A hierarchical, building-block-based computational scheme for protein structure prediction

by C.-J. Tsai
B. Ma
Y. Y. Sham
S. Kumar
H. J. Wolfson
R. Nussinov

Protein folding is a hierarchical event in which transiently formed local structural elements assemble to yield the native conformation. We outline the hierarchical *building block protein folding model*, which is based on two premises. First, while the local building block elements may be unstable, they nevertheless have higher population times than all alternate conformations; second, protein folding progresses through a combinatorial assembly of these elements. In accordance with this model, we describe a building block cutting algorithm, implementing its rationale. Through its automated iterative application to the native structure we obtain an *anatomy tree*, in terms of protein folding. The anatomy tree automatically yields the most likely folding pathway. In particular, we describe how, by using this algorithm and the building blocks which are obtained, we expect to reduce substantially the computational time involved in simulations of protein folding.

Introduction

The way in which a polypeptide chain folds to assume its three-dimensional (3D) shape is a fascinating problem, which is captivating to understand and solve from a purely intellectual standpoint and also has extremely important practical consequences. Despite the fact that for nearly forty years it has occupied a central place in chemistry, biophysics, and branches of the experimental and computational sciences in general, it still presents a major obstacle to understanding. Nevertheless, over the years progress has been made in improving the methodology for predicting protein structure from its sequence (e.g., [1]).

On the technical computational side, there are three basic approaches to the prediction of protein structure: homology modeling, threading, and *ab initio* folding. *Homology modeling* is the method with the highest success rate; it is highly reliable in cases where there is another sequence, or better still, several sequences, which are highly similar to the sequence whose structure is sought, and the structures of these other sequences have been solved. In the absence of such a situation, the method of choice to try is *threading*, or *inverse folding*. Here one "threads" the protein chain through available folds, searching for a closely related sequence. The success in threading is, in general, a function of the sequence similarity. The more closely related the sequences, the higher the chances of producing a reasonably folded chain. The extent of sequence similarity sought in this method is lower than the one used in homology modeling. The details of these two methods, and their successes and

0018-8646/01/$5.00 © 2001 IBM

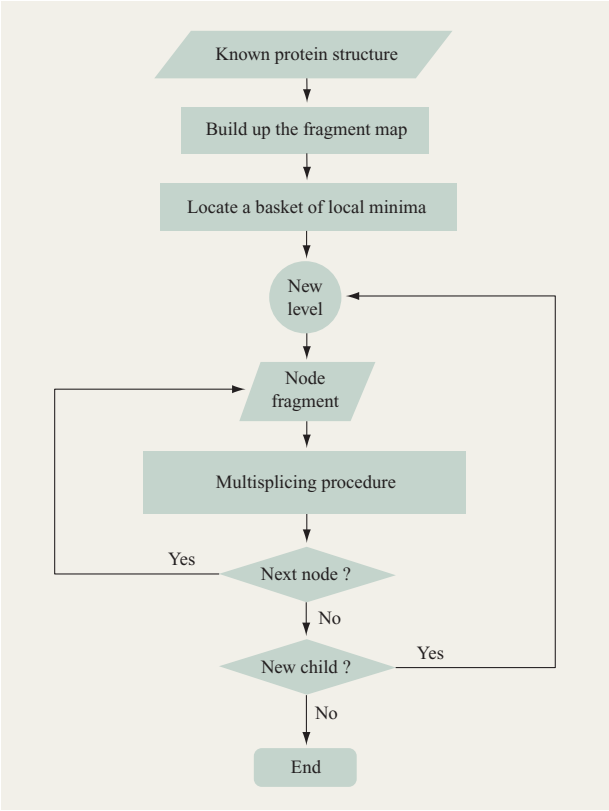

Figure 1

Flow chart of the dissection procedure that produces the building blocks.

failures, are given in a special issue of the journal *Proteins*, which was devoted to the recent CASP competition (Critical Assessment of Structure Prediction, *Proteins*, Volume 37, S3, 1999). If, however, no sequence similarity with a sequence whose structure has been solved is available, practically the only way to predict its folded shape is via simulations. However, unbiased simulation of protein folding with even 36 residues in length is currently infeasible, owing to the immense computation time required [2].

Here we describe a hierarchical protein folding model which makes it possible to visualize how a 1D protein chain folds into its 3D native state. We propose that through the application of such a model, the computational time for folding the protein could, in principle, be substantially reduced. We first describe this building block folding model and its principles. Its consistency with current experimental and theoretical results has already been described [3, 4]. We next present the algorithm itself, describing the procedure which is involved. We proceed to illustrate some examples of the anatomy trees produced through application of this model, showing the major folding pathway for each of these proteins. We present some of the results of the simulations of building blocks, showing that they are relatively stable. Finally, we describe our scheme as to how this model could be used to substantially reduce the computational time in folding. This part of the work is currently in progress.

For the purpose of clarification, **Figure 1** presents a schematic flow chart of our scheme. The figure illustrates the dissection procedure that produces the building blocks. The process is iterative, progressively cutting the native structure into shorter fragments by using a statistically based scoring function which calculates the stability of candidate building blocks [4]. This procedure has already been implemented. All native structures in the protein structure database (PDB) [5] have been cut, and their building block fragments are available (*http://protein3d.ncifcrf.gov/tsai/anatomy.html*).

The building block folding model: Principles and rationale

The building block folding model is a "practical" model for protein folding [3, 6]. The model states that protein folding is a hierarchical process [7], and that the units from which protein folds are constructed, i.e., the hydrophobic folding units (HFUs), are the outcome of a combinatorial assembly process of sets of building blocks. The HFUs subsequently associate to produce intramolecular domains. These, in turn, assemble to construct an intramolecular multidomain protein fold or, alternatively, an intermolecular quaternary structure. The "building block" itself is defined as a highly populated, contiguous fragment in a given protein structure. It may be composed of a single secondary structure element or a fragment consisting of chain-linked, interacting structural elements, such as those observed in supersecondary structures [8, 9]. This, however, is not the case for the hydrophobic folding unit. Such a unit is defined as an independent, compact, thermodynamically stable folding unit with a buried hydrophobic core [10, 11]. The "foldon" approach [12], in which a protein is built from an assembly of foldons (compact, independently folding units), also corresponds nicely to the building block concept.

The building block folding model is based on the idea that the native fold dictates the folding pathway. Hence, this implies that if we were to cut the building block from the protein chain and place it as a peptide in solution, the most highly populated conformation of this peptide would very likely be similar to that of the building block when it is in the native protein fold. Still, as we see later, a major difficulty in the protein folding prediction scheme derives from the fact that while the conformations of most

building blocks are preserved in the final native folded structure, this is not always the case. The mutually stabilizing association among the building blocks may result in alternate conformations being selected in the combinatorial assembly. Hence, in such cases, the conformations of the building blocks that we observe in the native protein structure will differ from their original stand-alone conformations.

Our algorithm is similar to the methods devised almost twenty years ago by Lesk and Rose [13] and by Wodak and Janin [14] to locate building blocks in a given protein tertiary structure. However, in contrast to those original methods, we do not confine ourselves to binary cuttings of the polypeptide chain; instead, the algorithm we have developed allows multiple dissections at each iterative level, resulting in a descending hierarchy of contiguous fragments. Each node in the anatomy tree is a one-segment building block, and the entire native structure of the protein is the starting root-node of this anatomy tree. The locations of the building blocks correspond to the end nodes of the top-down sprouting tree.

To be able to dissect the protein structure and create an anatomy tree, one must utilize a scoring function that is independent of the building block fragment size. We have constructed such a statistically based scoring function. This scoring function has been devised to measure the relative conformational stability of all candidate building blocks. This empirical fragment-size-independent scoring function is based on three elements: measurements of compactness, degree of isolation, and hydrophobicity. Compactness and isolation correspond to the "classical" visual criteria of a domain; on the other hand, hydrophobicity is well known to be the dominant driving force in protein folding [15]. We progressively cut the protein chain into sets of fragments with the highest stability score. Hence, at the end of the cutting process, the resulting anatomy tree straightforwardly yields the most likely folding micropathway. Furthermore, while the anatomy tree itself outlines the more probable folding routes, the minima among the cut-out fragments yield the number of alternate routes and their description. These are the less probable folding pathways. Further, trapped intermediates may be inferred. These may largely consist of misassociated, highest-population-time building blocks, or alternate local-minima building blocks, which are also present in our building block fragment pool. These correspond very nicely to experimental fragment CD and fluorescence spectra results (e.g., [16]).

Figure 2 [17, 18] illustrates two examples of the concept of the dissecting algorithm that progressively cuts the protein structure into building blocks and reveals its anatomical features. This figure presents side by side two proteins that share a common fold; these are inorganic pyrophosphatases (PDB codes: 2prd and 1ino). One protein is from a mesophilic organism (2prd), and the other is from a thermophilic organism (1ino). The melting temperatures (and stabilities) of the two proteins are substantially different from each other. As the figures show, the native state topologies of both proteins are similar; however, the sequence similarity is relatively low (48.5%). In particular, the figures show that despite this dissimilarity in sequence, the pattern of cutting and the anatomy trees are similar, illustrating that topology, rather than the details of the protein atomic interactions, largely determines the major folding pathway of the protein. In particular, this example further validates the notion that the native conformation and the native interactions present in the native state determine the folding pathways. We return to these crucial points below.

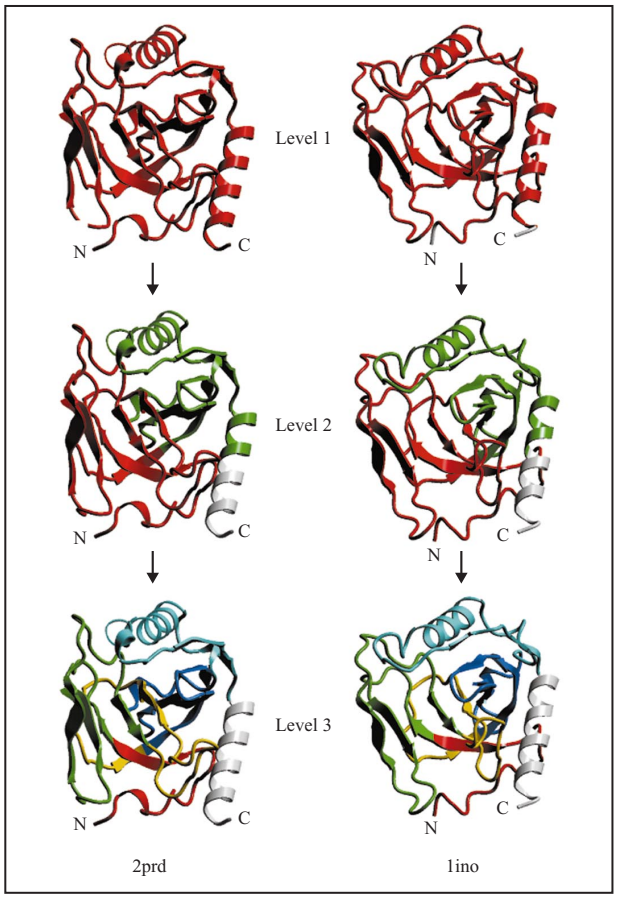

Figure 2

Two examples of the concept of the dissecting algorithm. The two protein structures are progressively cut into building blocks to reveal their anatomical features. The figure presents side by side two proteins sharing a common fold. The pair shown comprise a thermophile and its corresponding mesophile; they are inorganic pyrophosphatases with the PDB codes 2prd and 1ino [5], which were originally obtained from [17] (2prd) and [18] (1ino).

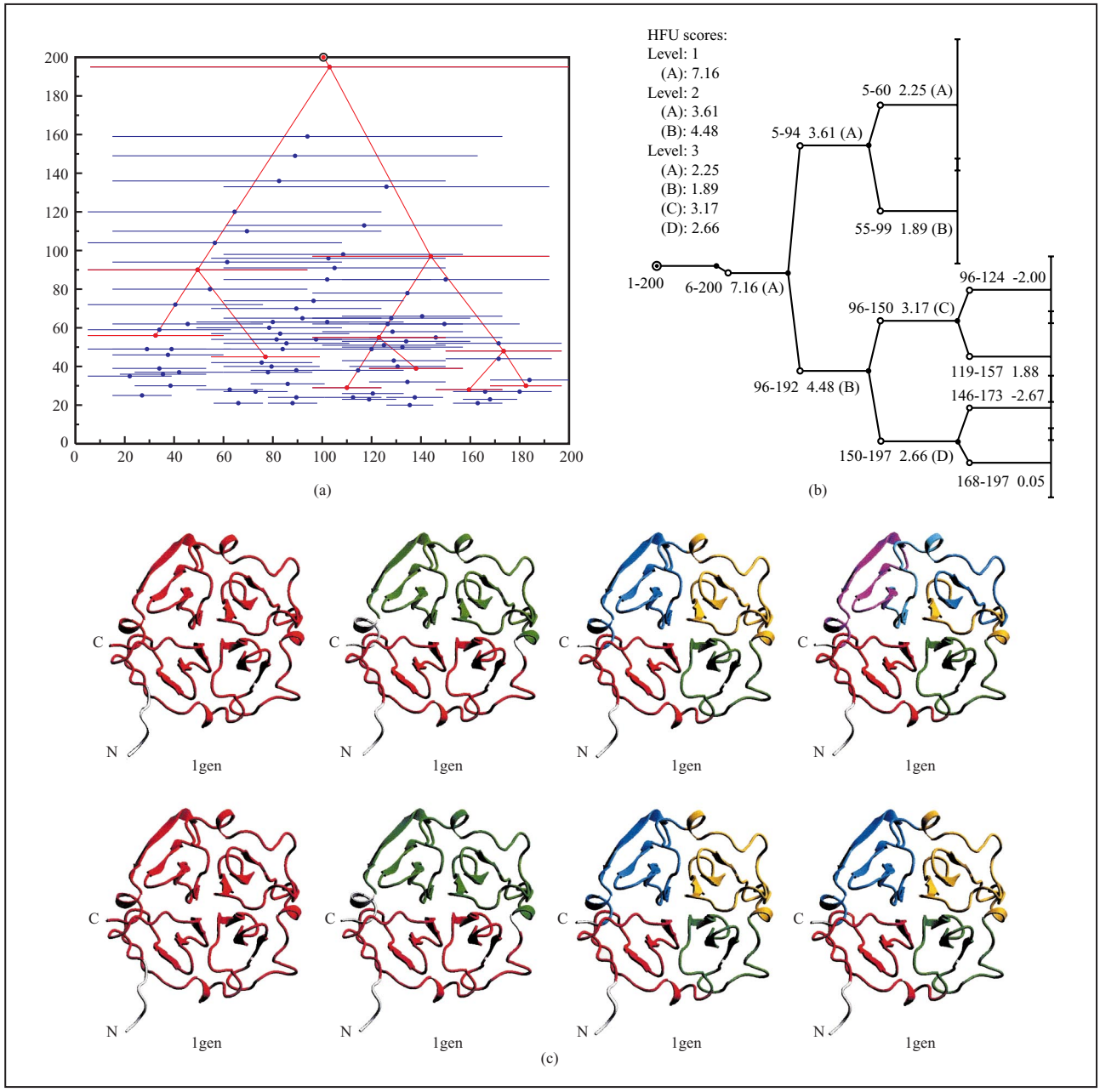

Figure 3

Building block cuttings and anatomy tree for a simple, sequentially folding protein, the C-terminal domain of Gelatinase A (1gen): (a) fragment map; (b) anatomy tree; (c) building block cuttings and their combinatorial assembly into hydrophobic folding units. In part (a), the x and y coordinates represent the fragment location and size. Local minima in the map are indicated by solid circles. The associated horizontal line for each minimum reflects its size. The detailed results of anatomy [red lines in part (a)] are given in the anatomy tree shown in part (b). Part (c) is a graphical representation of the anatomy tree at each level. The upper row shows the results of the building block assignments and the lower row their corresponding HFU assignments. Part (c) was obtained from the PDB [5] and was originally published in [19].

The information contained in a correct dissection of a protein tertiary structure into highly populated or stable building block components is likely to prove very useful: First, this information allows analysis and assessment of the folding complexity, i.e., classifying a protein structure and a folding pathway in terms of sequential/nonsequential folding in a more precise manner (**Figure 3** [19] and **Figure 4** [20], respectively) [21]. From the folding complexity we may

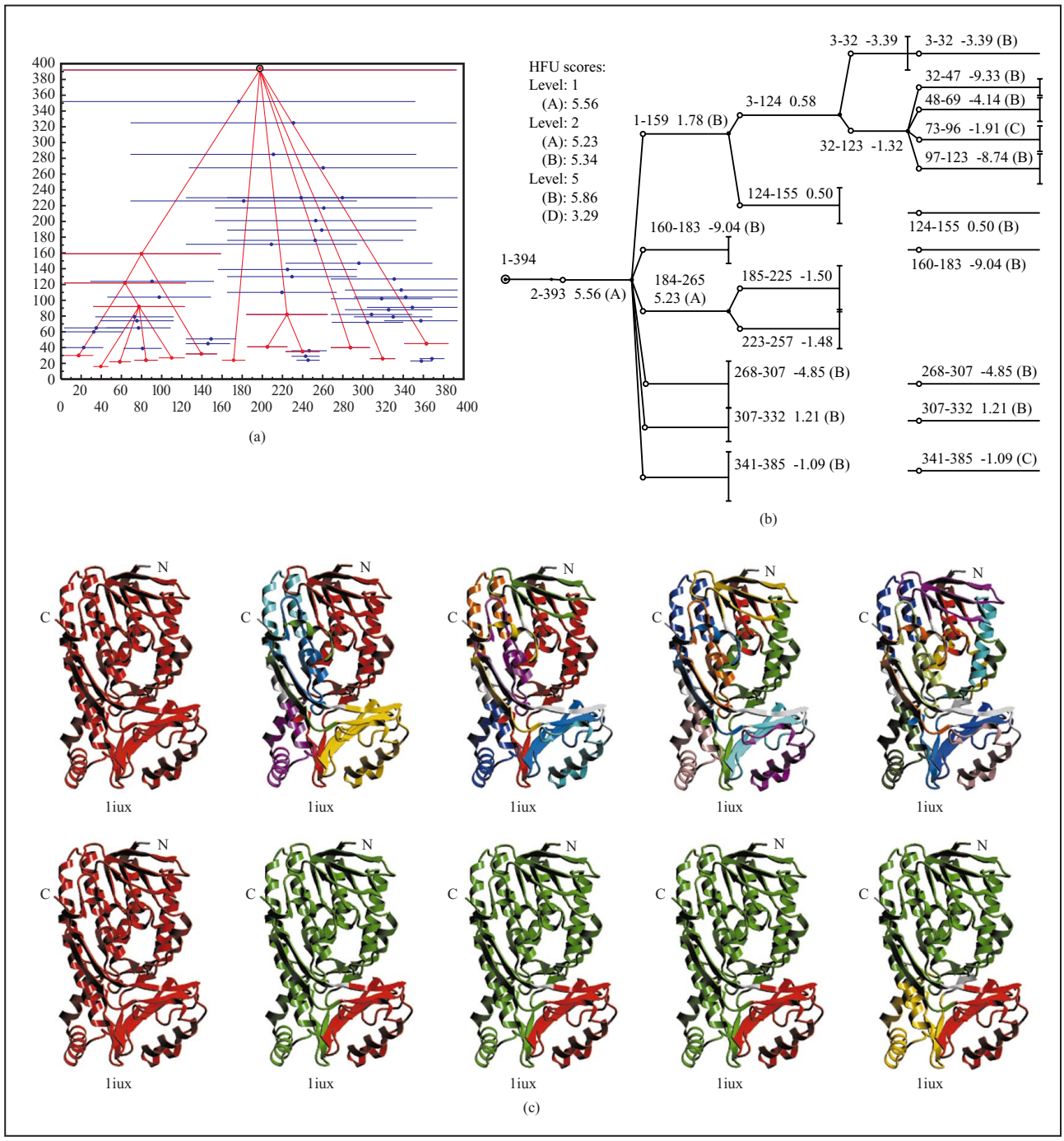

Figure 4

Building block cuttings and anatomy tree for a nonsequentially folding protein, the p-Hydroxybenzoate hydroxylase (1iux): (a) fragment map; (b) anatomy tree; (c) building block cuttings and their combinatorial assembly into hydrophobic folding units. See Figure 3 for the corresponding description. Part (c) was obtained from the PDB [5] and was originally published in [20].

infer the likelihood of misfolding and estimate whether the protein is a fast or a slow folder. In particular, this information can be obtained by analyzing the arrangement of building blocks in the native protein fold at the different levels of cuttings, and the way they associate with one another. Building blocks which are critical for correct

Figure 5

Critical building block. Building blocks at the second level of cutting in *Saccharomyces cerevisiae* adenylate kinase (PDB code 1aky [5, 22]), with each color representing a different building block. Shown in red is a critical building block that occurs at the N-terminus of adenylate kinase; it contains the ancient P-loop (a.k.a. giant anion hole) of adenylate kinase [23].

folding may be inferred from the anatomy tree (**Figure 5** [22, 23]), and likely intermediate states resulting from their "flipping out" during folding may be predicted. Second, the dissection yields a dataset of protein fragments. Such a library contains fragments ranging from complete tertiary folds to short pieces of the chain, with their associated favorable conformations. These provide a rich and very useful resource for secondary structure, or for *ab initio* tertiary structure prediction [24]. Third, we further note that since partial threading has proven useful in protein fold recognition, the availability of a complete, nonredundant library of known contiguous fragments, along with their likely conformations, should be particularly helpful in the correct prediction of protein structure.

The cutting algorithm

According to the building block folding model, the native conformation is the outcome of a combinatorial assembly of a set of building block fragments. In solution, the building blocks exist in an ensemble of conformations. Most of the associated building blocks are the stable, high-population-time conformers. During the folding process these conformers are preserved, and they are observed in the final native 3D structure. Nevertheless, as the building blocks assemble during the combinatorial assembly process to yield higher-order structures, the conformations which are selected are not necessarily those that have the highest population times in solution. The mutual stabilization of the building blocks may stabilize those conformations which as stand-alones are less stable and hence have lower population times. Owing to the mutual stabilization effect, the outcome of such an assembly process may be more stable hydrophobic folding units than those which would have been produced by the higher-population-time building blocks. While such a lower-population-time building block conformational selection scenario may occur relatively infrequently, with more highly populated conformers usually selected during the combinatorial assembly, it may nevertheless take place. It is this possibility which makes protein folding ventures so difficult.

To cut the native structure into a set of building blocks, we have devised a scoring function which estimates the stability of a candidate building block fragment. In order to be truly useful, such a scoring function must be *independent of the length of the building block*. The scoring function should reflect the population time of the building block in solution: the larger the stability, the higher the population time. The critical point about the scoring function is that it is size-independent. It is not biased by the length of a fragment when estimating its stability measurement. Therefore, two building blocks with the same score but of different lengths are not distinguished by the cutting algorithm.

Our goal is to locate building blocks in a native protein structure. To do that, we must find a set of non-overlapping fragments possessing the highest conformational stability of all possible candidate combinations. (In practice, a few residue overlaps between the fragments are permitted.) For a polypeptide chain with a size of N_e residues and with the size limit of a building block set to N_s residues, the total number of candidate fragments is $N_{\text{total}} = \Sigma(N_e - N_i + 1)$, where N_i runs the summation from N_s to N_e. Every fragment in the sequence is specified by two independent variables: size and position. The position of a fragment is assigned by the residue at the center of the fragment. The two-dimensional coordinate system (specified by the size of the fragment and its location in the protein chain) for all contiguous fragments in a given protein structure is a "fragment map." By associating with such a fragment map a scoring function that reflects the conformational stability of every fragment, the local minima in the map are automatically the locations of the building blocks of a given protein.

The scoring function has been described in detail elsewhere [4]. It is based on a previous scoring function

which has been successfully applied to locate hydrophobic folding units [10]. The hydrophobic folding unit scoring function comprises four elements: compactness, hydrophobicity, degree of isolatedness, and number of segments. Since a building block has by definition only one segment, only the first three ingredients are used here. The new function is expressed as a linear combination of these three measurements, with each quantity calculated as the deviation from the averaged value of known protein structures. The corresponding arithmetic averages of these quantities, and the standard deviations, are determined from a nonredundant dataset of 930 representative single-chain proteins from the protein structure data bank [5]. We have calculated the average and the standard deviation both with respect to fragment size and as a function of the fraction of the whole protein represented by the fragment. Their summation yields the scoring function.

There are two steps in the cutting procedure (Figure 1): In the first step, we locate a basket of building blocks (i.e., all relatively stable fragments). We do this by assigning a stability score to each trial fragment and collecting a basket of building blocks by locating all local minima on the fragment map. In general, lower energy values imply higher stability, and a local minimum refers to the lowest-energy fragment in the local region. However, our scoring function gives the more stable fragment a higher value; therefore, we define a local minimum as the highest value in a defined local region [4]. In the second step we perform a recursive top-down splitting process. (The "top" is the native structure with the cutting iteratively progressing to smaller structural units.)

The folding process does not follow a single pathway. Hence, in constructing an anatomy tree, we have two goals: First, the anatomy tree should straightforwardly yield the most likely folding pathway(s); second, the iterative cutting process should identify the most likely set of building blocks. These building blocks should, via a process of combinatorial assembly, form the native protein conformation. Thus, we organize the anatomy tree that we construct for the protein structure as a tree which grows upside down, with the starting root node of the native structure at the top. Each node represents a contiguous fragment. Via a multi-cutting procedure, multiple branches can originate from a single node. 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 which are needed to backtrack to the root node. The entire tree growth process stops when no new children nodes can be generated. The collection of end nodes is the set of the most likely building blocks, while the tree organization itself yields the most likely folding pathway.

Hence, in essence, our cutting algorithm sets no limit on the number of branches at any level: Starting with a node fragment and the basket of building blocks as described above, we carry out the search for multiple cutting. We sift through the basket, looking for a set of fragments that constitute the entire node fragment. In the search, we first allow a short overlap between building blocks (up to seven residues); second, if an "unassigned" segment is less than 15 residues, it is left unassigned. Otherwise, the segment is assigned to be a low-score building block. A short unassigned segment may be a linker between two building blocks. A long low-score fragment may be a conformationally unstable building block that has opened up. Third, with the exception of a root node, a node cannot have only one branch-child node. Fourth, a node is considered to be an end node if we cannot find two building blocks with scores above a defined threshold value. This last criterion is the stopping condition for a branch node in the recursive top-down splitting process. Whether a node must be cut further appears to depend on the setting of the defined threshold value. Further description of the cutting algorithm can be found in Tsai et al. [4].

Figures 3 and 4 represent examples of the progressive stepwise cutting and hierarchical assembly of the building blocks of two proteins. Figure 3 illustrates a "simple," sequential pattern of folding, whereas Figure 4 presents a complex pattern. In the latter, complex folding case, building blocks which are not adjacent to one another in the chain are in tertiary contact, while the sequentially intervening building blocks are flipped out, to contact other parts of the structure. It is worth noting that the mutual stabilization effect among building blocks has been implicitly implemented in the cutting algorithm, with the high scores observed for the building blocks in the top level and low scores in the bottom anatomy level. It is expected that the use of secondary structure information in the scoring function may help slightly in defining a stable building block. We did not use it in our scoring function, since cutting in the middle of a secondary structure, especially a β-sheet, usually produces two bad, low-score building blocks. We further note that at each anatomy branch, the cut building blocks undergo a combinatorial assembly search to locate hydrophobic folding units which are not contiguous pieces in sequence.

Simulations of building blocks

To further explore the building blocks and their associations, we have carried out molecular dynamics simulations. Two types of simulations have been performed: In the first, we have simulated some building blocks which were obtained through the cutting algorithm. We have chosen to simulate building blocks with relatively low stability scores. One of these is the fragment (Leu23–Asn53) of immunoglobulin-binding

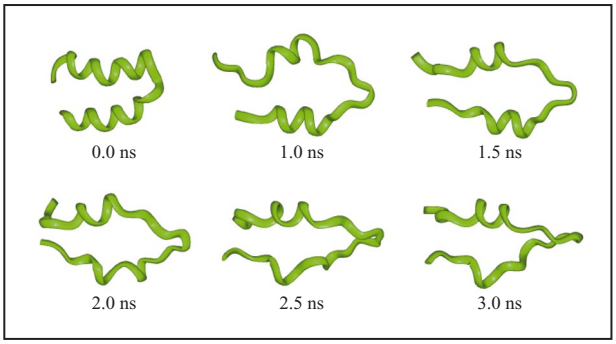

Figure 6

Snapshots from simulations of a building block (Leu23–Asn53 of immunoglobulin-binding protein, PDB code 1bdc [5]). The peptide is solvated with 2442 water molecules in a cubic box measuring 43 Å × 43 Å × 43 Å. The simulations were run at 273 K using DISCOVER 2.98 (MSI/Biosys) with CFF91 force field.

protein. This fragment has a cutting score of 1.3. **Figure 6** shows some snapshots extracted from its trajectory. These simulations were carried out at 273 K. As the figure shows, even these low-stability-score building blocks are relatively stable, implying reasonable population times. Figure 6 further illustrates that secondary structure is not necessarily formed prior to building block formation. This is an important difference between the building block folding model and the traditional hierarchical folding model, in which secondary structures form first.

In the second type of simulation, the amino-terminal building block has been cut out of the adenylate kinase protein from *Saccharomyces cerevisiae*. This building block (Ser1–Gly36) is a critical one (Figure 5), according to our definition (given in the next section). The simulation was carried out using the c27b1 version of the CHARMM program [25]. **Figure 7** provides the structures of the initial and final conformations of adenylate kinase in the absence of the critical building block. As the figure shows, without this building block the structure collapses upon itself. Hence, the two most important points here are, first, that in the absence of such a critical building block to mediate the interactions of most of the other building blocks, the native conformations of these other building blocks are retained; and second, that what is altered here is their association [27]. This is exactly as predicted by the building block folding model. Such a conformation may constitute an intermediate state.

Thus, both types of simulations consistently indicate that whether cut from the structure and simulated as a peptide (the first type of simulation, Figure 6), or remaining within the structure, but with a critical building block removed (Figure 7), the building blocks retain their native conformations.

Above we have described the building block folding model, the anatomy trees, and the practical procedure generating the building blocks through a multi-cutting process. Next, we describe how we envision its utilization to create more efficient folding routines.

Implications of the building block folding model for computational protein folding

Given the amino acid sequence of a protein chain, how can we use the building block folding model to substantially reduce computation time?

Three elements in the building block folding model are useful: First, we have a collection of building blocks, their sequences and their conformations, as observed in the native structure of the proteins from which they were derived. Furthermore, along with each building block we have a stability score, which provides an indication of the population time of the building block in solution. Second, the collection of building blocks is arranged according to the levels of the cuttings. In general, the closer to the root of the tree, the larger are the building block sizes. Third, through inspection of the way in which the building blocks interact in the native structure, we can identify those building blocks that occupy a critical position in the structure. A critical building block is in tertiary contact with a number of building blocks, burying an extensive amount of nonpolar surface area in these interactions. Further, a critical building block is involved in nonsequential interactions; i.e., it may be inserted between two sequentially connected building blocks, contacting both. The most important criterion in defining a critical building block relates to its overall mediating location. If it were to be pulled out of the structure, the other building blocks would alter their associations, collapsing to yield a stable though non-native conformation (Figure 7). Such a conformation may constitute an intermediate structure on the folding pathway. By inspection of the building block cuttings and the way they associate in the native structures, the critical building blocks can be identified.

When seeking to fold a protein chain in accordance with the hierarchical folding model, our first goal is to dissect the *sequence* into building block elements. Here our goal is twofold: Since according to this model folding is the outcome of a combinatorial assembly of building block fragments, we would like to cut it into as few building blocks as possible to reduce the combinatorics of their assembly. Further, the more stable the building block we identify, the higher its population time, and hence the smaller the number of alternate conformations we may need to deal with in the combinatorial assembly. In addition, if we could identify a critical building block

directly in the sequence, the combinatorial assembly process would be greatly shortened. This is particularly the case if the critical building block is a stable structure, similar to the so-called stable, high-population-time folding nuclei. Since we have a library of building blocks, and we are in possession of the information regarding the cutting levels in which they are produced in their corresponding anatomy trees, and since we also have a library of critical building blocks, all with associated stability scores, it appears that in principle we are well poised to tackle this task.

In practice, in order to carry out this extremely complex endeavor, we must be able to identify the building blocks in the sequences. To do this, we need to extract all available information from the collection of building blocks in our possession. This is the step on which we are currently focusing. We cluster the building blocks both by their conformations, i.e., according to the root mean squared distances between them, and by their topologies when they are more distantly related. Additionally, we consider their buried/exposed nature, stability, and size. Either the actual sequences are used, or their hydrophobic/polar (H/P) representation. In parallel, we cluster them by their similarity in their amino acid sequences. Here our goal is to obtain the range of variant conformations into which similar sequences can fold. This information will be particularly useful when alternate conformations are assigned to a sequence of a building block during the combinatorial assembly. Reducing the number of building blocks and the number of alternate conformations that each fragment can have will render the combinatorial assembly process computationally faster. However, we have not yet estimated their maximum allowed numbers to make the process feasible with current computational resources.

Conclusions

We have outlined the building block folding model and its rationale. The model is consistent with available experimental and computational data. This model is attractive in that it not only provides a logical and consistent framework for understanding protein folding, the major folding pathways, folding kinetics, and some potential intermediate states, but it further allows for a reduction in computational complexity.

The first part of this task has already been carried out. Every protein in the structure database has been cut, and its anatomy tree, building block fragments along with their stability scores, and major folding pathways have been elucidated and are available on line (*http://protein3d.ncifcrf.gov/tsai/anatomy.html*). We are currently extracting the information available in this rich collection in order to be able to apply it to protein sequences. Through hierarchical clustering of the building blocks at the different levels, and their comparisons, statistical information concerning building blocks and assembled hydrophobic folding units is currently being derived from a representative set of single-chain proteins.

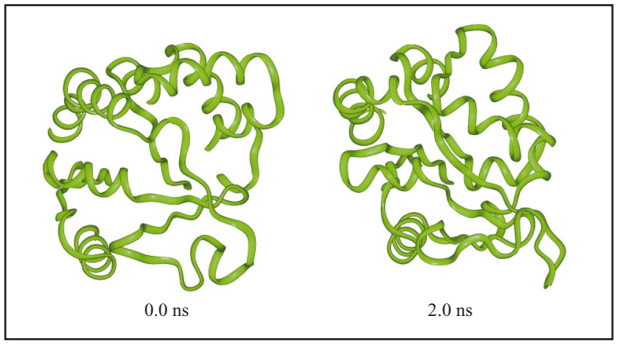

Figure 7

Structures of the initial and final conformations of the *Saccharomyces cerevisiae* adenylate kinase protein fragment. The N-terminus, consisting of residue 1-36, was removed from its native structure (PDB code 1aky [5]). The simulation was carried out with 1000 steps of adapted-basis Newton–Raphson (ABNR) minimization followed by 2 ns simulation at 2 fs time step at 300 K. The CHARMM 19 polar hydrogen potential energy function with effective implicit solvent energy function (EEF1) [26] was used.

It is particularly important to note the major difficulty that we foresee in the proposed folding scheme. The building blocks are conformationally fluctuating entities. While in most cases the conformation that we observe in the native state is the one that has the highest population time in solution, this is not always the case. Folding involves conformational selection. It is possible that during the combinatorial assembly, the conformation of the building block which is selected has a lower population time. Hence, assigning this conformation to a similar building block sequence during folding might result in an erroneously predicted conformation. While in principle the chances of such an occurrence might be estimated from the stability scores, it nevertheless still represents a major obstacle. Had it not been for this difficulty, there is a reasonable chance that the folding problem might have been solved years ago.

Acknowledgments

In particular, we thank Dr. Jacob V. Maizel for numerous helpful and very insightful discussions. The research of Drs. R. Nussinov and H. Wolfson in Israel has been supported in part by Grant No. 95-00208 from BSF, Israel, by a grant from the Israel Science Foundation administered by the Israel Academy of Sciences, by a Magnet grant, by a Ministry of Science grant, and by the Tel Aviv University Basic Research grants and the

Center of Excellence, administered by the Israel Academy of Sciences. This project was funded wholly or in part with federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. NO1-CO-56000. The content of this publication does not necessarily reflect the view or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organization imply endorsement by the U.S. government.

References

1. M. J. E. Sternberg, P. A. Bates, L. A. Kelley, and R. M. MacCallum, "Progress in Protein Structure Prediction: Assessment of CASP3," *Curr. Opin. Struct. Biol.* **9,** 368–373 (1999).
2. Y. Duan and P. Kollman, "Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution," *Science* **282,** 740–744 (1998).
3. C. J. Tsai, S. Kumar, B. Ma, and R. Nussinov, "Folding Funnels, Binding Funnels and Protein Function," *Prot. Sci.* **8,** 1181–1190 (1999).
4. C. J. Tsai, J. V. Maizel, and R. Nussinov, "Anatomy of Protein Structures: Visualizing How a 1-D Protein Chain Folds into a 3-D Shape," *Proc. Natl. Acad. Sci. USA* **97,** 12038–12043 (2000).
5. F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. E. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, "The Protein Data Bank: A Computer-Based Archival File for Macromolecular Structures," *J. Mol. Biol.* **112,** 535–542 (1977).
6. C. J. Tsai, D. Xu, and R. Nussinov, "Protein Folding via Binding, and Vice Versa," *Folding & Design* **3,** R71–R80 (1998).
7. R. L. Baldwin and G. D. Rose, "Is Protein Folding Hierarchic? I. Local Structure and Peptide Folding," *Trends Biochem. Sci.* **24,** 26–33 (1999).
8. M. Levitt and C. Chothia, "Structural Patterns in Globular Proteins," *Nature* **261,** 552–558 (1976).
9. G. M. Salem, E. G. Hutchinson, C. A. Orengo, and J. M. Thornton, "Correlation of Observed Fold Frequency with the Occurrence of Local Structural Motifs," *J. Mol. Biol.* **287,** 969–981 (1999).
10. C. J. Tsai and R. Nussinov, "Hydrophobic Folding Units Derived from Dissimilar Monomer Structures and Their Interactions," *Prot. Sci.* **6,** 24–42 (1997).
11. C. J. Tsai and R. Nussinov, "Hydrophobic Folding Units at Protein–Protein Interfaces: Implications to Protein Folding and Protein–Protein Association," *Prot. Sci.* **6,** 1426–1437 (1997).
12. A. R. Panchenko, Z. Luthey-Schulten, and P. G. Wolynes, "Foldons, Protein Structural Modules, and Exons," *Proc. Natl. Acad. Sci. USA* **93,** 2008–2013 (1996).
13. A. M. Lesk and G. D. Rose, "Folding Unit in Globular Proteins," *Proc. Natl. Acad. Sci. USA* **78,** 4304–4308 (1981).
14. S. J. Wodak and J. Janin, "Location of Structural Domains in Proteins," *Biochemistry* **20,** 6544–6552 (1981).
15. K. A. Dill, "Dominant Forces in Protein Folding," *Biochemistry* **31,** 7134–7155 (1990).
16. C. V. Gegg, K. E. Bowers, and C. R. Matthews, "Probing Minimal Independent Folding Units in Dihydrofolate Reductase by Molecular Dissection," *Prot. Sci.* **6,** 1885–1892 (1997).
17. A. Teplyakov, G. Oblomova, K. S. Wilson, K. Ishii, H. Kaji, T. Samejima, and I. Kuranova, "Crystal Structure of Inorganic Pyrophosphatase from *Thermus Thermophilus*," *Prot. Sci.* **3,** 1098–1107 (1994).
18. V. Yu. Oganessyan, S. A. Kurilova, N. N. Vorobyeva, T. I. Nazarova, A. N. Popov, A. A. Lebedev, S. M. Avaeva, and E. H. Harutyunyan, "X-Ray Crystallographic Studies of Recombinant Inorganic Pyrophosphatase from *Escherichia Coli*," *FEBS Lett*. **348,** 301–304 (1994).
19. A. M. Libson, A. G. Gittis, I. E. Collier, B. L. Marmer, G. I. Goldberg, and E. E. Lattman, "Crystal Structure of the *Haemopexin*-Like C-Terminal Domain of Gelatinase A," *Nat. Struct. Biol.* **2,** 938–942 (1995).
20. D. L. Gatti, B. Entsch, D. P. Ballou, and M. L. Ludwig, "pH-Dependent Structural Changes in the Active Site of p-Hydroxybenzoate Hydroxylase Point to the Importance of Proton and Water Movements During Catalysis," *Biochemistry* **35,** 567–578 (1996).
21. C. J. Tsai, J. V. Maizel, and R. Nussinov, "Distinguishing Between Sequential and Nonsequentially Folded Proteins: Implications for Folding and Misfolding," *Prot. Sci.* **8,** 1591–1604 (1999).
22. U. Abele and G. E. Schulz, "High-Resolution Structures of Adenylate Kinase from Yeast Ligated with Inhibitor Ap5A, showing the Pathway of Phosphoryl Transfer," *Prot. Sci.* **4,** 1262–1271 (1995).
23. D. Dreusicke and G. E. Schulz, "The Glycine-Rich Loop of Adenylate Kinase Forms a Giant Anion Hole," *FEBS Lett.* **208,** 301–304 (1986).
24. C. Bystroff and D. Baker, "Prediction of Local Structure in Proteins Using a Library of Sequence-Structure Motifs," *J. Mol. Biol.* **281,** 565–577 (1998).
25. B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, "CHARMM: A Program for Macromolecular Energy, Minimization and Dynamics Calculations," *J. Comput. Chem.* **4,** 187–217 (1983).
26. T. Lazaridis and M. Karplus, "Effective Energy Function for Proteins in Solution," *Proteins* **35,** 133–152 (1999).
27. S. Kumar, Y. Y. Sham, C. J. Tsai, and R. Nussinov, "Protein Folding and Function: The N-Terminal Fragment in Adenylate Kinase," *Biophys. J.* **80,** 2439–2454 (2001).

Received June 15, 2000; accepted for publication February 2, 2001

Chung-Jung Tsai *Laboratory of Experimental and Computational Biology, National Cancer Institute, Frederick Cancer Research and Development Center, Building 469, Room 151, Frederick, Maryland 21702 (tsai@protein3d.ncifcrf.gov).* Dr. Tsai received his Ph.D. in chemistry from the University of Pittsburgh in 1992, followed by a brief period as a Postdoctoral Fellow with Professor K. D. Jordan. He now works in the Laboratory of Experimental and Computational Biology, National Cancer Institute, National Institutes of Health. Dr. Tsai's research interests focus primarily on the study of protein folding, binding, and stability problems. He has published more than 30 papers in the fields of computational chemistry, physics, and biology.

Buyong Ma *Laboratory of Experimental and Computational Biology, National Cancer Institute, Frederick Cancer Research and Development Center, Building 469, Room 151, Frederick, Maryland 21702 (mab@fcindy13.ncifcrf.gov).* Dr. Ma received his B.Eng. degree in chemical engineering from the Hefei University of Technology in 1984 and his M.S. degree in polymer chemistry from the Hubei Research Institute of Chemistry in 1987. Working under the direction of Dr. H. F. Schaefer, he received his Ph.D. degree in physical chemistry from the University of Georgia at Athens in 1995. After three years of postdoctoral work on molecular mechanics with Dr. N. L. Allinger, he joined the National Cancer Institute as a Research Fellow. Dr. Ma's research interests include a computational approach to biochemical and biophysical problems, drug design, and protein structure and function.

Yuk Yin Sham *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (shamy@us.ibm.com).* Dr. Sham received his B.S. degree from Florida International University and his Ph.D. degree from the University of Southern California, both in chemistry. His Ph.D. thesis was on determining solvation, pK_a, and binding free energies in proteins using both implicit and all-atom solvent models. Dr. Sham joined the National Cancer Institute in 1999 as a Postdoctoral Fellow, focusing on the simulation of protein fragment stability and protein unfolding pathways; this paper was written while he was at NCI. He recently joined the Biomolecular Simulation and Stability group in the Computational Biology Center at the IBM Thomas J. Watson Research Center as a Postdoctoral Fellow to develop an efficient molecular dynamics algorithm for protein folding studies. Dr. Sham's current research interests focus on the understanding of biological function through free-energy calculation and simulation.

Sandeep Kumar *Laboratory of Experimental and Computational Biology, National Cancer Institute, Frederick Cancer Research and Development Center, Building 469, Room 151, Frederick, Maryland 21702 (kumarsan@ncifcrf.gov).* Dr. Kumar is a Postdoctoral Visiting Fellow in the Laboratory of Experimental and Computational Biology at the National Cancer Institute, Frederick Cancer Research and Development Center. He received his B.Sc. (Honours) degree in physics from the University of Delhi, India, in 1989, his M.Sc. degree in molecular biology and biotechnology from the G. B. Pant University of Agriculture and Technology, India, in 1992, and his Ph.D. degree in molecular biophysics from the Indian Institute of Science in 1998. He received the Department of Biotechnology (DBT) scholarship in 1990, a Council of Scientific and Industrial Research (CSIR) junior research fellowship in 1992, and a senior research fellowship in 1994. In 1998, Dr. Kumar received the National Institutes of Health Visiting Fellow Award. He is an author or coauthor of 21 technical papers in the field of protein structure and function.

Haim J. Wolfson *School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel (wolfson@post.tau.ac.il).* Dr. Wolfson is a Professor of Computer Science in the School of Computer Science at Tel Aviv University. He received his B.Sc., M.Sc., and Ph.D. degrees in mathematics from Tel Aviv University in 1973, 1975, and 1985, respectively. From 1975 to 1983 he was a Research Officer in the IDF, and from 1985 to 1989 he was an Associate Research Scientist and an Assistant Professor in the Robotics Research Laboratory and Computer Science Department of New York University. In 1989 he joined the Computer Science Department at Tel Aviv University, where he is conducting research in computer vision, spatial pattern discovery, and structural bioinformatics. Dr. Wolfson is the chair of the Tel Aviv University multidisciplinary steering committee on bioinformatics. He was a visiting scientist at New York University and the IBM Thomas J. Watson Research Center. Dr. Wolfson has received the year 2000 Juludan Prize awarded by the Technion, Haifa, for Outstanding Research in the Application of Exact Sciences or Advanced Technology to Medicine.

Ruth Nussinov *Medical School, Tel Aviv University, and SAIC, Laboratory of Experimental and Computational Biology, National Cancer Institute, Frederick Cancer Research Facility, Building 469, Room 151, Frederick, Maryland 21702 (ruthn@ncifcrf.gov).* Dr. Nussinov is a Professor in the Department of Human Genetics, School of Medicine, Tel Aviv University, and a Senior Scientist at the National Cancer Institute. She received her B.Sc. degree in microbiology from the University of Washington, Seattle, in 1967, and her M.Sc. in biochemistry in 1968 from Rutgers University. She received her Ph.D. in biochemistry from Rutgers in 1977. Dr. Nussinov was a Fellow at the Weizmann Institute, and a Visiting Scientist in the Chemistry Department at Berkeley and in the Biochemistry Department at Harvard. She joined the Medical School at Tel Aviv in 1985 as an Associate Professor, and in 1990 became a Full Professor. Her association with the National Institutes of Health began in 1983, first with the National Institute of Child Health and Human Development, and since 1985 with the National Cancer Institute. She is an author and coauthor of more than 180 scientific papers. Until 1990 her papers addressed RNA and DNA sequence and structure and nucleic acid–protein interactions. In 1990 she switched to proteins; her research currently focuses on protein folding and protein binding.