Supplementary Material for Tsai *et al*. (2000) Proc. *Natl. Acad. Sci. USA* **97** (22), 12038-12043.

Methods

The Rationale. According to the building block folding model, a native protein conformation is the outcome of a combinatorial assembly of a set of contiguous fragments (building blocks). The building blocks exist in an ensemble of conformations. During the combinatorial assembly process, the building block conformations selected for binding might be those with low population times, differing from the conformations observed for the corresponding peptides in solution. However, owing to the mutual stabilization of the building blocks in the assembly process, low stability stand-alone conformations may produce more stable hydrophobic folding units, as compared with the assembly of the more highly populated conformers. Nevertheless, whereas such lower-stability, lower-population-times building block conformers may be selected, most of the associating building blocks are the stable, high-population-times conformers. During the folding process, these conformers are preserved and are observed in the final, “static” native 3D structure.

It is reasonable to assume that a stability measurement of an isolated contiguous fragment in a particular conformation will reflect the population time of the conformation during the folding process. The larger the measurement, the larger the time that the fragment spends in this conformation. Hence, to locate building blocks from a native protein 3D structure, we need to find a set of nonoverlapping contiguous fragments that have the highest conformational stability among all other possible candidate combinations. In practice, a few-residue overlap between the fragments is permitted.

For a polypeptide chain with a size of N_e residues, and with a size limit of a building block set to N_s residues, the total number of candidate fragments is $N_{\text{total}} = \sum(N_e - N_i + 1)$, where N_i runs the summation from N_s to N_e. A contiguous fragment along a sequence can be specified by two independent variables: size and position. The position of a fragment can be assigned by either the starting residue or the residue at the center (as done here). We call this two-dimensional coordinate system for all contiguous fragments in a given protein structure a “fragment map.” Note that both coordinates in a fragment map use a common unit (in residue), although one refers to the size of the fragment and the other to its position. Now, if a fragment map is associated with a scoring function that can reflect the conformational stability of each fragment, regardless of its size, then the local minima of the map are the locations of the building blocks of a given protein.

The Scoring Function. The solvent accessible surface area (ASA) of a protein is calculated by following the Lee and Richards' definition (1). The algorithm of Shrake and Rupley (2) has been adopted to calculate the ASA. The solvent accessible surface area of a protein is calculated numerically by summing the ASA represented by discrete spherical points. These points are evenly distributed on a sphere with the van der Waals radius of the atom plus a probe ball of 1.4 Å in radius. We have adopted the look-up table approach (3) to speed up the calculation.

Compactness, $\underline{\boldsymbol{Z}}$. The ASA of a residue comprises of two quantities, $\mathrm{ASA}_{\mathrm{Buried}}$ and $\mathrm{ASA}_{\mathrm{Surf}}$, with the $\mathrm{ASA}_{\mathrm{Buried}}$ referring to the area buried by atoms belonging to other residues, and the $\mathrm{ASA}_{\mathrm{Surf}}$ referring to the exposed surface area. We use the Zehfus and Rose definition (4) to calculate the coefficient of compactness ($\underline{Z}$) of a protein fragment. $\underline{Z}$ is defined as the solvent accessible surface area of a fragment, divided by its minimum possible surface area, which is the surface area of a sphere with an equal volume of the fragment. Hence, $\underline{Z}$ is calculated as

$$\underline{Z} = \mathrm{ASA}_{\mathrm{Surf}} / (36\pi\mathrm{VOL}^{2})^{1/3}. \qquad \mathbf{[1]}$$

The $\mathrm{ASA}_{\mathrm{Surf}}$ is the exposed solvent accessible surface area of a fragment. The VOL is the corresponding volume, which is numerically calculated by an integration of all individual exposed solvent accessible surface areas.

Hydrophobicity, $\underline{\boldsymbol{H}}$. To calculate hydrophobicity, the ASA of a fragment is divided into two categories, nonpolar ASA (${}^{\mathrm{Non}}\mathrm{ASA}$) and polar ASA (${}^{\mathrm{Pol}}\mathrm{ASA}$), according to the atom type to which the ASA belongs. The hydrophobicity is defined as the fraction of the buried nonpolar area out of the total nonpolar area,

$$\underline{H} = {}^{\mathrm{Non}}\mathrm{ASA}_{\mathrm{Buried}} / ({}^{\mathrm{Non}}\mathrm{ASA}_{\mathrm{Buried}} + {}^{\mathrm{Non}}\mathrm{ASA}_{\mathrm{Surf}}) \qquad \mathbf{[2]}$$

where ${}^{\mathrm{Non}}\mathrm{ASA}_{\mathrm{Buried}}$ and ${}^{\mathrm{Non}}\mathrm{ASA}_{\mathrm{Surf}}$ are the buried and the exposed nonpolar ASA, respectively. Note that the hydrophobicity was not calculated as the fraction of the whole buried area, but just the buried nonpolar area.

Isolation, $\underline{\boldsymbol{I}}$. After a fragment is generated, the solvent accessible surface area ($\mathrm{ASA}_{\mathrm{B}\rightarrow\mathrm{E}}$) that was originally buried in the interior of a protein, but is exposed to the solvent after cutting, is calculated. To emphasize the hydrophobic effect, we exclude the polar ASA in the evaluation of the extent of the isolation of the fragment. Hence, the degree of isolation is expressed as

$$\underline{I} = {}^{\mathrm{Non}}\mathrm{ASA}_{\mathrm{B}\rightarrow\mathrm{E}} / \mathrm{ASA}_{\mathrm{frag}} \qquad \mathbf{[3]}$$

where ${}^{\mathrm{Non}}\mathrm{ASA}_{\mathrm{B}\rightarrow\mathrm{E}}$ is the nonpolar ASA that was originally buried in the interior of a protein but became exposed after cutting; $\mathrm{ASA}_{\mathrm{frag}}$ is the solvent accessible surface area of the isolated fragment. More detailed definitions and calculations regarding $\underline{Z}$, $\underline{H}$, and $\underline{I}$ are described in (5).

The scoring function gives an “empirical” statistical score, derived from a compilation of native structures. Although no direct physical term (except hydrophobicity) is used to reflect the absolute stability or free energy of a cut fragment, it is designed to give relative stabilities of fragments with variable sizes in a given native conformation. If the score of a fragment is better than the average score of the two best individual fragments that may compose it, the associated (uncut) fragment is considered more stable than the two separate fragments. Consider a functional dimer, trimer, or higher oligomer in which each isolated monomer is also stable. In our terminology, if the free energy of the associated oligomer is more favorable than the average free energy of individual monomers, the system prefers an oligomeric state. Although empirical, the fragment-size-independent scoring function we have designed is in light of this consideration.

While the cutting algorithm and the scoring function are sensitive to chosen empirical thresholds, their effects did not alter much the overall anatomy tree structure. Thus, the number of local minima in the fragment map is very sensitive to the parameter defining the area of a local region (7.5% of the size of a candidate fragment) as well as to the minimum fragment size of a building block (15 residues). Nevertheless, although the number of local minima increases substantially with a decrease in the area of local region or with a decrease in the minimum fragment size, the process of combinatorial assembly is consistent in the cutting of a node fragment. Further, whereas for highly similar protein structures our scoring function does not give the exact fragment range for the basket of local minima because side-chain orientations may differ, the differences are always within the allowed overlap length (7 residues, half of the minimum size of a building block). Therefore, this difference does not affect the process of combinatorial assembly.

The cutting procedure. *Step 1*: *Locating a basket of building blocks (relatively stable contiguous fragments).* For a given protein 3D structure, all fragment candidates are assigned a stability score calculated by the above building block scoring function. To collect a basket of building blocks, we locate all local minima on the fragment map. We define a local minimum in the simplest way: A local minimum must be the highest value in a defined local region. A local region can be specified in an absolute or relative way. In order to locate compact substructures in a fragment map, Zehfus (3) has defined an absolute local quasi-circle region, with a distance of 4 residues from a candidate fragment. However, we find that a relative local region is more suitable for locating building blocks. We adopt the Zehfus quasi-circle definition as the area of a local region in our algorithm. However, the radius of the circle is a variable that is 7.5% of the size of a candidate fragment. Thus, for example, the local region of a 100-residue candidate fragment with its termini at least 7 residues away from the termini of the chain will contain a total of 197 fragments in which either their size or position are within 7 residues of the candidate. To locate all of the building blocks, every fragment candidate is examined to see whether it is the highest scoring within its local region as defined above. If a candidate is the highest scoring one, we register it in the basket of building blocks. In the case of actin (6) with 373 residues, 78 local minima with score > -5.0 have been located out of a total 64,620 fragments. A plot of the score surface of the fragment map is given in Fig. 4.

Step 2: A recursive top-down splitting process. Currently, it is generally agreed that the folding process does not follow a single pathway. In constructing an anatomy tree, our goals are twofold. First, an anatomy tree straightforwardly yields the most likely folding pathway(s). And second, we wish to identify the set of the most likely building blocks that, via a process of combinatorial assembly, form the final, native protein conformation. Hence, in our scheme, the anatomy of the protein structure is organized as a tree that grows upside down, with the starting node at the top. The root node is the first, starting node, and it constitutes the entire native protein. Each node represents a contiguous fragment. A node can sprout multiple branches to create child nodes. These nodes are generated via a multi-splicing procedure described below. If a new node does not produce a child, it is an end node. The level of a node is determined by counting the number of steps that are needed to trace back to the root node. The entire tree growth process stops when no new children nodes can be generated. At the end, the collection of end nodes is the set of the most likely building blocks, and the tree organization itself depicts the most likely folding pathway. Our success in generating such anatomy trees further indicates that the "building block" folding model leads to a hierarchical folding process.

The multi-splicing procedure. Unlike other top-down binary splicing algorithms (7,8,9), our recursive top-down multi-splicing procedure sets no limit on the number of branches at any level. Starting with a node fragment and a basket of building blocks created above, the search for multiple cutting is very simple: We search the basket for a set of fragments that constitute the entire node fragment. Then, among all combinatorially assembled candidates, the average score of the best two fragments (if there are more than two fragments) is used to rank the dissections for the node fragment. The search algorithm follows several rules. First, a short overlap between building blocks is allowed, with the overlapping segment not exceeding 7 residues. Second, if an unassigned segment is fewer than 15 residues, it is left unassigned. Otherwise, the segment is assigned to be a low-score building block, not listed in the original building block basket. A short unassigned segment may be a short linker between two large units. A long low-score fragment may be a long fragment linker or a building block that has opened up. Third, except for the root node, a node cannot have only one branch-child node. Fourth, a node is considered as an end node if we cannot find two building blocks with scores above a threshold value.

Assembly of hydrophobic folding units. Furthermore, in addition to the above, at each branching level, we locate the hydrophobic folding units (HFUs) by combinatorially assembling the collection of building blocks. This procedure is straightforwardly implied from the building block folding model (10). Preliminary results indicate a significant improvement compared with our previous HFU cutting algorithm (5). The detailed procedure will be reported elsewhere (C.-J.T. and R.N., unpublished data).

The Usefulness of the Anatomy Tree. Being able to construct an anatomy tree for any protein structure is particularly useful. First, by inspecting the trees we are able to see whether proteins fold through multiple routes. In such a multiple-route case, building blocks at different locations assemble separately; Only at later stages these units combine to form larger structural elements and ultimately the entire fold. The routes can be observed by inspecting the fragment map, with the local minima indicated and the lines connecting them depicting the most likely pathways (e.g., Fig. 5 *A*). Alternatively, we can examine the more detailed map of Fig. 5 *B*. Analysis of such plots immediately indicates that protein folding involves multiple pathways, rather than a single path, with a step-by-step addition of a single building block at a time.

Second, anatomy trees narrate a sequential vs. a nonsequential folding pathway story of the protein. Inspection of a fragment map, such as in Fig. 5 *A*, immediately suggests what is the likelihood that the protein is a sequential, or a nonsequential folder. In general, if more than two branches descend from a node, the protein is likely to manifest nonsequential folding already at that nodelevel. The larger the number of branches growing from a given node, the higher the chances that the protein has a complex fold. A fragment map, which consistently shows binary branches at every node, such as in the case of hemopexin (1hxn) in Fig. 7 *A*, immediately suggests that the protein is a sequential folder. As we shall see below, this suggestion is verified in Fig. 7 *B* and *C*. On the other hand, a plot such as that observed for the case of the ribosomal protein L1 (1ad2) in Fig. 8 *A*, suggests that this protein is likely to fold in a nonsequential way to produce a complex fold. However, from the fragment map alone, for the multiple-branches case, in order to verify that this is the situation, we need to inspect the detailed pathways, as in Fig. 8 *B*, and the corresponding structures, in Fig. 8 *C*.

Third, at the end of the dissection procedure, we may climb up, to combinatorially assemble the building blocks, into the compact, independently folding hydrophobic folding units. Inspection of the detailed anatomy trees, along with the hydrophobic folding units into which the building blocks are assigned via the combinatorial assembly, indicates whether the protein is a sequential or a non-sequential folder, regardless of the number of branches that descend from a node. Fig. 7 *B* illustrates clearly the situation for 1hxn. The letters in parentheses, showing the specific HFU assignments, combined with the tree, verify that this protein folds via a sequential pathway. At every level in this figure, the corresponding hydrophobic folding unit contains two sequentially connected fragments. Inspection of the detailed anatomy tree for 1ad2 in Fig. 8 *B*, shows that as suspected from the multiple branches shown in Fig. 8 *A*, 1ad2 has a complex fold. Thus, for example, the HFU labeled as B in Fig. 8 *B* contains both covalently, as well as noncovalently connected building block fragments. We see fragments 1-31; 29-67; then 159-208 and 209-224. On the other hand, fragment 68-153 forms a compact, independent unit on its own, labeled here as A. Fig. 7 *C* and 8 *C* illustrate the step-by-step process visually, coloring the fragments, both for the dissection into the building blocks (top of each figure) and their assembly into HFUs. Again, the difference between the two sequential and nonsequential types is immediately apparent.

Fourth, anatomy trees straightforwardly suggest which proteins are fast folding chains. Fast-folding proteins are likely to be sequentially folding proteins. In sequentially folding proteins, building blocks which are connected on the chain interact with each other. Since these building block elements are next to each other, a fold in which they interact involves a shorter search time. On the other hand, folds in which the interacting building blocks are not covalently connected, would still initially go through the connected-interaction trials. Such folding routes, which culminate in complex folds, take a longer time to fold, and are more prone to errors, that is, to misfolding. Hence, the anatomy trees provide a clue as to the likelihood of misfolding.

Fifth, here we illustrate only the predominant pathway down the funnel slope. That is, at each barrier that the down-gliding conformation encounters, we choose, and depict, the most likely one. Yet, as all fragments have been generated, and the stability of each assessed, the anatomy trees can narrate the story of the folding funnels. Specifically, the number of mutually-consistent local minima correlate with the number of pathways; And, the stability of the fragment at the minima indicates the probability associated with the route the chain would choose to descend. If the fragment has a higher stability score, the folding pathway has a higher chance of gliding in its direction. Hence, by weighing the relative scores at each node, we can choose alternate routes. Fast folders may be expected to have a predominant folding pathway.

Sixth, through inspection of the building blocks in the fragment map, we may obtain an insight into folding intermediates, trapped along the funnel walls. The conformers residing in these wells largely represent native building blocks in their native conformations. However, their associations involve non-native contacts. These may involve either the building blocks in the most traveled paths, or alternate ones, also showing up in our local minima map-listing. While these are misfolded, trapped intermediates, they nevertheless serve a role in the folding of the chain. The traps catch these conformers, and thereby aid in increasing the concentration of the native conformations of the building blocks. As the non-native contacts break, the conformers climb out of their wells, and, through the combinatorial assembly process, finally attain the native 3D shape of the protein. Thus, these traps are on- rather than off-pathway in the folding process.

References

1. Lee, B. & Richards, F. M. (1971) *J. Mol. Biol.* **55**, 379-400.

2. Shrake, A. & Rupley, J. A. (1973) *J. Mol. Biol.* **79**, 351-371.

3. Zehfus, M. H. (1993) *Proteins* **16**, 293-300.

4. Zehfus, M. H. & Rose, G. D. (1986) *Biochemistry* **25**, 5759-5765.

5. Tsai, C. J. & Nussinov, R. (1997) *Protein Sci.* **6**, 24-42.

6. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol.* **112**, 535-542.

7. Lesk, A.M. & Rose, G.D. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 4304-4308.

8. Wodak, S. J. & Janin, J. (1981) *Biochemistry* **20**, 6544-6552.

9. Rose, G. (1979) J. Mol. Biol. 134, 447-470.

10. Tsai, C. J., Kumar, S., Ma, B. & Nussinov, R. (1999) *Protein Sci.* **8**, 1181-1190.