated on top of the graph, "B" designates β -sheets and "H" indicates where the α -helices are in the protein sequence. The secondary structure information was taken directly from the coordinate set 2sns deposited in the PDB. The experimental free energies were calculated from the hydrogen exchange protection factors measured by Loh et al. (1993). The correlation between the data sets is statistically significant, $r = 0.53$, $n_f = 66$, $p < 0.01$.

This three-dimensional domain is also clearly seen in Figure 11 where we have colored the most protected parts of the protein in blue, intermediate protection in green, and least protected regions in red. The most stable folding unit is close to the calcium binding site and the groove containing the active site of the protein. The two outer helices are not strongly stabilized, nor are the outer coil/loop regions indicated in Figure 11. The low cooperative protection afforded residues 42–56 in Figure 10 correlates with this region's flexibility, which is greater than that of the native protein (Kay et al., 1989).

A number of experimental studies on staphylococcal nuclease-like proteins has ascertained the correlation between denatured states being characterized by an ensemble of native-like fragments. Thus, the mutational variant K78C studied (James et al., 1992) using guanidinium hydrochloride denaturation suggested that non-native states contain large portions of the native structure, comprising residual structure or incompletely disrupted hydrophobic cores. The structure of the nuclease in denatured form was not found to resemble a random coil. Wang and Shortle investigated the deletion mutant $\Delta 131\Delta$ encompassing residues 1–3 and 13–140 of the native sequence with respect to urea denaturation (Wang & Shortle, 1995). While this protein is unstable under native conditions for the full-length sequence, it does fold in the presence of substrate. This protein must by homology be closely related to the denatured states accessible to the full-length protein under native conditions. Key features of the native state, such as the $\beta 2$ – $\beta 3$ turn with adjacent groups and the $\alpha 1$ and $\alpha 2$ helices, are observed. At pH 5.3 and 37°C the population of the $\alpha 2$ helix is 30% compared to 10–15% for the $\alpha 3$ helix, a difference predicted by our theoretical calculation and indicating that the $\alpha 2$ helix is more stable than the $\alpha 3$ helix under these conditions.

Fig. 11. A schematic picture of staphylococcal nuclease indicating the theoretically folding core in the three-dimensional structure. For visual orientation, a complexed thymidine 3',5'-bisphosphate inhibitor molecule and a Ca^{2+} atom are included. These molecules are located in the back of the picture where the active site is. The folding core, colored blue, comprises three structural elements, two strands of the β -barrel and helix 1. Less protected regions are colored in green and include the rest of the β -barrel and helix 2. Loop regions in the exterior of the protein, together with the third helix, are less protected against denaturation.

Conclusions

Available experimental data indicate that one strategy in nature for folding proteins is based on the assembly of native state configurations, i.e., on-pathway folding. Thus, intermediates in the folding process can show more or less the structural characteristics of the native state itself. These states are not disordered as in a random coil; rather, they are characterized by a significant fraction of secondary structure elements ordering themselves in their respective native three-dimensional spatial arrangement. For a sufficiently stable arrangement, we designate these structural elements as a folding core or, synonymously, an autonomous folding unit. This core corresponds to the parts of the protein that unravel last under denaturing conditions, and conversely this unit should be the first native-like structural element to appear in a refolding experiment. This is a cooperative effect, by which we mean that the structural information embedded in the amino acid sequence of these folding units alone does not completely determine the spatial relations in the folded polypeptide. Such context-dependent folding is also observed for smaller peptide segments. It is the entire sequence that determines the final structure of the protein.

The methodology used here to study amino acid residue protection against unfolding incorporates cooperativity by considering each residue in an ensemble of denatured states constructed from a combination of native structural elements. Though these methods

are obviously limited to equilibrium folding pathways, the current identification of important structural regions correlates well with the available experimental data and indicates the correspondence of these regions to autonomous folding units. As of yet, we cannot make any general statements about the structure of these folding cores; their context dependence may even preclude a structural identification and classification across protein families. If this is the case, a general protein folding code may not exist.

From the small data set examined here it appears that, in terms of evolution, folding cores are as conserved in the amino acid sequence as residues involved in function. Thus, both the overall structural arrangement, necessary for function, and possible folding pathways themselves are preserved in the evolution of proteins. Given the difficult task of folding large polypeptide chains into native configurations, this preservation is an excellent strategy for minimizing the costs of developing proteins in nature.

Methods

Knowledge-based scoring scheme

Although it is theoretically possible to calculate the free-energy change associated with the folding-unfolding transition of proteins, such a procedure is not currently computationally feasible. Since a wealth of information about protein structures in their native, or near-native, states from X-ray- or NMR-derived structures deposited in the Protein Data Bank, an alternative approach is to incorporate this knowledge directly into a method for studying protein structures. Such a knowledge-based scheme is fundamentally different from the more basic, atomically detailed, potential-based simulation approach, i.e., a knowledge-based scheme describes a free energy surface whereas atomic interaction potentials describe an internal energy surface. In order to recoup free energy quantities from the internal energy potential, one must sample an ensemble of systems, generated by molecular dynamics or Monte Carlo methods. The advantage of such an approach is that it is a complete method, i.e., both energetics and structural quantities correspond directly to experimentally measurable quantities. A knowledge-based approach is more limited, e.g., it is concerned only with the native states that correspond to the free-energy minima for the experimental conditions under which the structure was determined. There is no information about the denatured states; indeed, any state that is not native is a denatured state regardless of how much it resembles the native state. Thus denatured states are many; e.g., random coils, collapsed molten globules, and kinetically trapped misfolded states. Of course without the knowledge of the denatured state one cannot calculate the free energy associated with the process,

$$\text{unfolded} \xrightleftharpoons{\Delta G_{d \rightarrow n}} \text{folded}. \quad (3)$$

However, the study of structural correlations and their dependence on sequence in the known database of protein structures is still a useful tool to investigate amino acid preferences for adopting some recognizable structural features.

These preferences can be quantified in terms of a "free energy" or probability of adopting certain structural features for a given sequence. These "free energy" quantities do not correspond to any directly measurable quantity. Indeed, one can define a free-energy-like quantity based on the probability of observing any quantity, e.g., the frequency of occurrence of letters on this page. The mean-

ing of such a free energy is not clear, yet it could be used to identify authorship of this paragraph quite accurately.

A knowledge-based scoring scheme represents an average of the various physical conditions for each system included in the database. As long as the native state is populated to a significant level, the structural correlations extracted from the native state apply. A knowledge-based scoring scheme does not include the effects of temperature, ionic strength, pressure, pH, salt, co-solvents, etc. We cannot use it, without further parameterizations, to calculate the thermodynamic stability of proteins. If a sequence-structure correlation in a given protein exists, we can extract useful information about the importance of a given sequence for a given structure and vice versa. This procedure relies on two premises: That the structural database be a fair representation of an unbiased sampling of all protein folds and that we pick the relevant correlations to study. These issues are addressed in the following sections.

Structural database

The Protein Data Bank (Bernstein et al., 1977; Abola et al., 1987) now consists of over 4,000 protein entries, each containing one or more structures. It is expected that this number will increase exponentially, yet the true amount of information about sequence-structure relations will remain relatively constant for the given folds currently identified (Orengo et al., 1994). In order to form unbiased averages over the data set, one must select so as to reduce the redundancy of sequences as well as folds. Fischer et al. (1995) have used a procedure that selects high-quality structures, which are non-homologous both in sequence and structure, to reduce the number of unique proteins in the data bank. We have used this subset of non-homologous proteins (Naor et al., 1996) as our structural database. Thus our scoring scheme will be based on the database of $N_{dbs} = 235$ members given in Table 1.

Structural correlations

We have chosen to study structural properties based solely on C_α coordinates. Therefore, all the information about the backbone structure and possible side-chain conformations for a given residue is condensed into a single point. All the information necessary to regenerate the full, atomically detailed backbone and side chains has been shown to be contained in the C_α -trace (Levitt, 1992). Since the functional form describing such a projection and the interaction between such points are far too complex to write down analytically, we have elected to discretize space and work only with a numerical form of correlations and scores.

Given the still limited amount of information in the database it is only numerically warranted to develop pair-wise preferences between amino acid types; for example, the virtual angle between three connected C_α -atoms requires 20^3 or 8,000 individual potential types. Instead we employ a Kirkwood approximation (McQuarrie, 1976) for higher-order correlations such that these correlations are constructed from lower-order correlations,

$$p(A_1, A_2, A_3) \sim F[p(A_1, A_2), p(A_2, A_3)] \quad (4)$$

and can be used to reconstruct the necessary three- or four-atom correlation for torsion angles.

We focus on correlations beyond nearest neighbors, namely virtual angles θ between connected triplets of amino acids, dihedral angles ϕ defined by connected quadruplets of amino acids, structural preferences based on distances r between pairs of "non-bonded" sites, and a pairwise local density term ρ defining available

Table 1. Selected structural database^a

laaf	laak	laba	labh	laca	lala	laoz(a)
lapm(e)	lapm(i)	lapo	latx	lbac	lbal	lbbi
lbbo	lbbp(a)	lbbt(2)	lbbt(4)	lbja	lbmv(2)	lbn2(1)
lbst	lbtc	lbus	lc5a	lcbh	lcbn	lcil
lcmb(a)	lcob(a)	lcol(a)	lcpc(a)	lcro(a)	lcsc	lcse(e)
lctf	ld66(a)	ldpi	leaa	leca	lepi	leps
leri(a)	lezm	lfcb(a)	lfid2	lfia(a)	lfkb	lfus
lgai	lgd1(o)	lgf1	lgla(g)	lglu(a)	lgpl(a)	lgpb
lgrc(a)	lgsg(p)	lgst(a)	lhcl	lhdd(c)	lhfh	lhge(a)
lhgc(b)	lhip	lhis(a)	lhit(b)	lhiv(a)	lhas(a)	lige(a)
ligf(m)	l158	l1ab	l1ap	l1ec	l1fi	l1mb(3)
l1ts(a)	l1ts(d)	l1z1	l1mb1	l1mco	l1min	l1mli
l1xb	l1omp	l1ova(a)	l1pda	l1pdc	l1pfa(a)	l1pgx
l1phh	l1phs	l1pii	l1plc	l1poc	l1ppn	l1ppt
l1prc(c)	l1prc(h)	l1pya(a)	l1pyp	l1rhd	l1rtc	l1rve(a)
l1sc	l1sha(a)	l1stp	l1tab(i)	l1tfd	l1tgl	l1thg
l1thi	l1tme(1)	l1tnf(a)	l1tpt	l1trb	l1tta(a)	l1ubq
l1ula	l1utg	l1vsg(a)	l1wsy(a)	l1wsy(b)	l1xis	l1ycc
l1zaa(c)	l1znf	256b(a)	2abk	2ada	2ak3(a)	2baa
2bat	2bb2	2bbm(a)	2bbq(a)	2bpa(1)	2bpa(2)	2bpa(3)
2cba	2cdv	2cmd	2cyp	2dnj(a)	2ech	2end
2er7(e)	2gls(a)	2gn5	2had	2hsd(a)	2hsp	2kai(a)
2lbp	2ltn(b)	2meg(2)	2mev(4)	2mhu	2mrt	2ms2(a)
2msb(a)	2nrd	2ovo	2pcd(a)	2pec	2pgd	2phy
2pia	2plv(4)	2por	2por	2pte(e)	2reb	2rig
2rn2	2rns	2rsl(a)	2sas	2sic(i)	2sn3	2sns
2stv	2tbv(a)	2tmv(p)	2ts1	2wrp(r)	3b5c	3cbh
3cd4	3dpa	3dpv(a)	3gap(a)	3gfl	3gly	3grs
3hhr(b)	3il8	3mon(a)	3mon(b)	3pgk	3pmg(a)	3pte
3rub(s)	3sc2(a)	3sdp(a)	4bp2	4clg(a)	4dfr(a)	4enl
4gcr	4icb	4ins(a)	4ins(b)	4ptp	4sgb(i)	5cpa
5fbp(a)	5p21	5pti	5rub(a)	5tim(a)	6taa	7aat(a)
7acn	7api(b)	7icd	7rsa	8abp	8adh	8atc(a)
8atc(b)	8fab(a)	8rxn(a)	9wga(a)			

^a The PDB code refers to the name of the deposited coordinate sets in the Protein Data Bank (Bernstein et al., 1977; Abola et al., 1987). The (x)-notation gives the non-redundant chain identifier. The table was collated in the summer of 1996.

regions in space. From these correlations we will construct a score, W , based on the structures in the database. The total score for a sequence $\{A\}_N$ with coordinates $\{\mathbf{X}\}_N$ is then given as

$$W(\{A\}_N; \{\mathbf{X}\}_N) = \alpha_\theta W^\theta(\{A\}_N; \{\mathbf{X}\}_N) + \alpha_\phi W^\phi(\{A\}_N; \{\mathbf{X}\}_N) + \alpha_r W^r(\{A\}_N; \{\mathbf{X}\}_N) + \alpha_p W^p(\{A\}_N; \{\mathbf{X}\}_N) \quad (5)$$

where each individual component has been estimated from frequencies of occurrence in the non-homologous data set and is expressed in units of RT. It is useful to transform the calculated probabilities to an energy-like scoring function so as to be able to use techniques developed for minimizations as well as for simulations.

A minimization of the score W corresponds to maximizing the joint probability of observing the most likely sequence-structure relation for all known structures in the data bank.

Angles, W^θ

The probability of occurrence of a given virtual angle $\cos \theta$ between connected residues $i, i+1, i+2$, $p^\theta(A_i, A_{i+1}, A_{i+2})$, is estimated from two pair probabilities, i.e.,

$$p^\theta(A_i, A_{i+1}, A_{i+2}) \equiv [p^\theta(A_i, A_{i+1}, \langle A_{i+2} \rangle) p^\theta(\langle A_i \rangle, A_{i+1}, A_{i+2})]^{1/2} \quad (6)$$

where the $\langle A_j \rangle$ indicates that we have averaged the j th amino acid type over all 20 naturally occurring amino acids. The frequency of occurrence is calculated as

$$p^\theta(A_i, A_{i+1}, \langle A_{i+2} \rangle; \cos \theta) = \sum_m \sum_{n=1}^{N_{\text{obs}} - 2} \delta(A_i) \delta(A_{i+1}) \delta(\cos \theta - \cos \theta_{i,i+1,i+2}). \quad (7)$$

A sampling window of $\Delta \cos \theta = 0.04$ rad is used. This count is normalized and used to estimate a residue and angle-specific potential of mean force, w^θ :

$$w^\theta(A_i, A_{i+1}, A_{i+2}; \cos \theta) = -kT \ln p^\theta(A_i, A_{i+1}, A_{i+2}; \cos \theta) + c_{i,i+1,i+2} \quad (8)$$

where we have calibrated the constant $c_{i,i+1,i+2}$ so as to coincide with the Miyazawa/Jernigan (MJ) energy scale (Miyazawa & Jernigan, 1996). Thus, if the MJ contact energy between residues A_i and A_{i+1} is denoted $\epsilon_{i,i+1}$, we force the total score in our scheme to be equal to this number,

$$\epsilon_{i,i+1} = \int w^\theta(A_i, A_{i+1}, A_{i+2}; \cos \theta) p^\theta(A_i, A_{i+1}, A_{i+2}; \cos \theta) d\Omega; \quad (9)$$

this defines the constants c used in Equation 8. The score arising from the virtual-angle terms of a given sequence $\{A\}_N$ with coordinates $\{\mathbf{X}\}_N$ is then given as

$$W^\theta(\{A\}_N; \{\mathbf{X}\}_N) = \sum_{n=1}^{N-2} w^\theta(A_n, A_{n+1}, A_{n+2}; \cos \theta_{n,n+1,n+2}) \quad (10)$$

and where the virtual angle between residues A_n, A_{n+1}, A_{n+2} is calculated from

$$\cos \theta_{n,n+1,n+2} = \frac{(\mathbf{X}_n - \mathbf{X}_{n+1}) \cdot (\mathbf{X}_{n+2} - \mathbf{X}_{n+1})}{|\mathbf{X}_n - \mathbf{X}_{n+1}| |\mathbf{X}_{n+2} - \mathbf{X}_{n+1}|}. \quad (11)$$

Dihedrals, W^ϕ

The probability of occurrence of a given virtual dihedral angle ϕ between connected residues $i, i+1, i+2, i+3$, $p^\phi(A_i, A_{i+1}, A_{i+2}, A_{i+3})$ is estimated from three pair probabilities, i.e.,

$$p^\phi(A_i, A_{i+1}, A_{i+2}, A_{i+3}) \equiv [p^\phi(A_i, A_{i+1}, \langle A_{i+2} \rangle, \langle A_{i+3} \rangle) \times p^\phi(\langle A_i \rangle, A_{i+1}, A_{i+2}, \langle A_{i+3} \rangle) \times p^\phi(\langle A_i \rangle, \langle A_{i+1} \rangle, A_{i+2}, A_{i+3})]^{1/3} \quad (12)$$

where $\langle A_i \rangle$ indicates that we have averaged the j th amino acid over all amino acid types. The pairwise frequency of occurrence is then calculated as

$$p^\phi(A_i, A_{i+1}, \langle A_{i+2} \rangle, \langle A_{i+3} \rangle; \phi) \\ = \sum_m \sum_{n=1}^{N_{\text{dbs}} N_m - 3} \delta(A_i) \delta(A_{i+1}) \delta(\phi - \phi_{i,i+1,i+2,i+3}). \quad (13)$$

A sampling window $\Delta\phi$ of 0.14 rad is used. The count is normalized and the complete dihedral probability distribution function is reconstructed using Equation 12. The residue and angle-specific potential of mean force, w^ϕ :

$$w^\phi(A_i, A_{i+1}, A_{i+2}, A_{i+3}; \phi) \\ = -kT \ln p^\phi(A_i, A_{i+1}, A_{i+2}, A_{i+3}; \phi) + c_{i,i+1,i+2,i+3} \quad (14)$$

is calibrated with the constant $c_{i,i+1,i+2,i+3}$ so as to coincide with the MJ energy scale (Miyazawa & Jernigan, 1996). The score arising from the virtual dihedral terms of a given sequence $\{A\}_N$ with coordinates $\{\mathbf{X}\}_N$ is thus

$$W^\phi(\{A\}_N, \{\mathbf{X}\}_N) = \sum_{n=1}^{N-3} w^\phi(A_n, A_{n+1}, A_{n+2}, A_{n+3}; \phi_{n,n+1,n+2,n+3}) \quad (15)$$

where the virtual dihedral angle between residues $A_n, A_{n+1}, A_{n+2}, A_{n+3}$ is calculated from

$$\phi_{n,n+1,n+2,n+3} = \text{sign}(\phi) \arccos(\mathbf{n}_1 \cdot \mathbf{n}_2) \quad (16)$$

$$\mathbf{n}_1 = \frac{(\mathbf{X}_n - \mathbf{X}_{n+1})}{|\mathbf{X}_n - \mathbf{X}_{n+1}|} \times \frac{(\mathbf{X}_{n+2} - \mathbf{X}_{n+1})}{|\mathbf{X}_{n+2} - \mathbf{X}_{n+1}|} \quad (17)$$

$$\mathbf{n}_2 = \frac{(\mathbf{X}_{n+2} - \mathbf{X}_{n+1})}{|\mathbf{X}_{n+2} - \mathbf{X}_{n+1}|} \times \frac{(\mathbf{X}_{n+2} - \mathbf{X}_{n+3})}{|\mathbf{X}_{n+2} - \mathbf{X}_{n+3}|}. \quad (18)$$

Non-bonded interactions, W^r

The radial distribution of C_α -pairs that are within 12.5 Å of each other was calculated for all amino acid types. In order to reduce the intramolecular correlations present among closely linked residues, we include in the count only those separated by eight or more residues in the sequence. Since a radial description is devoid of directional information, we have tried, by considering virtual dihedral angles between non-connected strands, to describe a local handedness (Braun, 1983; Covell, 1992) via the non-bonded potential. The pair-distribution depends both on the radial distance between residues A_i and A_j and on the relative orientation of the local chain directions around residue i , $\mathbf{X}_{i-1} - \mathbf{X}_i$ and residue j , $\mathbf{X}_{j+1} - \mathbf{X}_j$. We calculate the distribution over the selected database as

$$p^r(A_i, A_j; r_{ij}, \phi_{i-1,i,j,j+1}) = \sum_m \sum_{n=1}^{N_{\text{dbs}} N_m - 7} \sum_{n'=n+7}^{N_M} \delta(A_i) \delta(A_j) \\ \times \delta(r_{ij} - r_{n,n'}) H(12.5 - r_{n,n'}) \\ \times \delta(\phi_{i-1,i,j,j+1} - \phi_{n-1,n,n',n'+1}) \quad (19)$$

where H is a Heaviside step-function and the virtual dihedral angle is calculated as in Equations 17 and 18. We used a window of $\Delta r =$

0.25 Å and $\Delta\phi = 0.62$ rad to collect the data. A potential of mean force is constructed to be analogous to the angle and dihedral terms defined above:

$$w^r(A_i, A_j; r_{ij}, \phi_{i-1,i,j,j+1}) = -kT \ln p^r(A_i, A_j; r_{ij}, \phi_{i-1,i,j,j+1}) + c_{ij}. \quad (20)$$

Each constant c_{ij} is determined using the requirement that the total score for residues A_i and A_j in the specified volume is equal to the MJ contact energy ϵ_{ij} . Thus, the score from the non-bonded terms for the sequence $\{A\}_N$ with coordinates $\{\mathbf{X}\}_N$ is written as

$$W^r(\{A\}_N, \{\mathbf{X}\}_N) = \sum_{n=1}^{N-4} \sum_{n'=n+4}^N w^r(A_n, A_{n'}; r_{nn'}, \phi_{n-1,n,n',n'+1}). \quad (21)$$

If we cannot define the non-bonded virtual dihedral angle, we substitute the $w^r(r, \phi)$ score with an angle-independent potential $w^r(r, \langle \phi \rangle)$ averaged over all angles.

Density mask, W^p

The non-pairwise density correction term found in many knowledge-based potentials is substituted here by a pairwise additive term, denoted a “density mask.” This term is evaluated for each amino acid-type pair and includes a global handedness (Braun, 1983) not present in the non-bonded score W^r . For a connected triplet of residues with amino acid A_i at the center we define a local coordinate system

$$\mathbf{z}' = (\mathbf{X}_{i+1} - \mathbf{X}_i) \times (\mathbf{X}_{i-1} - \mathbf{X}_i) \quad (22)$$

$$\mathbf{x}' = (\mathbf{X}_{i-1} - \mathbf{X}_i) \times \mathbf{z}' \quad (23)$$

$$\mathbf{y}' = \mathbf{z}' \times \mathbf{x}'. \quad (24)$$

The spatial distribution of amino acid A_j is then evaluated in the new coordinate system, which is defined by the three normalized vectors $(\mathbf{x}', \mathbf{y}', \mathbf{z}')$ for each triplet of connected amino acids as

$$p^p(A_i, A_j; x_{ij}, y_{ij}, z_{ij}) = \sum_m \sum_{n=1}^{N_{\text{dbs}} N_m - 7} \sum_{n'=n+7}^{N_M} \delta(A_i) \delta(A_j) H(12.5 - r_{n,n'}) \\ \times \delta(x_{ij} - x_{nn'}) \delta(y_{ij} - y_{nn'}) \delta(z_{ij} - z_{nn'}). \quad (25)$$

The binning size in the data collection is 2.5 Å. A potential formulation is defined as

$$w^p(A_i, A_j; x_{ij}, y_{ij}, z_{ij}) = -kT \ln p^p(A_i, A_j; x_{ij}, y_{ij}, z_{ij}) + c \quad (26)$$

where the constant c is determined from the condition that the total score for all possible residues A_i and A_j in the specified volume is zero. The score from these terms for the sequence $\{A\}_N$ with coordinates $\{\mathbf{X}\}_N$ is then

$$W^p(\{A\}_N, \{\mathbf{X}\}_N) = \sum_{n=1}^{N-4} \sum_{n'=n+4}^N w^p(A_n, A_{n'}; x_{nn'}, y_{nn'}, z_{nn'}). \quad (27)$$

If it is not possible to define a local coordinate system for a residue, no contribution is made to the density mask score.

Threadings

A threading procedure was used to determine the relative weights assigned to the individual scores in Equation 5. In threading, the sequence $\{A\}_N$ of a given protein structure is incrementally assigned to all coordinates $\{X\}_1 \dots \{X\}_{N_{\text{dbs}}}$ in the data set (Sippl, 1990; Casari & Sippl, 1992; Sippl & Weitckus, 1992; Bryant & Lawrence, 1993; Sippl, 1993; Wodak & Rومان, 1993; Bauer & Beyer, 1994; Kocher et al., 1994; Sippl, 1995). If the sequence information can rank the score of the native structures as being the most favorable, the scoring scheme is well suited to study sequence-

structure relationships. Although there is a nominal agreement between the energy scale introduced by Miyazawa/Jernigan (Miyazawa & Jernigan, 1996) and the one employed here, the different structural correlations considered in the two schemes result in both quantitative and qualitative differences.

The largest protein in the threading set (1L58) has 22,113 possible threadings with the database used. Larger proteins were not considered in order to maximize the number of possible threadings. A weighting scheme is used that reduces the importance of the secondary structure elements in favor of the non-bonded contacts with $\alpha_\theta = \alpha_\phi = 0.30/N$, $\alpha_r = 2.80/N$, and $\alpha_\rho = 0.50/N$. The results of this threading procedure are summarized in Table 2, which lists ranks, deviations, and the individual components of the

Table 2. Threading results

PDB ^a	Rank	N	W ^b	$\Delta W/\delta W$	$\alpha_\theta W^\theta$	$\alpha_\phi W^\phi$	$\alpha_r W^r$	$\alpha_\rho W^\rho$
1bn2	1	86	-5.60	-1.67	-0.81	-0.28	-6.42	1.93
1cbn	1	46	-6.24	-2.17	-1.05	-1.01	-5.33	1.16
1cob	1	151	-11.56	-2.36	-0.94	-0.84	-11.18	1.40
1eca	1	136	-11.49	-2.15	-1.12	-1.23	-9.96	0.83
1fkb	1	107	-11.03	-2.53	-0.97	-0.92	-10.31	1.18
1fus	1	105	-8.64	-2.15	-0.85	-0.68	-8.34	1.25
1hfh	1	120	-7.91	-1.97	-0.84	-0.79	-7.92	1.65
1l58	1	164	-12.56	-2.25	-1.09	-1.16	-11.30	0.99
1lzl	1	130	-10.52	-2.44	-1.02	-0.99	-10.08	1.57
1nxb	1	62	-6.02	-2.36	-0.77	-0.76	-6.15	1.68
1plc	1	99	-10.22	-2.14	-0.94	-0.72	-9.59	1.04
1poc	1	134	-9.03	-2.11	-0.97	-0.95	-8.73	1.63
1sha	1	103	-11.19	-2.48	-0.98	-0.99	-10.29	1.08
1tnf	1	152	-11.10	-1.89	-0.97	-0.89	-10.37	1.14
1ubq	1	76	-10.18	-2.28	-1.00	-0.98	-8.91	0.72
1utg	1	70	-9.38	-2.41	-1.05	-1.20	-7.69	0.57
1ycc	1	108	-8.76	-1.91	-0.89	-0.66	-8.48	1.28
256b	1	106	-12.12	-2.50	-0.99	-1.12	-10.74	0.73
2end	1	137	-11.81	-2.13	-1.13	-1.12	-10.51	0.96
2ovo	1	56	-8.09	-2.92	-0.97	-0.92	-7.31	1.11
2sns	1	141	-11.93	-2.33	-0.87	-0.80	-11.38	1.14
2tmv	1	154	-10.08	-1.60	0.95	-0.99	-9.07	0.95
3rub	1	123	-10.10	-2.09	-1.06	-1.06	-9.14	1.17
4dfr	1	159	-11.17	-2.05	-0.95	-0.84	-10.59	1.21
4icb	1	76	-11.87	-2.55	-1.02	-1.05	-10.45	0.66
5pti	1	58	-8.27	-2.84	-0.99	-0.94	-7.82	1.48
7rsa	1	124	-9.65	-2.45	-0.86	-0.84	-9.16	1.21
8rxn	1	52	-8.32	-2.96	-0.91	-0.94	-7.49	1.04
1aaf	>10	55						
1lab	>10	80						
1mli	>10	96						
2bbm	>10	148						
Structures not in the training set								
1mbc	1	153	-10.11	-1.56	-1.08	-0.96	-11.20	3.13
2lyz	1	129	-6.74	-1.36	-0.97	-0.43	-9.65	4.32
4fxn	1	138	-9.35	-1.54	-0.95	-0.39	-11.84	3.83
4hvp	1hiv	99	-9.74	-1.88	-1.02	-0.83	-8.92	1.04
4cpv	3	108	-10.71	-1.81	-1.05	-1.00	-11.00	2.35
2trx	6	108	-6.81	-1.74	-1.03	-0.87	-10.52	2.51
1asu	>10	162						
1nef	>10	136						

^aThe PDB code refers to the name of the deposited coordinate sets in the Protein Data Bank (Bernstein et al., 1977; Abola et al., 1987) with the chain identifiers as in Table 1.

^bEnergy units are in RT.

score. The deviation of the score from the mean, expressed in units of the standard deviation, σ_w , of the threading scores for that sequence, is calculated as

$$\Delta W/\delta W \equiv \frac{W - \langle W \rangle}{\sigma_w}. \quad (28)$$

In Table 2, results are given for a subset of structures in the database and for some that were not part of the parameterization of the score W . The individual scores W range from a low of -5.6 to a high of -12.5 , with the $\Delta W/\delta W$ measured from -1.4 to -3.0 . The total score is not strongly dependent on the size of the molecule. The threadings of sequences that were not in the database show some reduction of the $\Delta W/\delta W$, primarily through their higher W^p components. Although the 4hvp structure was not included in the threading configurations, the sequence correctly identified the similar 1hiv structure as the native configuration. In contrast to the Miyazawa/Jernigan scheme (Miyazawa & Jernigan, 1996), which ranks the 5pti sequence 44th in the 5pti fold, our scheme ranks the sequence correctly in first place. This discrepancy is because of the different number of points assigned to each amino acid and the inclusion of angularly dependent correlations in our scheme. Further studies indicate that both Monte Carlo annealing (Andricioaei & Straub, 1996) and segment-growth procedures have proven resilient in generating structures with higher scores and high root-mean-square deviations from the native structure for some small proteins.

Calculation of unfolding penalties

We cannot within a reasonable time generate all possible denatured states of a given sequence. Instead, we limit ourselves to a small but important subset based on native structure itself. We generate an ensemble of states by forming all possible segment combinations where the "native" (n) segments interact via the scoring scheme in Equation 5, and all "denatured" (d) segments have no interactions either among themselves or with the "native" segments (Hilser & Freire, 1996, 1997). For a given amino acid sequence of length N , A_N , we divide the sequence space into N/N_w segments, with N/N_w rounded up so that trailing residues are assigned to a smaller $\text{mod}(N, N_w)$ -sized segment. Each window can be either native (n) or denatured (d), i.e., on or off, so that there are a total of $2^{N/N_w}$ possible combinations of these segments. This defines the sequence mask, schematically pictured in Figure 12. Since the enumeration of the amino acids in each window need not coincide with the enumeration of the sequence mask, we can start the enumeration in N_w places by simply sliding the entire sequence one step at a time.

The probability of finding a residue i in the unfolded state in the ensemble is

$$p_i^d = \sum_{k=1}^{2^{N/N_w}} \sum_{l=1}^{N_w} e^{-W(kT)\delta(s_{kl}(i) - d)} \quad (29)$$

where $\delta(s_{kl}(i) - d)$ signifies that only states of the specified sequence mask kl that contain states in which residue i is denatured are included in the sum. Likewise, the probability of finding a residue i in the native state is

$$p_i^n = \sum_{k=1}^{2^{N/N_w}} \sum_{l=1}^{N_w} e^{-W(kT)\delta(s_{kl}(i) - n)}. \quad (30)$$

Sequence Masks

1	1	0	1	0	0	0	0
0	1	1	0	0	1	1	1
1	0	1	1	0	1	1	1
0	0	1	0	0	1	0	1

Windows

N	Q	W	A	R	S	A	A	F	G	G	G
Q	W	A	R	S	A	A	F	G	G	G	P
W	A	R	S	A	A	F	G	G	G	P	I

Fig. 12. The sequence masks define the regions of the sequence that are allowed to interact with each other. All residues within an active window are deemed to be "native." The complete generation of all possible combinations of "native" and "denatured" states yield an ensemble of $2^{N/N_w}$ structures. Each active window of the sequence encompasses N_w amino acid residues—in this case 12 residues. For each such window, we push the sequence N_w times one residue at a time to slide over all possible starting positions of this sequence. This procedure yields a total of $2^{N/N_w} N_w$ structures to include in the unfolding penalty calculation.

The cooperative unfolding penalty of residue i is then defined as

$$\Delta W_i^F \equiv -RT \ln \frac{p_i^d}{p_i^n}. \quad (31)$$

The quantity ΔW_i^F in Figure 13 is given as a function of residues and window size for the bovine pancreatic trypsin inhibitor (Wlodawer et al., 1984). The convergence of the unfolding penalties as a function of window size, or more properly the number of structures included in the ensemble $N_e = 2^{58/N_w} N_w$, indicates that gross features of the curve are observable with a relatively large window size. Thus, the most protected residues around F22 are already discernible with the $N_w = 10$ or $N_e = 6.4 \times 10^2$. The relative ordering of the four most prominent peaks is already in place for the $N_w = 6$ or $N_e = 6.1 \times 10^3$. As more conformations are included in the calculations, the ΔW_i^F -curve becomes more resolved. A satisfactory value of the N_e is empirically found to be 10^6 – 10^7 for a number of both small and large proteins. If two strongly protected residues are close in sequence it may not be possible to resolve their individual peaks, unless the window size is comparable to the difference in sequence positions of the residues. In general, the reliability of identifying one residue as being located at a peak position is about $\pm N_w/5$.

The importance of performing an average over the entire ensemble of native and denatured chain segments is evident from calculating just the individual amino acid contributions to the total score $W(\{A\}_N; \{X\}_N)$. The difference in score when changing each residue to a glycine,

$$\Delta W_i = W(\{A'\}_N; \{X\}_N)_{A_i \rightarrow G} - W(\{A\}_N; \{X\}_N) \quad (32)$$

is given at the bottom of Figure 13 for BPTI. The ΔW_i quantity bears little resemblance to the unfolding penalties ΔW_i^F . Not only

Fig. 13. The convergence of ΔW_i^F as a function of window size, N_w , for the bovine pancreatic trypsin inhibitor (5pti). Gross features of the protection pattern are observable even at low resolutions, whereas convergence appears around 10^6 included configurations. For the BPTI-molecule this occurs at a window size of 3–4, in the data presented here a window size $N_w = 3$ have been consistently used for all the BPTI calculations. The graph at the bottom gives the change in score ΔW_i for a single amino acid substitution into glycine of the raw score W .

are these contributions too small, they also fail to identify the important cluster of residues at position 20–23 as being protected against unfolding. Thus, by using the denatured ensemble averaging technique, we can predict behavior from the calculated unfolding penalties that is not simply a function of side-chain burial.

If the generated ensemble of denatured states has a strong correspondence to the actual collection of protein configurations in equilibrium with the native state, then the unfolding penalties calculated in Equation 31 will also be congruent to the experimentally measured hydrogen exchange protection factors. They are, however, not the same quantity. It is, of course, only possible to access experimentally protection factors for buried residues, as any surface-exposed amide hydrogen can freely and rapidly exchange with the solvent on the time scale of global folding/unfolding events. Furthermore, if there exists in the folding/unfolding process intermediates that have well-defined and energetically important structural properties and that do not resemble the native state, the protection factors calculated by our procedure may not correspond to the measured data.

The experimentally accessible quantity used to describe partially unfolded structures under native conditions is the amide-hydrogen exchange measured by NMR. All amides in a polypeptide chain exchange their protons with water protons if they are exposed to solvent. From a theoretical point of view, the unfolding penalties of Equation 31 are not the same as the free energy change measured by the hydrogen exchange experiment. Yet there is a strong qualitative similarity between the calculated results and the experimental data on appropriately buried amides. This shape similarity between the calculated and the measured values allows us to gauge protection patterns and identify folding units. In order to emphasize this comparison we have scaled the ΔW_i^F values to correspond to an overall folding penalty for each particular protein at the given experimental conditions, which we do not account for in our simplified model. Both the unscaled ΔW_i^F and the scaled values are given when comparing $\Delta G_{op,i}$ and ΔW_i^F values. As is

detailed in Results and discussion, the limitations of the scoring scheme and the assumption about the partially denatured states do not crucially impinge on their interpretative values.

As our aim is to identify cooperative folding units we need to determine if the experimental delineation of folding units agrees with the structural regions determined by the theoretical unfolding penalties. The correlation coefficient between two data sets allows us to determine whether the observation is statistically significant (Crow et al., 1960). This correlation, r , is not sensitive to the scaling of the ΔW_i^F values. As the level of significance employed in this work is at the 1% level, we cannot reject the hypothesis that the data sets are correlated if the p value is less than 0.01.

Sequence similarities

To investigate positional variability, we assembled a group of functionally similar sequences from a search of a non-redundant protein database¹ with a candidate sequence. The MPSRCH² program was used for the similarity search. This collection of sequences were then aligned using the PileUp multiple sequence alignment program.³ Finally a running average, employing a window of four residues, of similarity among the aligned sequences was generated.

Acknowledgments

The staff of the Biomedical Supercomputing Center, Frederick Cancer Research and Development Center, Frederick, Maryland, is thanked for its assistance and for access to its computer facilities. We thank the NIH Intramural Targeted Anti-viral AIDS Program for support. This research is also sponsored by the National Cancer Institute, Department of Health and Human Services, under contract with Science Applications International Corporation. The contents of this publication do not necessarily reflect the views or policies of the DHHS, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

References

- Abola EE, Bernstein FC, Bryant SH, Koetzle TF, Weng J. 1987. Protein data bank. In: Allen FH, Bergerhoff G, Sievers R, eds. *Crystallographic databases—Information content, software systems, scientific applications*. Bonn: Data Commission of the International Union of Crystallography. pp 107–132.
- Almasy RJ, Dickerson RE. 1978. *Pseudomonas* cytochrome c_{551} at 2.0 Å resolution: Enlargement of the cytochrome c family. *Proc Natl Acad Sci USA* 75:2674–2678.
- Andricioaei I, Straub JE. 1996. Generalized simulated annealing algorithms using Tsallis statistics: Application to conformational optimization of a tetrapeptide. *Phys Rev E* 55:3055–3058.
- Anfinsen CB. 1973. Principles that govern the folding of protein chains. *Science* 181:223–230.
- Bai Y, Milne JS, Mayne L, Englander SW. 1993. Primary structure effects on peptide group hydrogen exchange. *Proteins Struct Funct Genet* 17:75–86.
- Bai Y, Sosnick TR, Mayne L, Englander WS. 1995. Protein folding intermediates: Native-state hydrogen exchange. *Science* 269:192–197.
- Bauer A, Beyer A. 1994. An improved pair potential to recognize native protein folds. *Proteins Struct Funct Genet* 18:254–261.
- Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rogers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A

¹NRP (Non-Redundant Protein) Database. Distributed on the Internet via anonymous FTP from ftp.ncifcrf.gov, under the auspices of the National Cancer Institute's Frederick Biomedical Supercomputing Center.

²Release 2.1D. John F. Collins, Biocomputing Research Unit. Copyright 1993, 1994, 1995, University of Edinburgh, United Kingdom. Distribution rights by IntelliGenetics, Inc.

³Program Manual for the Wisconsin Package, Version 8, September 1994. Genetics Computer Group, 575 Science Drive, Madison, Wisconsin 53711.

- computer-based archival file for macromolecular structures. *J Mol Biol* 112:535–542.
- Braun W. 1983. Representation of short- and long-range handedness in protein structures by signed distance maps. *J Mol Biol* 163:613–621.
- Bryant SH, Lawrence CE. 1993. An empirical energy function for threading protein sequence through the folding motif. *Proteins Struct Funct Genet* 16:92–112.
- Bushnell GW, Louie GV, Brayer GD. 1990. High-resolution three-dimensional structure of horse heart cytochrome *c*. *J Mol Biol* 214:585–595.
- Casari G, Sippl MJ. 1992. Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structure of globular proteins is able to identify native folds. *J Mol Biol* 224:725–732.
- Chazin WJ, Wright PE. 1988. Complete assignment of the ^1H nuclear magnetic resonance spectrum of French bean plastocyanin. Sequential resonance assignments, secondary structure, and global fold. *J Mol Biol* 202:623–636.
- Cotton FA, Hazen EE Jr, Legg MJ. 1979. Staphylococcal nuclease: Proposed mechanism of action based on structure of enzyme-thymidine 3',5'-bisphosphate-calcium ion complex at 1.5 Å resolution. *Proc Natl Acad Sci USA* 76:2551–2555.
- Covell DG. 1992. Folding protein α -carbon chains into compact forms by Monte Carlo methods. *Proteins Struct Funct Genet* 14:409–420.
- Creighton TE. 1992. *Protein folding*. New York: Freeman.
- Crow EL, Davis FA, Maxfield MW. 1960. *Statistics manual*. New York: Dover.
- Danishesky AT, Housset D, Kim KY, Tao F, Fuchs J, Woodward C, Wlodawer A. 1993. Crevice-forming mutants in the rigid core of bovine pancreatic trypsin inhibitor: Crystal structures of F22A, Y23A, N43G, and F45A. *Protein Sci* 2:577–587.
- Dickerson RE. 1980. Evolution and gene transfer in purple photosynthetic bacteria. *Nature* 283:210–212.
- Dickerson RE, Timkovich R. 1975. Cytochromes *c*. *The enzymes* 11:397–547.
- Dyson HJ, Sayer JR, Merutka G, Shin HC, Lerner RA, Wright PE. 1992. Folding of peptide fragments comprising the complete sequence of proteins. Models for initiation of protein folding. II. Plastocyanin. *J Mol Biol* 226:819–835.
- Dyson HJ, Wright PE. 1993. Peptide conformation and protein folding. *Curr Opin in Struct Biol* 3:60–65.
- Elöve GA, Chaffotte AF, Roder H, Goldberg ME. 1992. Early steps in cytochrome *c* folding probed by time-resolved circular dichroism and fluorescence spectroscopy. *Biochemistry* 31:6876–6883.
- Englander SW, Mayne L. 1992. Protein folding studied using hydrogen-exchange labeling and two-dimensional NMR. *Annu Rev Biophys Biomol Struct* 21:243–265.
- Finucane MF, Jardetzky O. 1995. Mechanism of hydrogen-deuterium exchange in *trp* repressor studied by ^1H - ^{15}N NMR. *J Mol Biol* 253:576–589.
- Fischer D, Tsai CJ, Nussinov R, Wolfson H. 1995. A 3D sequence-independent representation of the protein bank. *Protein Eng* 8:981–997.
- Fisher A, Taniuchi H. 1992. A study of core domains, and the core domain-domain interaction of cytochrome *c* fragment complex. *Archives of Biochemistry and Biophysics* 296:1–16.
- Goldenberg DP. 1992. Native and non-native intermediates in the BPTI folding pathway. *TIBS* 17:257–261.
- Goldenberg DP, Zhang JX. 1993. Small effects of amino acid replacements on the reduced and unfolded state of pancreatic trypsin inhibitor. *Proteins Struct Funct Genet* 15:322–329.
- Goodman EM, Kim PS. 1989. Folding of a peptide corresponding to the α -helix in bovine pancreatic trypsin inhibitor. *Biochemistry* 28:4343–4347.
- Gross EL. 1993. Plastocyanin: Structure and function. *Photosynthesis Research* 37:103–116.
- Guss JM, Freeman HC. 1983. Structure of oxidized poplar plastocyanin at 1.6 angstroms resolution. *J Mol Biol* 169:521–563.
- Hampsey DM, Das G, Sherman F. 1986. Amino acid replacements in yeast iso-1-cytochrome *c*. *J Biol Chem* 261:3269–3271.
- Hilser VJ, Freire E. 1996. Structure-based calculation of the equilibrium folding pathway of proteins. Correlation with hydrogen exchange protection factors. *J Mol Biol* 262:756–771.
- Hilser VJ, Freire E. 1997. Predicting the equilibrium protein folding pathway: Structure-based analysis of staphylococcal nuclease. *Proteins Struct Funct Genet* 27:171–183.
- Hvidt A, Nielsen SO. 1966. Hydrogen exchange in proteins. *Adv Protein Chem* 21:287–386.
- Hynes TR, Fox RO. 1991. The crystal structure of staphylococcal nuclease refined at 1.7 Å. *Proteins Struct Funct Genet* 10:92–105.
- Jaenicke R. 1991. Protein folding: Local structure, domain, subunits, and assemblies. *Biochemistry* 30:3147–3161.
- James E, Wu PG, Stites W, Brand L. 1992. Compact denatured state of a staphylococcal nuclease mutant by guanidinium as determined by resonance energy transfer. *Biochemistry* 31:10217–10225.
- Jeng MF, Englander SW. 1991. Stable submolecular folding units in non-compact form of cytochrome *c*. *J Mol Biol* 221:1045–1061.
- Kauzmann W. 1959. Some factors in the interpretation of protein denaturation. *Adv in Prot Chem* 14:1–63.
- Kay LE, Torchia DA, Bax A. 1989. Backbone dynamics of protein as studied by ^{15}N inverse detected heteronuclear NMR spectroscopy: Application to staphylococcal nuclease. *Biochemistry* 28:8972–8979.
- Kim KS, Fuchs JA, Woodward CK. 1993a. Hydrogen exchange identifies native-state motional domains important in protein folding. *Biochemistry* 32:9600–9608.
- Kim KS, Tao F, Fuchs J, Danishefsky AT, Housset D, Wlodawer A, Woodward C. 1993b. Crevice-forming mutants of bovine pancreatic trypsin inhibitor: Stability changes and new hydrophobic surface. *Protein Sci* 2:588–596.
- Kocher JPA, Rooman MJ, Wodak SJ. 1994. Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *J Mol Biol* 235:1598–1613.
- Koide S, Dyson HJ, Wright PE. 1993. Characterization of a folding intermediate of apoplastocyanin trapped by proline isomerization. *Biochemistry* 32:12299–12310.
- Kosen PA, Creighton TE, Blout ER. 1981. Circular dichroism spectroscopy of bovine pancreatic trypsin inhibitor and five altered conformational states. Relationship of conformation and the refolding pathway of the trypsin inhibitor. *Biochemistry* 20:5744–5754.
- Kraulis P. 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structure. *J Appl Crystallogr* 24:946–950.
- Kuroda Y. 1993. Residual helical structure in the C-terminal fragment of cytochrome *c*. *Biochemistry* 32:1219–1224.
- Levitt M. 1992. Accurate modeling of protein conformation by automatic segment matching. *J Mol Biol* 226:507–533.
- Linderström-Lang KU. 1955. Deuterium exchange between peptides and water. *Chem Soc (London) Spec Publ* 2:1–20.
- Lodish H, Baltimore D, Berk A, Zipursky SL, Matsudaira P, Dranell J. 1995. *Molecular Cell Biology*. New York: Scientific American Books.
- Loh SN, Prehoda KE, Wang J, Markley JL. 1993. Hydrogen exchange in unligated and ligated staphylococcal nuclease. *Biochemistry* 32:11022–11028.
- Loll PJ, Lattman EE. 1989. The crystal structure of the ternary complex of staphylococcal nuclease, Ca^{2+} , and the inhibitor pdTp, refined at 1.64 Å. *Proteins Struct Funct Genet* 5:183–201.
- Louie GV, Brayer GD. 1990. High-resolution refinement of yeast iso-1-cytochrome *c* and comparisons with other eukaryotic cytochromes *c*. *J Mol Biol* 214:527–555.
- Makhatadze GI, Kim KS, Woodward C, Privalov PL. 1993. Thermodynamics of BPTI folding. *Protein Sci* 2:2028–2036.
- McQuarrie DA. 1976. *Statistical mechanics*. New York: Harper Row.
- Miyazawa S, Jernigan RJ. 1996. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol* 256:623–644.
- Molday RS, Englander SW, Kallen RG. 1972. Primary structure effects on peptide group hydrogen exchange. *Biochemistry* 11:150–158.
- Moore JM, Case DA, Chazin WJ, Gippert GP, Havel TF, Powls R, Wright PE. 1988. Three-dimensional solution structure of plastocyanin from the green alga *scenedesmus obliquus*. *Science* 240:314–317.
- Moore JM, Lepre CA, Gippert GP, Chazin WJ, Case DA, Wright PE. 1991. High-resolution solution structure of reduced French bean plastocyanin and comparison with the crystal structure of poplar plastocyanin. *J Mol Biol* 221:533–555.
- Naor D, Fischer D, Jernigan RL, Wolfson HJ, Nussinov R. 1996. Amino acid pair interchanges at spatially conserved locations. *J Mol Biol* 256:924–938.
- Oas TG, Kim PS. 1988. A peptide model of a protein folding intermediate. *Nature* 336:42–48.
- Orengo CA, Jones DT, Thornton JM. 1994. Protein superfamilies and domain superfolds. *Nature* 372:631–634.
- Perrett S, Clarke J, Hounslow AM, Fersht AR. 1995. Relationship between equilibrium amide proton exchange behavior and the folding pathway of barnase. *Biochemistry* 34:9288–9298.
- Richarz R, Sehr P, Wagner G, Wüthrich K. 1979. Kinetics of the exchange of individual amide protons in the basic pancreatic trypsin inhibitor. *J Mol Biol* 130:19–30.
- Schulz GE, Schirmer RH. 1979. *Principles of protein structure*. New York: Springer-Verlag.
- Shortle D. 1995. Staphylococcal nuclease: A showcase of *m*-value effects. *Adv Prot Chem* 46:217–247.
- Sippl MJ. 1990. Calculations of conformational ensembles from potential of mean force. An approach to the knowledge-based prediction of the local structure in globular proteins. *J Mol Biol* 213:859–883.
- Sippl MJ. 1993. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J Comp-Aided Mol Desg* 7:473–501.

- Sippl MJ. 1995. Knowledge-based potentials for proteins. *Curr Opin in Struct Biol* 5:229–235.
- Sippl MJ, Weitckus S. 1992. Detection of native-like models for amino acid sequences of unknown three-dimensional structures in a database of known protein conformations. *Proteins Struct Funct Genet* 13:258–271.
- Sosnick TR, Mayne L, Englander SW. 1996. Molecular collapse: The rate-limiting step in two-state cytochrome *c* folding. *Proteins Struct Funct Genet* 24:413–426.
- Staley JP, Kim PS. 1992. Complete folding of bovine pancreatic trypsin inhibitor with only a single disulfide bond. *Proc Natl Acad Sci USA* 89:1519–1523.
- Takano T, Dickerson RE. 1981a. Conformation change of cytochrome *c*. I. Ferrocyanochrome *c* structure refined at 1.5 Å resolution. *J Mol Biol* 153:79–94.
- Takano T, Dickerson RE. 1981b. Conformation change of cytochrome *c*. II. Ferrocyanochrome *c* refinement at 1.8 Å and comparison with the ferrocyanochrome structure. *J Mol Biol* 153:95–115.
- Vincent JP, Lazdunski M. 1972. Trypsin-pancreatic trypsin inhibitor association. Dynamics of the interaction and role of disulfides bridges. *Biochemistry* 11:2967–2977.
- Vincent JP, Peron-Renner M, Pudles J, Lazdunski M. 1974. The association of anhydrotrypsin with the pancreatic trypsin inhibitor. *Biochemistry* 13:4205–4211.
- Wagner G, Stassinopoulou CI, Wüthrich K. 1984. Amide-proton exchange studies by two-dimensional correlated ¹H NMR in two chemically modified analogs of the basic pancreatic trypsin inhibitor. *Eur J Biochem* 145:431–436.
- Wang Y, Shortle D. 1995. The equilibrium folding pathway of staphylococcal nuclease: Identification of the most stable chain-chain interaction by NMR and CD spectroscopy. *Biochemistry* 34:15895–15905.
- Weissman JS, Kim PS. 1992. The pro region of BPTI facilitates folding. *Cell* 71:841–851.
- Wlodawer A, Walter J, Huber R, Sjölin L. 1984. Structure of bovine pancreatic trypsin inhibitor. Results of joint neutron and X-ray refinement of crystal form II. *J Mol Biol* 180:301–329.
- Wodak SJ, Rooman MJ. 1993. Generating and testing protein folds. *Curr Opin in Struct Biol* 3:247–259.
- Wu LC, Grandori R, Carey J. 1994. Autonomous subdomains in protein folding. *Protein Sci* 3:369–371.
- Wu LC, Laub PB, Elöve GA, Carey J, Roder H. 1993. A noncovalent peptide complex as a model for an early folding intermediate of cytochrome *c*. *Biochemistry* 32:10271–10276.