Protein Folding and Function: The N-Terminal Fragment in Adenylate Kinase

Sandeep Kumar,* Yuk Yin Sham,* Chung-Jung Tsai,† and Ruth Nussinov†‡

*Laboratory of Experimental and Computational Biology and †Intramural Research Support Program, SAIC Frederick, National Cancer Institute, Frederick Cancer Research and Development Center, Frederick, Maryland 21702 USA; and ‡Department of Human Genetics and Molecular Medicine, Sackler Institute of Molecular Medicine, Sackler School of Medicine, Tel Aviv University, Tel Aviv 69978, Israel

ABSTRACT Three-dimensional protein folds range from simple to highly complex architectures. In complex folds, some building block fragments are more important for correct protein folding than others. Such fragments are typically buried in the protein core and mediate interactions between other fragments. Here we present an automated, surface area-based algorithm that is able to indicate which, among all local elements of the structure, is critical for the formation of the native fold, and apply it to structurally well-characterized proteins. In particular, we focus on adenylate kinase. The fragment containing the phosphate binding, P-loop (the "giant anion hole") flanked by a β-strand and an α-helix near the N-terminus, is identified as a critical building block. This building block shows a high degree of sequence and structural conservation in all adenylate kinases. The results of our molecular dynamics simulations are consistent with this identification. In its absence, the protein flips to a stable, non-native state. In this misfolded conformation, the other local elements of the structure are in their native-like conformations; however, their association is non-native. Furthermore, this element is critically important for the function of the enzyme, coupling folding, and function.

INTRODUCTION

How contiguous fragments in the one-dimensional sequence are arranged in the three-dimensional structure of a protein is an important aspect of the protein folding problem. Protein folding is often described as a hierarchical process initiated by the formation of local interactions (Baldwin and Rose, 1999), with subsequent spontaneous intramolecular recognition and stabilization of the local substructures (Wu et al., 1994). Among these local elements, some may be expected to be more critical than others in reaching the native fold. Here our goal is twofold: first, to be able to identify these critically important folding elements; and second, to see whether the critical folding elements are also those essential for fulfilling biological function. Both require high sequence and structure conservation through evolution. Mutational events in critically important folding elements will lead to misfolded conformations, abolishing function. Hence, utilizing a given segment in both roles may confer evolutionary advantage.

Two types of examples have already indicated that such a trend may apply. The first is the intramolecular chaperone, where the proregion plays such a dual role. There, it is absolutely required for correct folding of the enzyme; However, in addition, it plays the functionally important role of an inhibitor. The second example is the dihydrofolate reductase, where it has recently been shown that the N-terminus fragment is critical for reaching the native fold (Ma et al., 2000; Sham et al., 2001). In the absence of this N-terminal fragment, the C-terminal fragment consisting of residues 37–159 of *Escherichia coli* dihydrofolate reductase forms a stable non-native structure (Gegg et al., 1997). At the same time, this N-terminal fragment also forms an integral part of the active site. To address this folding-function question, we first identify a critical element for folding. We next proceed to examine its functional role, if any.

A recent model described a folded protein as consisting of a set of hydrophobic folding units with buried hydrophobic cores capable of independent, thermodynamically stable existence (Tsai and Nussinov, 1997a, b; Tsai et al., 1998, 1999a). The hydrophobic folding units associate into domains, which in turn assemble to form a multi-domain protein fold or intermolecular multi-subunit quaternary structure. A hydrophobic folding unit is the outcome of a combinatorial assembly of a set of building blocks. A building block is defined as a highly populated fragment in a given protein structure with a continuous sequence. The sequence length of a building block is $\geq$15 amino acid residues. It may be composed of a single secondary structure or a combination of contiguous secondary structures (super-secondary structures). In contrast to a hydrophobic folding unit, a building block may not have a stable, well-defined conformation in solution by itself, and may flip among several conformations with different population times. The building block conformation seen in the native state of the protein is likely to be the one with the highest population in solution for the isolated fragment peptide.

Received for publication 16 November 2000 and in final form 9 February 2001.

Address reprint requests to Dr. Ruth Nussinov, SAIC Frederick NCI-FCRDC, Bldg. 469, Rm. 151, Frederick, MD 21702. Tel.: 301-846-5579; Fax: 301-846-5598; E-mail: ruthn@ncifcrf.gov.

S.K. and Y.Y.S. have contributed equally to this article.

0006-3495/01/05/2439/16 $2.00

However, this is not always the case. The building block conformations are stabilized by mutual interactions. Recently, these ideas have been implemented in a computer program called Anatomy (Tsai et al., 2000). The input for this program is a protein structure whose atomic coordinates have been obtained by crystallography and are available in the Protein Data Bank (PDB) (Bernstein et al., 1977). Using an iterative, top-down dissecting procedure, native protein tertiary structure is first cut into domains, then into hydrophobic folding units, and at the end of a multi-level process, into a set of building blocks. The resulting anatomy tree-like organization describes the most likely folding pathway(s) of the protein. It further illustrates its folding complexity and kinetics, and its likelihood to misfold.

Although all building blocks and their combinatorial assemblies are required for a protein to yield its complete three-dimensional fold, formation and interaction of one or a few building blocks with their sister building blocks may be essential for the protein to fold correctly. This may be particularly true for large proteins that fold in a complex manner. Recently, Tsai et al. (1999b) have described the folding complexity of a protein in terms of the arrangement of the building blocks in the protein tertiary structure. If building blocks adjacent in the primary sequence of the protein are also adjacent in the three-dimensional structure, then the protein is thought to fold in a sequential manner. Otherwise, it is a nonsequential folder. Different levels of protein folding complexity can be described within these two classes (Tsai et al., 1999b). A building block that is in contact with several other building blocks at one or more hierarchical levels of the protein anatomy tree may be critical for correct protein folding. A critical building block may be expected to fulfill three conditions. First, it should be in contact with most other building blocks in the structure. Second, it is likely to be inserted between sequentially connected building blocks, mediating their tertiary interactions. And third, most importantly, in its absence, the remaining building blocks are likely to mis-associate. Under such circumstances, the conformations of the assembled building blocks are likely to remain native-like. Nevertheless, it is also possible that alternate less stable conformations will be selected in the combinatorial assembly, mutually stabilizing each other. The sequence of a critical building block is likely to be conserved in different organisms. This suggests that mutations occurring in critical building blocks are more likely to have deleterious effects on the protein conformation than those occurring elsewhere in the protein sequence. This may suggest one reason as to why although most mutations have little effect on protein structure, some have more drastic consequences (Lim and Sauer, 1991; Lim et al., 1992; Matthews, 1993). Owing to the complexity of the fold, proteins containing critical building block(s) may be more prone to misfolding.

We have designed an algorithm to identify critical building blocks in the protein structure. Its input is the set of building blocks, obtained at different levels of cuttings of the native structure, in the generation of the anatomy tree (Tsai et al., 2000). The algorithm assigns a numerical value to each building block, based on its location in the protein, the identity and number of other building blocks it contacts, and its surface area buried by such contacts. A building block whose *critical building block index* (CIndex) is greater by at least two standard deviations than the average building block CIndex value at a given hierarchical level is identified as critical for that level. If this building block is not cut further into component building blocks at lower levels, and if it has significantly high CIndex values at more than one level, it is considered as a critical building block (CBB) for the protein. We have applied this algorithm to 10 small nonhomologous nonsequentially folded proteins. Two of these, adenylate kinase and purine nucleoside phosphorylase, contain potential critical building blocks. Both proteins fold in complex nonsequential manner (Tsai et al., 1999b). We focus on adenylate kinase due to the availability of substantial sequence, structural, and functional information.

Adenylate kinases (ADK, M_r = 21–25 kDa) catalyze the reaction

$$Mg^{2+}ATP + AMP \rightleftharpoons Mg^{2+}ADP + ADP.$$

The enzyme is known to involve large synergistic domain movements with the binding of each substrate (Schulz et al., 1990; Gerstein et al., 1993; Matte et al., 1998). Several crystal structures of adenylate kinases in the presence of different ligands and from different sources (Dreusicke et al., 1988; Muller and Schulz, 1992; Berry et al., 1994; Abele and Schulz, 1995; Schlauderer et al., 1996; Muller et al., 1996; Vonrhein et al., 1998; Berry and Phillips, 1998) are available in the PDB (Bernstein et al., 1977). *Saccharomyces cerevisiae* ADK contains an ~30-residue-long building block near its N-terminus. This building block contains the phosphate binding loop (P-loop) between a β-strand and an α-helix. The P-loop is characteristic of adenylate kinases and of a variety of ATP- and GTP-binding proteins (Schulz et al., 1990; Saraste et al., 1990; Matte et al., 1998). It has also been described as a giant anion hole (Dreusicke and Schulz, 1986). This N-terminal building block precedes the AMP binding domain in the sequence. It exhibits a high degree of sequence and structural conservation. Our algorithm identifies this building block as critical to the protein structure.

To study the role of this building block in the folding of adenylate kinases, we have carried out 2.0 ns molecular dynamics simulations of the *S. cerevisiae* ADK with its first 36 residues removed. These simulations indicate that the rest of the protein quickly shrinks, resulting in a stable, more compact, non-native conformation. In particular, examination of the shrunk structure indicates that the conformations of the individual building blocks are largely un-

changed. Consistent with the model, the non-native contacts originated mostly from the mis-association of the other building blocks, owing to the absence of the critical building block. Because the ATP binding site of adenylate kinase is abolished by the removal of this building block, the enzyme is likely to be inactive in this conformation. We have further removed three other building blocks, one at a time, and repeated the simulations. None of these building blocks has been identified as critical by our algorithm. Two of these building blocks are in the protein core and are important for protein structure, as indicated by their high CIndex values. The first building block is smaller than the critical building block, while the second building block is larger than the critical building block. At the lowest level of anatomy cutting, the second building block splits into two smaller building blocks. The third building block has the lowest CIndex value and is not located within the protein core. It corresponds to the LID domain of yeast adenylate kinase. The results of the simulations indicate that removal of the first or the third building block does not result in an appreciable change in the overall structure of yeast ADK. Removal of the second building block causes a significant perturbation in the native structure. However, the extent of the perturbation is smaller than that observed by the removal of the critical building block.

MATERIALS AND METHODS

Cutting into building blocks

Building blocks at different hierarchical levels of the protein structure are identified by using an in-house program, Anatomy, which cuts the protein at several levels (Tsai et al., 2000). The cutting procedure is based on a scoring function that has been successfully applied to locate compact hydrophobic folding units (Tsai and Nussinov, 1997a, b). The scoring function for a protein fragment (candidate building block) is given as

$$\mathrm{Score}^{\mathrm{BB}}(Z, H, I) = \frac{Z^1_{\mathrm{Avg}} - Z}{Z^1_{\mathrm{Dev}}} + \frac{H^1_{\mathrm{Avg}} - H}{H^1_{\mathrm{Dev}}} + \frac{I^1_{\mathrm{Avg}} - I}{I^1_{\mathrm{Dev}}} + \frac{Z^2_{\mathrm{Avg}} - Z}{Z^2_{\mathrm{Dev}}} + \frac{H^2_{\mathrm{Avg}} - H}{H^2_{\mathrm{Dev}}} + \frac{I^2_{\mathrm{Avg}} - I}{I^2_{\mathrm{Dev}}} \quad (1)$$

The corresponding arithmetic average and standard deviation are determined from a nonredundant dataset of 930 representative single-chain proteins. The standard deviation and average with superscript 1 are calculated with respect to the fragment size, while the values with superscript 2 are calculated as a function of the fraction of the fragment size to the whole proteins.

Z, the compactness of a protein fragment, is given by

$$Z = \mathrm{ASA}_{\mathrm{surf}}/(36\pi \mathrm{Vol}^2)^{1/3} \quad (2)$$

where Vol is evaluated by the integration of all the solvent-exposed accessible area, $\mathrm{ASA}_{\mathrm{surf}}$.

I, the isolatedness of the fragment, is the ratio of the change in nonpolar solvent-accessible surface area to the total accessible surface area, when the fragment is exposed. It is given by

$$I = \mathrm{ASA}^{\mathrm{Non}}_{\mathrm{B}\rightarrow\mathrm{E}}/\mathrm{ASA}_{\mathrm{frag}} \quad (3)$$

H, the hydrophobicity of the fragment, is given as the fraction of the buried nonpolar area out of the total nonpolar area

$$H = \mathrm{ASA}^{\mathrm{Non}}_{\mathrm{Buried}}/(\mathrm{ASA}^{\mathrm{Non}}_{\mathrm{Buried}} + \mathrm{ASA}^{\mathrm{Non}}_{\mathrm{Surf}}) \quad (4)$$

All possible protein fragments are generated from a given protein sequence and their corresponding scores are evaluated. Candidate fragments with high scores are classified as local minima on the scoring surface and are considered as building blocks. The procedure is re-applied to each building block at each level of the cutting until no further cutting is possible. Fig. 1 shows an example of building block cutting at different hierarchical levels in yeast adenylate kinase.

Identification of critical building block(s)

Table 1 lists the building blocks at different levels of yeast adenylate kinase along with the parameters showing their significance for the adenylate kinase structure. Table 2 indicates the critical building blocks identified in 15 crystal structures of adenylate kinases from different organisms. These parameters and the procedure to identify critical building blocks in proteins are described below.

Consider building block *j* that interacts with two different building blocks, *k* and *l*. We compute the differential contacting surface area for fragment *j* as

$$\mathrm{Diffcontsa}(j) = \mathrm{contsa}(j, k) + \mathrm{contsa}(j, l) - \mathrm{contsa}(k, l) \quad (5)$$

where contsa(*j*, *k*) is the surface area buried between building blocks *j* and *k*. The surface areas are calculated using the method described by Tsai and Nussinov (1997a, b). For sequentially connected building blocks, surface areas buried by residues at the junction between the two building blocks are not considered. This keeps the interaction between two building blocks independent of their sequential separation. We consider the following cases.

Diffcontsa(*j*) < 0, i.e., the interactions between building blocks *k* and *l* are stronger than the sum of their interactions with building block *j*. For example, in the top panel of Fig. 1, the interactions between building blocks shown in green (residues 30–108) and cyan (residues 180–199) are stronger than their interactions with the building block shown in magenta (residues 200–220) (Table 3). The top panel of Fig. 1 shows the building block cutting of yeast ADK at the second hierarchical level. In such cases, diffcontsa(*j*) is set to zero. This helps to keep diffcontsa for permutations of *j*, *k*, and *l* independent of each other.

When diffcontsa(*j*) > 0, i.e., the combined interactions of building block *j* with building blocks *k* and *l* are stronger than the interactions between building blocks *k* and *l*. In such cases, diffcontsa(*j*) is multiplied by different weights, as follows.

If building block *j* contacts building blocks *k* and *l*, and *k* and *l* are sequentially connected but do not interact with each other except at the junction, then diffcontsa(*j*) is multiplied by 4.0. Two building blocks are considered not to interact if they bury <20 Å^2 area between them. In the top panel of Fig. 1, the building block shown in red (residues 3–32) mediates interaction between the building blocks in cyan (residues 180–199) and magenta (residues 200–220). The surface areas buried between these building blocks are given in Table 3. The surface area buried between the building blocks shown in cyan and magenta at the junction is neglected in our calculations.

If building blocks *k* and *l* are not sequentially connected and do not interact with each other, and building block *j* is also not sequentially connected with either *k* or *l*, then diffcontsa(*j*) is multiplied by 4.0. For example, in the middle panel of Fig. 1, the building block shown in red

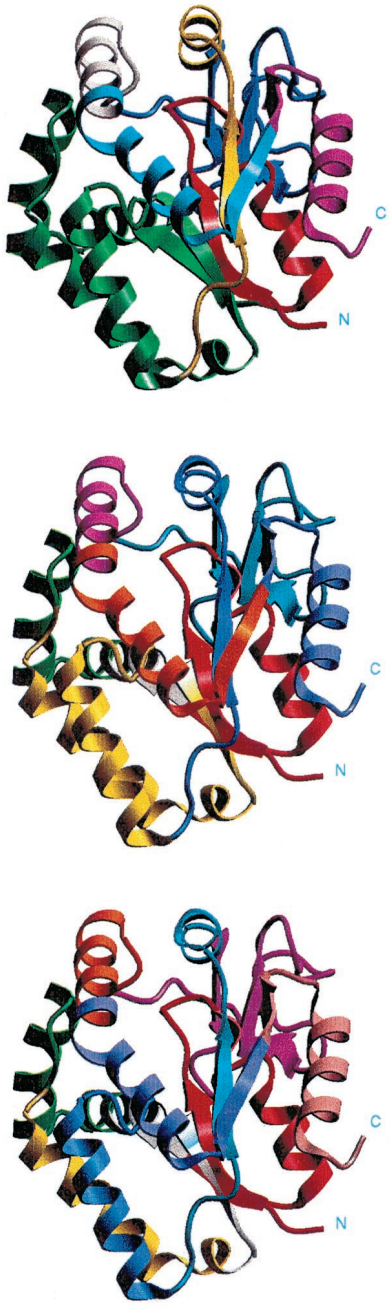


FIGURE 1 The yeast adenylate kinase (1aky) anatomy at different hierarchical levels. At a given level, each color represents a building block. The residues that cannot be assigned to any building block are shown in white. The building block shown in red at each level is the critical building block for ADK. The levels shown in the figure are 2 (*top*), 3 (*middle*), and 4 (*bottom*). Level 1 (not shown) is the whole ADK. Levels 2–4 contain, respectively, six, eight, and nine building blocks. The building blocks at each level of Anatomy are given in Table 1 along with their CIndices and statistical significance.

(residues 3–32) mediates the interaction between the building blocks shown in yellow (residues 62–112) and in cyan (residues 131–165). The surface areas buried between these building blocks are given in Table 4.

TABLE 1 Building blocks in *Saccharomyces cerevisiae* adenylate kinase (PDB code 1aky) at different levels of Anatomy

Level Number	Building block(s)	CIndex	*Z*-Score
1	Glu-3–Asn-220	—	—
2	Glu-3–His-32	25.74	1.75
	Ala-30–Gln-108	1.35	−0.62
	Gly-109–Thr-130	14.62	0.67
	Gly-131–Arg-165	0.11	−0.74
	Tyr-180–Gly-199	3.16	−0.45
	Val-200–Asn-220	1.59	−0.60
3	Glu-3–His-32	62.79	2.27
	Leu-39–Met-68	9.87	−0.37
	Leu-62–Leu-112	12.03	−0.26
	Gly-109–Thr-130	28.40	0.56
	Gly-131–Arg-165	0.23	−0.85
	Ser-166–Glu-185	8.57	−0.43
	Tyr-180–Gly-199	8.14	−0.45
	Val-200–Asn-220	7.94	−0.46
4	Glu-3–His-32	65.84	2.32
	Leu-39–Met-68	3.72	−0.69
	Leu-62–Cys-82	6.97	−0.53
	Leu-88–Leu-112	18.79	0.04
	Gly-109–Thr-130	35.46	0.85
	Gly-131–Arg-165	0.28	−0.85
	Ser-166–Glu-185	9.61	−0.40
	Tyr-180–Gly-199	10.10	−0.38
	Val-200–Asn-220	10.20	−0.37

Building blocks in *Saccharomyces cerevisiae* adenylate kinase (1aky) at different levels of Anatomy. At each level, a building block is denoted by its beginning and end residue names and numbers. Building block termini may overlap. The building blocks at levels 2–4 are depicted in Fig. 1. For each building block, its critical building block index (CIndex) and *Z*-score are also given. The two values are not calculated at level 1 because ADK is uncut at this level. At levels 2–4 the first building block Glu-3–His-32 at the N-terminal has the highest CIndex. At levels 3 and 4 the CIndices for this building block are significant at 95% level of confidence (*Z*-score $>$ 1.96). This building block is shown in red in Fig. 1 at all levels. Building block Gly-109–Thr-130 has the second highest CIndex at levels 2–4. This building block is shown in yellow in the top, blue in the middle, and cyan in the bottom panel of Fig. 1. Building block Gly-131–Arg-165 has the lowest CIndex and the least *Z*-score at each level. This building block is shown in blue in the top, cyan in the middle, and magenta in the bottom panel of Fig. 1.

The middle panel of Fig. 1 shows the building block cutting of yeast ADK at the third hierarchical level.

If building blocks k and l are not sequentially connected and do not interact with each other, and building block j is sequentially connected with either k or l (but not with both), then diffcontsa(j) is multiplied by 2.0. An example of this case is seen in the top panel of Fig. 1. The building block shown in red (residues 3–32) mediates the interaction between the building blocks shown in green (residues 30–108) and magenta (residues 200–220). The surface areas buried between these building blocks are given in Table 3.

Hence, larger weights are given to cases where building block j mediates the interactions between building blocks that are not in direct contact. In all other combinations of j, k, and l, diffcontsa(j) gets a weight of 1.0.

At a given level the critical building block index (CIndex(j)) for building block j is the sum of diffcontsa(j) computed for all combinations of k and l divided by the total surface area of building block j, totsa(j).

$$\mathrm{CIndex}(j) = \sum \mathrm{diffcontsa}(j)/\mathrm{totsa}(j) \qquad (6)$$

TABLE 2 Critical building blocks in adenylate kinase structures

PDB Entry and Resolution (Å)	Source Organism	Oligomeric State	Bound Substrate(s) and/or Mutation(s)	Critical Building Block
1AKE (1.90)	*E. coli*	Dimer	AP_5A	A Met-1–Ile-26 B Met-1–Gln-28
1ANK (2.00)	*E. coli*	Dimer	AMP, AMPPNP	A Met-1–Ile-26 B Met-1–Gln-28
2ECK (2.80)	*E. coli*	Dimer	AMP, ADP	A Met-1–Gly-25* B Met-1–Gln-28
4AKE (2.20)	*E. coli*	Dimer		A Met-1–Gln-28* A Val-103–Val-121 B Met-1–Gln-28* B Val-103–Val-121*
1AKY (1.63)	*S. cerevisiae*	Monomer	AP_5A, IMD	Glu-3–His-32
2AKY (1.96)	*S. cerevisiae*	Monomer	AP_5A, Mg^{2+}	Glu-3–His-32
3AKY (2.23)	*S. cerevisiae*	Monomer	AP_5A, IMD, I213F	Glu-3–His-32
1DVR (2.36)	*S. cerevisiae*	Dimer	ATF, D89V, R165I	A Glu-3–His-32 B Glu-3–His-32*
1ZIN (1.60)	*B. stearothermophilus*	Monomer	AP_5A, Zn^{2+}	Met-1–Gly-25
1ZIO (1.96)	*B. stearothermophilus*	Monomer	AP_5A, Zn^{2+}, Mg^{2+}	Met-1–Gly-25
1ZIP (1.96)	*B. stearothermophilus*	Monomer	AP_5A, Zn^{2+}, Mn^{2+}	Met-1–Gly-25
1ZAK (3.50)	*Zea mays*	Dimer	AP_5A	A Lys-7–Leu-31* B Lys-7–Leu-31*
2AK2 (2.10)	*Bos taurus*	Monomer	SO_4^{2-}	Pro-14–Val-42*
3ADK (2.10)	Porcine Muscle	Monomer	SO_4^{2-}	Val-13–His-36*
1NKS (2.57)	*S. acidocaldarius*	Trimer†	ADP, AMP	—

Critical building blocks in adenylate kinase crystal structures. The building blocks that have significant CIndex values, at 95% or more level of confidence, in at least one subunit are shown. The corresponding building blocks in the other subunits also have high, but not significant, CIndex values. Such building blocks are indicated by an asterisk.

†The crystal asymmetric unit of adenylate kinase from *Sulfolobus acidocaldarius* contains two trimers. In ADK structures from *Zea mays*, *Bos taurus*, and porcine muscle, no building block has a significant CIndex value. However, ~25–30-residue-long building blocks have high CIndex values. These building blocks are also indicated by an asterisk. The building block termini may overlap. Approximately 30-residue-long building blocks at the N-termini of ADK from *E. coli*, *S. cerevisiae*, and *B. stearothermophilus* are critical.

where totsa(j) is used as a normalization factor to facilitate comparison among different building blocks. If similar weights were used for the interaction of j with other building blocks, CIndex(j) would have always been <1. The total surface area of building block j, totsa(j), has two terms: the surface area buried by the rest of the protein (protburysa(j)) and the surface area exposed to water (solvent), solvexpsa(j). Their ratio indicates the location of j in the protein. The final critical building block index for block j is given by

$$\mathrm{CIndex}(j) = \mathrm{CIndex}(j) * \frac{\mathrm{protburysa}(j)}{\mathrm{solvexpsa}(j)} \qquad (7)$$

If j is largely buried, the ratio protburysa(j)/solvexpsa(j) is >1. If j is largely on the surface, protburysa(j)/solvexpsa(j) is <1. A building block that is buried in the protein core and that interacts with several other building blocks gets a high CIndex. In contrast, a building block that is solvent-exposed with little interaction with other building blocks gets a low CIndex. All building blocks at all levels are assigned a CIndex. At each level we compute the average and standard deviation for the building block

TABLE 3 Surface areas (Å²) buried among building blocks at level 2

BB	30–108	109–130	131–165	180–199	200–220
3–32	989	1235	69	374	820
30–108		207	115	677	0
109–130			81	458	493
131–165				0	17
180–199					0

Surface areas buried among various building blocks of *Saccharomyces cerevisiae* adenylate kinase (1aky) at the second level of Anatomy. The building blocks (BB) are indicated by their beginning and end residue numbers. The value of the surface area buried between the two building blocks j and k is the average of the surface areas of building block j buried by the building block k, and vice versa. At each level, the building block termini are reassigned to remove the overlap among the building blocks (see Tables 4 and 5). This may lead to slightly different values of surface areas buried between two building blocks at different levels. Furthermore, the surface areas buried at the junction of two successive building blocks are not taken into account. Details are given in the Methods section.

TABLE 4 Surface areas (Å²) buried among building blocks at level 3

BB	39–68	62–112	109–130	131–165	166–185	180–199	200–220
3–32	0	1406	1069	67	175	172	820
39–68		2029	0	23	328	0	0
62–112			1	0	42	459	0
109–130				81	207	308	493
131–165					13	0	17
166–185						6	0
180–199							0

See legend to Table 3 and Methods for details.

CIndex values. The statistical significance of each building block is measured by its *Z*-score

$$Z\text{-score}(j) = (\text{CIindex}(j) - \mu)/\sigma \quad (8)$$

where μ is the average building block CIndex value at the given level and σ is the standard deviation about the average CIndex value.

A building block is considered to be critical for the protein if it satisfies the following criteria: 1) the building block is found at most levels below hydrophobic folding unit level of the protein anatomy; 2) the building block has a consistently high CIndex at different levels; 3) the building block's CIndex is significant by at least two standard deviations (95% level of confidence, *Z*-score > 1.96) about the average, at least in one hierarchical level of the protein anatomy.

Multiple sequence alignments

Multiple sequence alignments are performed using CLUSTALW, a non-GCG extension to SEQLAB in the GCG Wisconsin package version 10.0-UNIX. All the default parameters were used for gap opening and extension. The protein scoring matrix used is BLOSUM. The sequences for adenylate kinases are extracted from SWISSPROT database release 38.0 (June, 1999).

Molecular dynamics simulations

Five molecular dynamics simulations are performed on yeast ADK (PDB: 1aky, 1.63 Å resolution) at 300 K using the c27b1 version of CHARMM (Brooks et al., 1983). An implicit solvent model combined with CHARMM 19 polar hydrogen energy function (EEF1) is used (Lazaridis and Karplus, 1999). All crystallographic waters and substrate molecules are deleted. In each simulation hydrogen atoms are added using the HBUILD algorithm (Brunger and Karplus, 1988). Each structure is first initialized with 1000 steps of adapted basis Newton-Raphson (ABNR) minimization followed by a 2-ns simulation with a 2-fs time step at 300 K. The trajectories are saved at 1-ps time intervals. Both root-mean-square deviation for the C^{α} atoms (C^{α}-RMSD) and the number of contacts are used to determine the structural changes that occur during the molecular dynamics simulation. Two nonconsecutive residues, whose C^{α} atoms fall within 6.0 Å distance, are considered to be in contact.

The first molecular dynamics simulation involves the whole structure of ADK (residues Glu-3–Asn-220). In the second simulation, the N-terminal building block is removed from the ADK structure (simulating residues Asp-37–Asn-220). In the third simulation, we have removed the building block Gly-131–Arg-165. Hence, the third simulation consists of 183 residues in two fragments, Glu-3–Thr-130 and Ser-166–Asn-220. In the fourth simulation, we have removed the building block Gly-109–Thr-130. Hence, this simulation consists of 196 residues in two fragments, Glu-3–Gln-108 and Gly-131–Asn-220. In the fifth simulation we have removed Leu-62–Leu-112. This is a single building block at level 3 but breaks into two different building blocks at level 4 (Table 1). Hence, the fifth simulation consists of 167 residues in two fragments, Glu-3–Gly-61 and Glu-113–Asn-220.

For each molecular dynamics simulation the initial ($t = 0.0$ ns) and final ($t = 2.0$ ns) conformers are superimposed by performing a minimum RMSD alignment of the two. This removes deviations due to rigid body rotation and translation produced during the simulation. C^{α}-RMSD between the two conformers is used as a measure of structural divergence at the beginning and end of the simulation.

RESULTS AND DISCUSSION

The procedure to identify critical building blocks in proteins

Protein folding is thought to initiate locally and gradually fold into the native structure via interaction between the local structural elements. This is referred to as the hierarchical model of protein folding (Baldwin and Rose, 1999). A second model is the hydrophobic collapse, where protein folding is initiated by hydrophobically driven collapse of the unfolded polypeptide, followed by the formation of local structure (Chan and Dill, 1990). If, however, the hydrophobic collapse involves initial collapse of the local elements, with subsequent binding of these units, the two models are reconciled with each other. The building block model of protein folding is based on the concept of hierarchical folding, but considers hydrophobicity as its driving force. Based on this model, we have devised a procedure to iteratively dissect native protein structures into smaller compact units. At each level the stability of the units arises largely from intra-unit interactions. This iterative algorithm facilitates examination of protein anatomy at several hierarchical levels (Tsai et al., 2000). At the lower levels of the dissection the algorithm yields a set of building blocks for the protein. Fig. 1 presents such a dissection by cutting the yeast ADK (PDB entry 1aky) into building blocks at different hierarchical levels. Table 1 lists these building blocks at all levels. The cutting is based on three measurements: hydrophobicity, compactness, and the "isolatedness," i.e., the surface area that is buried before the cutting, and subsequently becomes exposed. The cutting procedure and the new algorithm to identify those building blocks whose role is likely to be critical for protein folding are detailed in Materials and Methods.

To identify CBB for each protein chain we use the Anatomy program to produce building blocks at several hierarchical levels. If a given level contains three or more building blocks, we compute a CIndex for each of the building blocks based on its interactions with its building block neighbors, and on its location in the protein. The interactions between the building blocks are measured in terms of the buried surface area (hydrophobic and polar). A building block that is buried in the protein core and that interacts with several other building blocks gets a high CIndex, as it does if it mediates interactions between sequentially connected building blocks. In contrast, if it is solvent-exposed with little interaction with other building blocks, it gets a low CIndex. At each level, we compute the average and standard deviation for the CIndices.

A building block is considered to be critical for the protein if it satisfies the following criteria: 1) the building block is found at most levels below the hydrophobic folding unit level of the protein anatomy; 2) the building block has a consistently high CIndex at different levels; and 3) the building block's CIndex is significant by at least two stan-

dard deviations (95% level of confidence, Z-score > 1.96) about the average CIndex, at least in one hierarchical level of the protein anatomy.

The proteins

Critical building blocks relate to the complexity of the protein fold. Here we focus on small proteins. We have applied this procedure to 10 nonsequential folders picked from the list compiled by Tsai et al. (1999b). They are monomers (200–300 residues), have high-resolution crystal structures, and are structurally and sequentially nonhomologous. These include ADK (PDB:1aky), chloroperoxidase (1cpo), dihydrofolate reductase (1dls and 7dfr), Gp32III (1gpc), cytochrome F (1hcz), concanavalin A (1jbc), leukemia inhibitory factor (Lif) (1lki), purine nucleoside phosphorylase (1pbn), endoglucanase V (2eng), and type III chloramphenicol acetyltransferase (3cla). We identified CBBs in ADK and purine nucleoside phosphorylase at 95% level of confidence. The CBBs are consistent with visual examination facilitated by assignment of different colors to the building blocks. Although there are important building blocks at the different hierarchical levels in the remaining eight proteins, none has a high enough CIndex value to qualify as a CBB.

Different structures for the same protein

Our criteria for identifying CBBs rely on the surface areas buried between the building blocks. The buried surface areas indicate the extent of interaction between two building blocks in an empirical manner. However, the calculated buried surface areas may vary due to several reasons. First, the protein crystal structures determined in the presence (or absence) of substrates may show movements between two parts due to hinge-bending motions. Second, the surface areas may depend somewhat upon the resolution of the structure. Third, different organisms can have significant sequence variations due to point mutations, insertions, and deletions at different locations. A CBB has a significantly high CIndex value irrespective of these. Hence, we apply this procedure to several crystallographic structures of the same protein, solved in the presence of different substrates and from different organisms. ADK serves as a good test of the algorithm because it is known to be a highly flexible protein (Schulz, 1992) undergoing extensive conformational changes in the presence of different substrates. In particular, *E. coli* and yeast ADKs are among the structurally best-studied proteins.

The yeast ADK consists of a five-stranded parallel β-sheet surrounded by α-helices (Fig. 1 in Abele and Schulz, 1995). The protein can be subdivided into three distinct domains. The CORE (residues 5–33, 64–130, and 169–218) domain consists of five-stranded β-sheet and adjacent helices. This domain contains the ATP binding site. NMPbind (residues 34–63) domain contains the AMP binding site. The LID (residues 131–168) domain covers the ATP and phosphoryl transfer region (Abele and Schulz, 1995). The *E. coli* ADK has a similar fold, except that the LID domain is referred to as the INSERT domain (Schulz et al., 1990). Fig. 1 shows the anatomy cutting for yeast adenylate kinase (1aky) at different hierarchical levels. Table 1 lists the building blocks in the yeast ADK structure at various levels. The number of building blocks increases from the top (level 1) to the bottom (level 4). The cutting program allows a seven-residue overlap of consecutive building blocks. At the lowest level, the building block Leu-39–Met-68 coincides with the NMPbind domain and Gly-131–Arg-165 coincides with the LID domain. Building blocks Glu-3–His-32, Leu-62–Cys-82, Leu-88–Leu-112, Gly-109–Thr-130, Ser-166–Glu-185, Tyr-180–Gly-199, and Val-200–Asn-220 constitute the CORE domain.

Table 1 also lists the building block CIndex values along with their significance at each level. A search for ADK (EC 2.7.4.3) in the PDB has yielded 15 crystal structures. Table 2 lists the CBBs we have identified. These structures are for adenylate kinases from *E. coli*, *S. cerevisiae*, *Bacillus stearothermophilus*, *Zea mays*, *Bos taurus*, porcine muscle, and *Sulfolobus acidocaldarius*. A building block consisting of ~30 residues near the N-terminus has a significantly high CIndex (at 95% level of confidence) in the structures of ADKs from *E. coli*, yeast, and *B. stearothermophilus*. The corresponding building block has high, but not significant enough, CIndices in ADKs from *Z. mays*, *B. taurus*, and porcine muscle. For ADK from *S. acidocaldarius*, the building block containing this motif does not have a high CIndex. The detailed anatomy tree for this hyperthermophilic ADK is simpler than those for ADKs from *E. coli*, yeast and *B. stearothermophilus* (Fig. 2, *a* and *b*). There is also a considerably larger difference in their amino acid sequences. The hyperthermophilic ADK clusters with methanococcal adenylate kinases and is only distantly related to other ADK sequences (Vonrhein et al., 1998). The sequence homology between these archaeal enzymes and other structurally known NMP kinases is mostly restricted to the P-loop region (Haney et al., 1997).

The N-terminus building block contains the P-loop that binds the β-phosphate of ATP flanked by a β-strand and an α-helix, and is part of the CORE domain (Yan and Tsai, 1999). ADKs belong to the nucleoside monophosphate (NMP) kinase family. The P-loop is the most ancient motif conserved in all (Dreusicke and Schulz, 1986). The consensus sequence for the P-loop, GXXXXGK(S/T), contains structurally and catalytically important residues (Reinstein et al., 1990). The lysine in the P-loop is conserved, orienting the nucleoside phosphate to bind to the active site residues (Byeon et al., 1995). The C^{α}-RMSD for the whole enzymes from *E. coli* (1ake chain A) and yeast (1aky) is 1.21 Å, and from *E. coli* and *B. stearothermophilus* (1zin) it is 1.21 Å.

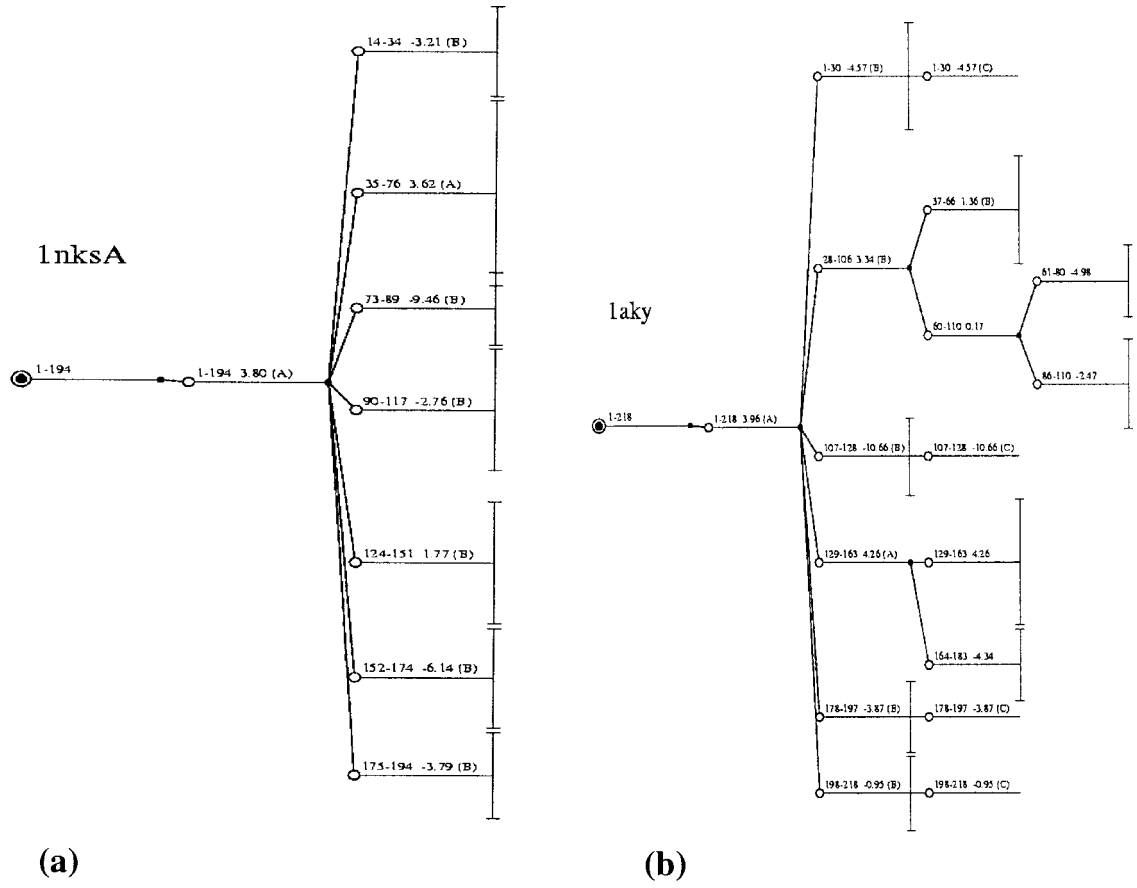


FIGURE 2 Anatomy trees for adenylate kinases from (*a*) *Sulfolobus acidocaldarius* (chain A, 1nks) and (*b*) yeast. The archaeal ADK are only distantly related to yeast ADK. Such anatomy trees for proteins with known structures are available at the web page http://protein3d.ncifcrf.gov:1025/tsai/

The corresponding N-terminal building blocks in the three structures (residues M7–G36 in 1aky, L3–G32 in 1zin, and I3–G32 in 1ake chain A) have average C^{α}-RMSD values between each other around 0.455 Å, indicating a strong structural conservation. Fig. 3 shows a superposition of the three building blocks. Fig. 4 gives the multiple sequence alignment of all known ADK sequences in the SWISS-PROT database aligned using the CLUSTALW program. P-loop sequences are highlighted. The region flanking the P-loop is also highly conserved. In Fig. 1, this building block is shown in red at all levels of the hierarchical cutting. Tables 3–5 show surface areas buried among different building blocks at each hierarchical cutting level of the yeast ADK. This building block (Glu-3–His-32) interacts extensively with most building blocks at each level.

Table 1 shows that building block Gly-109–Thr-130 has the second highest CIndex at levels 2–4. This building block is also part of the CORE domain of the yeast enzyme, and is also important for correct folding. However, its CIndex does not qualify it to be a CBB. It interacts with many other building blocks, however not as extensively as the red fragment. Building blocks Glu-3–His-32 and Gly-109–Thr-130 also interact with each other (Tables 3–5 and Fig. 5).

The building block Leu-62–Leu-112 ranks third in terms of CIndex value at level 3 (Table 1). At the lowest (fourth)

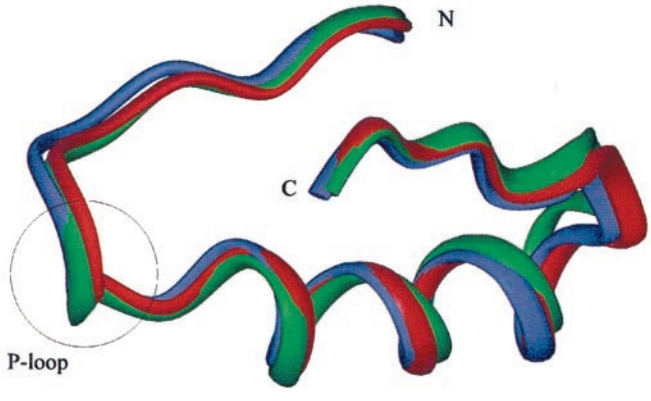


FIGURE 3 Superposition of critical building blocks identified in adenylate kinases from *E. coli* (*blue*), yeast (*green*), and bacillus (*red*). The building blocks are found at the N-terminus and contain the P-loop motif.

FIGURE 4 Multiple sequence alignment of adenylate kinase sequences in SWISSPROT using the CLUSTALW program in the GCG Wisconsin package. The region corresponding to the P-loop near the N-terminus is shown in red. There is a high degree of conservation in the region containing the P-loop motif and the flanking residues. Only the N-terminal portion of the alignment is shown in this figure.

level, this building block is further split into two building blocks, namely Leu-62–Cys-82 and Leu-88–Leu-112. Leu-88–Leu-112 still retains the third highest CIndex at level 4. Leu-62–Leu-12 also represents an important but noncritical region of the CORE domain of adenylate kinase.

The building block Gly-131–Arg-165 provides an interesting example. It has the lowest CIndex at each level and has the least significance. At each level, the interaction between this and other building blocks is the least (Tables 3–5). It is relatively isolated from the rest of the structure. This region of ADKs shows less sequence conservation. It coincides with the LID domain (Gly-131–Asp-168) of the yeast ADK (Abele and Schulz, 1995). The LID is a highly mobile part of ADK, closing over the bound ATP substrate (Schlauderer et al., 1996; Muller et al., 1996). 1aky is the liganded, closed form. In the open form, the LID is further separated.

Molecular dynamics simulations

The third and most important criterion in the definition of a critical building block is that its absence leads to non-native association between the other building blocks. We have therefore carried out molecular dynamics simulations of the yeast ADK. The aim of these simulations was to analyze the conformational changes in the rest of adenylate kinase structure caused by removing the N-terminal candidate crit-

TABLE 5 Surface areas (Å²) buried among building blocks at level 4

BB	39–68	62–82	88–112	109–130	131–165	166–185	180–199	200–220
3–32	0	116	553	1069	67	175	172	820
39–68		130	89	0	23	328	0	0
62–82			558	0	0	0	0	0
88–112				0	0	0	459	0
109–130					81	207	308	493
131–165						13	0	17
166–185							6	0
180–199								0

See legend to Table 3 and Methods for details.

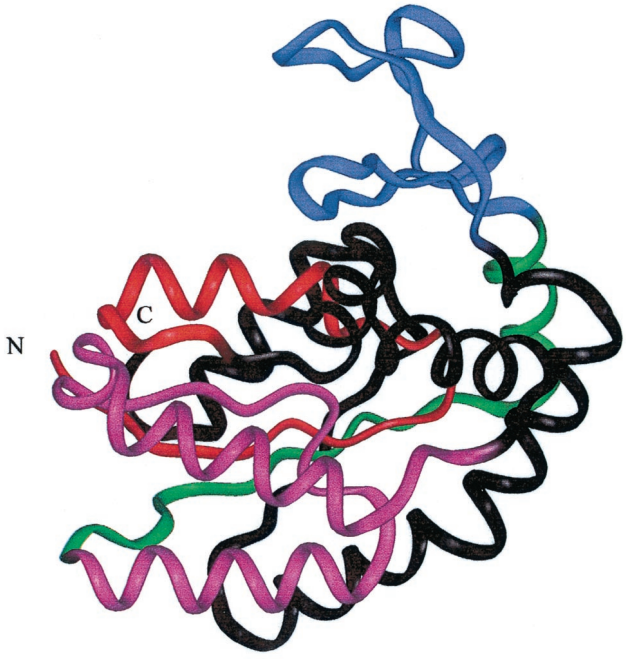


FIGURE 5 Ribbon diagram of yeast adenylate kinase highlighting the positions of the three building blocks, namely Glu-3–His-32 (*red*), Gly-109–Thr-130 (*green*), and Gly-131–Arg-165 (*blue*). The building block Glu-3–His-32 is the candidate critical building block. Gly-109–Thr-130 has the second highest CIndex values and Gly-131–Arg-165 has the lowest CIndex value. Leu-62–Leu-112 (*magenta*) is a single building block at level 3 but breaks into two building blocks Leu-62–Cys-82 and Leu-88–Leu-112 at the fourth level. This building block has the third highest CIndex at level 3. At level 4, Leu-88–Leu-112 has the third highest CIndex value.

ical building block. Previously, molecular dynamics simulations have been performed on *E. coli* ADK to study its nucleotide binding properties and global motions (Kern et al., 1994; Elamrani et al., 1996).

We have performed five molecular dynamic simulations using CHARMM (Brooks et al., 1983). We use an implicit solvent model combined with CHARMM 19 polar hydrogen energy function (Lazaridis and Karplus, 1999). In these simulations the solute (protein) is represented in atomic details, while the solvent (water) is represented only in terms of its bulk properties. Simulations using this model have been shown to yield a good agreement with explicit solvent simulation (Lazaridis and Karplus, 1999) and are able to discriminate between native and misfolded proteins (Lazaridis and Karplus, 1998). The use of such a model greatly enhances the speed of the simulation and allows for longer simulation time. All simulations have been performed at 300 K.

Each simulation was performed for 2.0 ns. The first simulation is carried out on the whole adenylate kinase. This simulation is used as a benchmark for the other four simulations. Below, we refer to it as ADKW. The other four simulations are performed on fragments of the adenylate kinase obtained after removing four different building blocks. The second molecular dynamics simulation is performed on the adenylate kinase with its N-terminal candidate critical building block removed. This simulation is for ADK fragment Asp-37–Asn-220. We denote this fragment ADKΔCBB. The third simulation is performed on ADK with the LID (Gly-131–R165) building block removed. The third simulation is for fragments Glu-3–Thr-130 and Ser-166–Asn-220. We refer to it as ADKΔLID. This simulation acts as a negative control, since the building block Gly-131–R165 is largely exposed, and has the lowest CIndex value at all the levels. The fourth simulation is for fragments Glu-3–Gln-108 and Gly-131–Asn-220, obtained after removing the building block Gly-109–Thr-130. We refer to this simulation as ADKΔ109–130. This building block is part of the CORE domain and largely buried. It has the second highest CIndex value at all levels. The fifth simulation is for fragments Glu-3–Gly-61 and Glu-113–Asn-220 obtained by removing the region Leu-62–Leu-112 from the CORE domain of the yeast adenylate kinase. This simulation is referred to as ADKΔ62–112. The fourth and fifth simulations are positive controls in our experiment because important but noncritical fragments have been removed from the protein core. Furthermore, Gly-109–Thr-130 (22 residues) is smaller and Leu-62–Leu-112 (51 residues) is larger than the protein fragment containing the candidate critical building block (36 residues). The structural arrangements of the N-terminal candidate critical building block, Leu-62–Leu-112, Gly-109–Thr-130, and Gly-131–R165 in the yeast ADK are shown in Fig. 5.

Effect of removing the critical building block

The C^{α}-RMSD trajectories for the simulations of the native protein, ADKW, ADKΔCBB, ADKΔLID, ADKΔ109–130, and ADKΔ62–112 are shown in Fig. 6. The C^{α}-RMSD between the initial (t = 0 ns) and final (t = 2.0 ns) conformations of ADKW, ADKΔCBB, ADKΔLID, ADKΔ109–130, and ADKΔ62–112 are 2.9, 6.2, 3.8, 4.2, and 5.0 Å, respectively. It can be seen that removal of the N-terminal building block causes the largest structural change.

Simulations performed on the native structure (ADKW) and ADKΔLID show a similar behavior. In both simulations, the initial C^{α}-RMSD is ~2 Å. For ADKW, the deviations rise quickly to >2.5 Å and fluctuate between 2.5 and 3.0 Å for the rest of the simulation time. For ADKΔLID the C^{α}-RMSD gradually rises to ~3.5 Å within the first 1 ns. In the final 1-ns run, C^{α}-RMSD fluctuates between 3.5 and 4.0 Å. Hence, removal of the building block Gly-131–Arg-165 (LID domain) affects the enzyme structure. However, the differences are small, with the native fold largely retained.

Removal of building block Gly-109–Thr-130 affects the rest of ADK structure to a smaller extent as compared to the

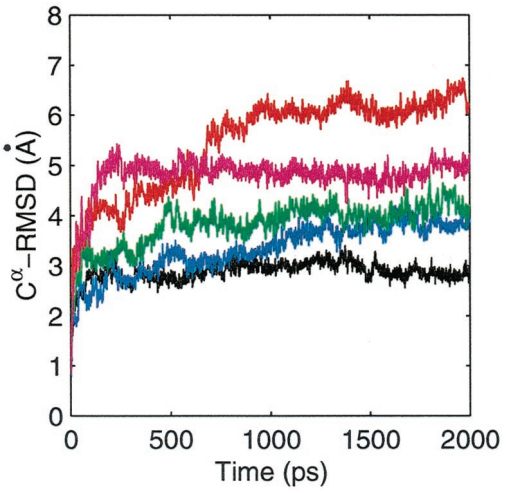


FIGURE 6 Molecular dynamics simulation trajectories showing C^{α}-RMSDs for the whole yeast adenylate kinase (ADKW) (*black*), ADKΔLID (*blue*), ADKΔCBB (*red*), ADKΔ109–130 (*green*), and ADKΔ62–112 (*magenta*). The *x*- and *y*-axes denote time and C^{α}-RMSD, respectively.

removal of the N-terminal candidate critical building block. ADKΔ109–130 has an initial C^{α}-RMSD of ~2.0 Å followed by a rise to ~3.5 Å within the first 1 ns, and fluctuations between 3.5 and 4.0 Å for the rest of the simulation. It can be argued that the smaller deviation between the initial and the final conformations of ADKΔ109–130 may be due to the smaller size of the Gly-109–Thr-130 building block. Hence, we have performed another simulation in which a larger fragment (Leu-62–Leu-112) has been removed from the protein core. ADKΔ62–112 shows an initial C^{α}-RMSD of ~1.0 Å. However, the C^{α}-RMSD rises quickly to ~5.0 Å within 200–300 ps and fluctuates at this value for the rest of the simulation. Hence, removal of the larger noncritical Leu-62–Leu-112 region in the ADK core also perturb the rest of the ADK structure to a smaller extent than the removal of the N-terminal candidate critical building block.

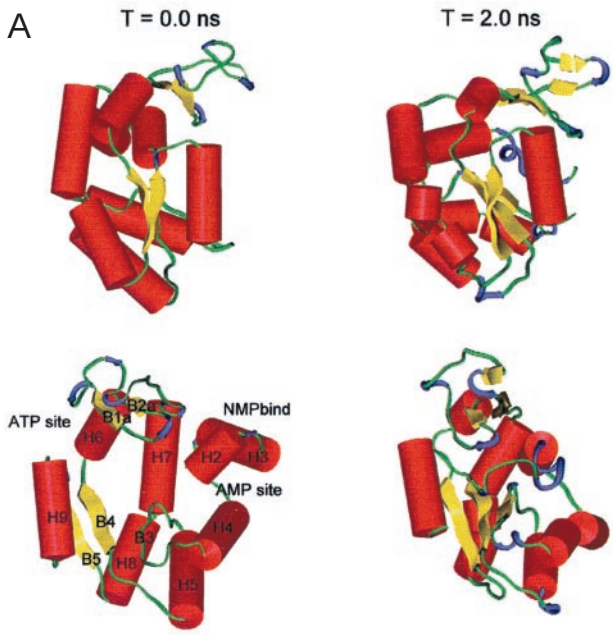
ADKΔCBB has an initial C^{α}-RMSD of ~3.5 Å followed by a gradual rise to 6 Å within the first 1 ns of the simulation. In the next 1 ns, the C^{α}-RMSD fluctuates between 6 and 7 Å. Fig. 7 *a* shows two perpendicular views of the initial and final conformations of ADKΔCBB. A significant amount of secondary structure is retained in the final conformation. These observations are in agreement with those of Kern et al. (1994) who performed 300-ps molecular dynamic simulations of *E. coli* adenylate kinase in vacuum and in explicit solvent. The secondary structures elements (assigned using DSSP; Kabsch and Sander, 1983) move closer in the final conformation.

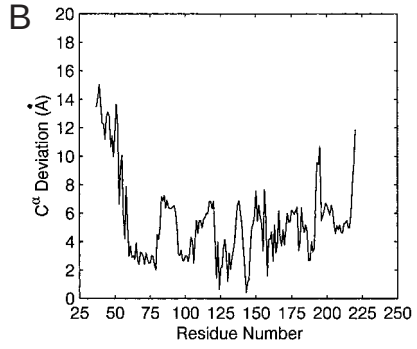
The final (t = 2.0 ns) conformation of ADKΔCBB is more compact than the initial (t = 0.0 ns) one. The compactness (Tsai and Nussinov, 1999a, b) for the initial and final ADKΔCBB conformations are 2.0 and 1.6, respectively. For the whole protein it is 1.8. Removal of the N-terminal building block creates a cavity in the CORE domain of ADK. The accessible surface area (ASA) of the whole enzyme is 11,767 $Å^2$. The ASA of the initial conformation of ADKΔCBB is 12,242 $Å^2$, indicating exposure of 475 $Å^2$ of surface area, previously buried. The final conformation of ADKΔCBB has an ASA of 10,727 $Å^2$. Hence, 1515 (12,242 − 10,727) $Å^2$ of the exposed surface is buried at the end of the simulation for ADKΔCBB. The C^{α} deviation and change in accessible surface area between the initial and the final conformations of ADKΔCBB are plotted in Fig. 7 *b*. Almost all residues move by >2 Å. Most of the residues also show a decrease in their ASAs. As expected, the residues at the termini show larger C^{α} deviations. Fig. 7 *c* plots the radius of gyration for ADKΔCBB. For ADKΔCBB, this radius decreases from ~17.7 Å to ~16.5 Å in the initial phase of the simulation and stabilizes between 16.3 and 16.5 Å for the rest of the simulation. When computed using only the C^{α} atoms, it decreases from 17.4 Å in the initial conformation to 16.2 Å in the final conformation of ADKΔCBB. The C^{α} atom radius of gyration for the whole ADK is 16.7 Å. Taken together, these observations indicate that in the absence of the critical N-terminal building block, the structure (ADKΔCBB) shrinks into a more compact form. Figs. 6 and 7 *c* illustrate that this collapse occurs within the first nanosecond of the simulation time.

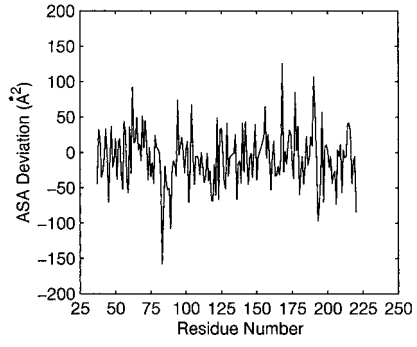
Our simulations take into account solvent molecules implicitly. The implicit solvation simulation protocol does not contribute toward the protein contraction (Lazaridis and Karplus, 1999). Hence, it is reasonable to conclude that ADKΔCBB adopts a different, non-native conformation in the absence of the critical building block. This non-native conformation is likely to be stable and more compact than the native conformation. Removal of noncritical fragments from the protein core also affects the adenylate kinase structure, but to smaller extents.

Non-native association of other building blocks in the absence of the critical building block

At the lowest level, the program Anatomy identifies nine building blocks in ADK (Table 1). Table 6 lists the C^{α}-RMSDs between the conformations of eight building blocks (excluding the CBB) at the start and end of the simulation of ADKΔCBB. The initial and final conformations of the individual fragments have lower RMSDs than the whole ADKΔCBB. Fig. 8 shows the superpositions of the individual building blocks, illustrating that their conformations are largely preserved. The largest changes occur in the building block Leu-39–Met-68, which coincides with the NMP binding domain of ADK. An α-helix present at the N-terminus of this building block is unfolded in the final conformation (Figs. 7 *a* and 8). This building block is adjacent to the N-terminal critical building block. Apart from this, the







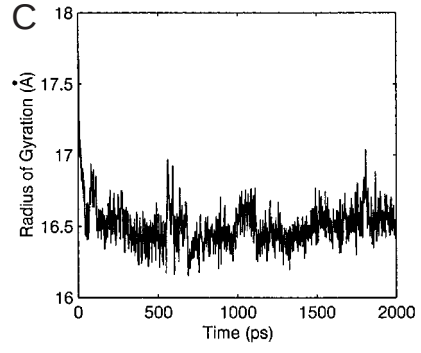


TABLE 6 Cα-RMSD between initial and final conformations of different building blocks

Building block	r.m.s.d. (Å)
Leu-39–Met-68	4.273
Leu-62–Cys-82	1.690
Leu-88–Leu-112	2.417
Gly-109–Thr-130	2.995
Gly-131–Arg-165	2.106
Ser-166–Glu-185	2.608
Tyr-180–Gly-199	3.068
Val-200–Asn-220	1.960
ADKΔCBB	6.186

individual building blocks are quite stable during the course of the simulation. Thus, the collapse in ADKΔCBB occurs mostly due to mis-association of the other building blocks in the absence of the critical N-terminal building block.

Gerstein et al. (1993) have described four joints at the N- and C-termini of α-helices 6 and 7 in *E. coli* and beef mitochondrial ADKs, helping in the closure of the INSERT domain upon substrate binding. In the ADKΔCBB these joints help the LID domain to move closer to the truncated CORE domain at the end of the simulation. The NMPbind domain also moves closer to the truncated CORE domain.

The conformational change in ADKΔCBB is reflected in the "non-native" contacts formed at the end of the simulation. Here, we monitor only the movement in the C^{α} traces of the initial and final conformations. The starting conformation of ADKΔCBB has 322 residue contacts. These are "native" contacts. The number of residue contacts in the final conformation of ADKΔCBB is 357; 233 of these 357 (65.3%) contacts are native contacts, present in the initial conformation as well. The remaining 124 (34.7% of 357) are non-native contacts, formed owing to the shrinkage of the ADKΔCBB fragment; 89 (of 322, 27.6%) native contacts are broken in the final conformation.

FIGURE 7 (*a*) Initial (t = 0.0 ns) and final (t = 2.0 ns) conformations of ADKΔCBB. The structure of adenylate kinase fragment Asp-37–Asn-220 (ADKΔCBB) is more compact at the end of the simulation. At the bottom left panel, we identify secondary structure elements in the initial conformation of ADK (α-helices H2 (residues 38–44), H3 (48–58), H4 (65–78), H5 (85–108), H6 (122–130), H7 (170–183), H8 (185–194), and H9 (206–217)) and β-strands (B3 (86–89), B4 (115–119), B5 (197–201), B1a (132–134), B2a (141–143)). The secondary structure labeling is following Abele and Schulz (1995). (*b*) Plots showing residue-wise C^{α} and accessible surface area (ASA) deviations between the initial (t = 0.0 ns) and final (t = 2.0 ns) conformations of ADKΔCBB. The deviations are computed with respect to the initial conformation of ADKΔCBB. In each plot, the x-axis indicates the residue number. In the plot showing the ASA deviation, a negative deviation indicates that the residue is more buried in the protein core in the final conformation as compared to the initial conformation. The reverse is true for the positive ASA deviation. (*c*) The radius of gyration of ADKΔCBB as function of simulation time. The x- and y-axes denote the time and radius of gyration, respectively.

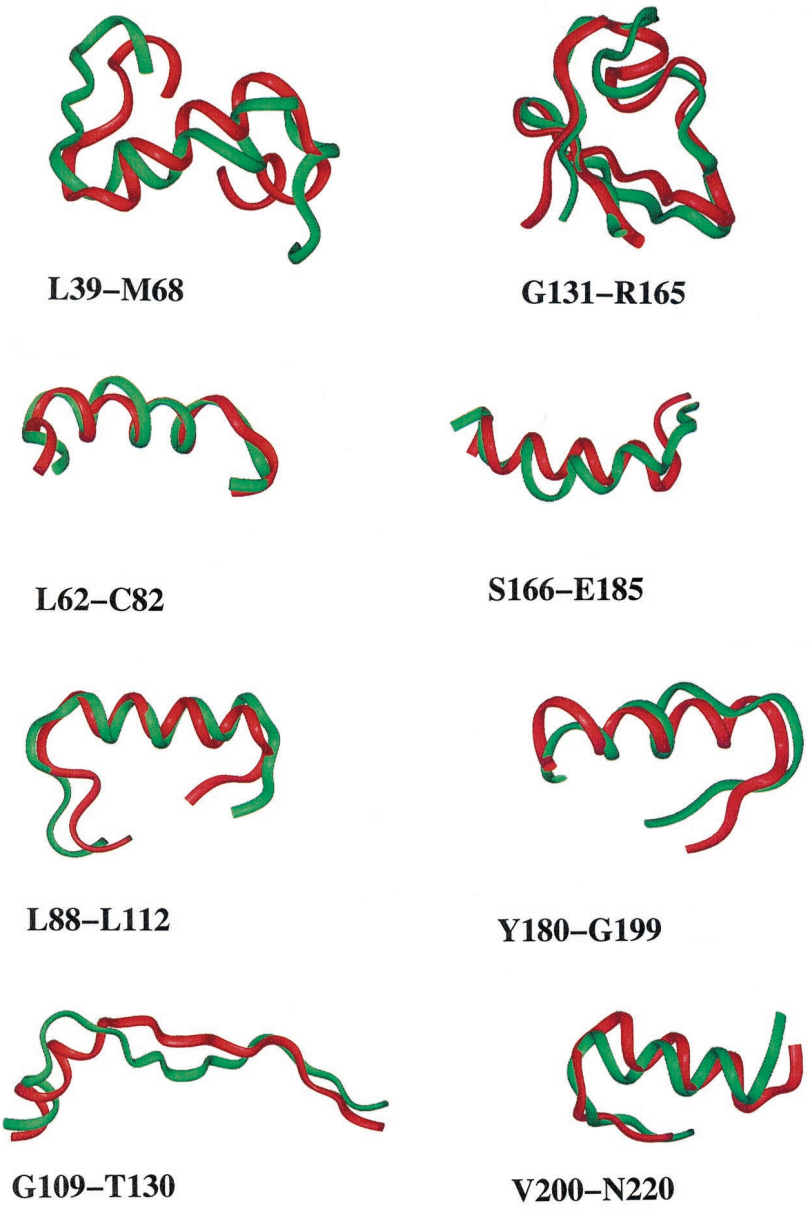


FIGURE 8 Superpositions of the initial (*red*) and final (*green*) conformations of eight building blocks (other than the N-terminal critical building block) at the lowest level of the anatomy of yeast ADK. These conformations are derived from the simulation of ADKΔCBB. Each building block is denoted by its beginning and end residue names in single letter code, and their numbers.

Taken together, the results from these molecular dynamics simulations indicate that our algorithm is able to correctly identify a critical building block in adenylate kinase. Furthermore, this simple surface area-based approach can discriminate between critical and noncritical elements in the protein core. Thus, presence in the protein core is insufficient for a building block to be critical for correct protein folding. It must interact with most of the other building blocks and mediate their interactions. As required by the definition, in the absence of the N-terminal critical building block the structure appears to quickly adopt a compact non-native state, but retaining the conformations of the local elements. Because the N-terminal building block contains the P-loop, its absence impairs the active site.

Consistent indications from experimental studies

Our proposition that the N-terminal building block in adenylate kinase may be critical for its correct folding is based on three consistent observations. First, our algorithm, which searches for such building blocks in protein structures, assigns this building block the highest CIndex value, significantly greater than for other building blocks (Table 1). Second, this building block shows a high degree of structural and sequence conservation among adenylate kinases (Table 2, Figs. 3 and 4). Third, molecular dynamics simulations indicate that its removal leads to the largest extent of disruption of the native adenylate kinase structure. Although there is no assurance that our molecular dynamics

simulations accurately model non-native structures in adenylate kinase, they do provide qualitative hints as to what may happen to the native adenylate kinase structure in the absence of the critical and other building blocks. Available experimental data are consistent with our observations.

Rose et al. (1991) have studied the structural and catalytic properties of a deletion derivative ($\Delta_{133–157}$) of *E. coli* adenylate kinase. This deletion removes the LID/INSERT domain of the enzyme. The modified ADK has a similar structure and thermal stability to the wild-type *E. coli* ADK. This is consistent with our observation that ADKW and ADKΔLID show similar behavior.

Saint Girons et al. (1987) have cleaved *E. coli* ADK into two fragments, 1–76 and 77–214, by exposure to alkaline pH of the cyanilated enzyme at the single Cys-77. They were able to recover both the catalytic activity and nucleotide binding properties only upon re-mixing the purified samples of both fragments. The isolated fragments were inactive, leading to the conclusion that the isolated fragments do not acquire their proper conformations. Cleaving the *E. coli* ADK at position 77 breaks its structure. The first 76 residues contain the AMP binding domain and a minor part of the CORE domain, including the N-terminal building block and a portion of the building block Leu-62–Cys-82. The second fragment contains the INSERT domain and a major part of the CORE domain including the noncritical but important building blocks Leu-88–Leu-112 and Gly-109–Thr-130. Saint Girons et al. (1987) have further concluded that the second fragment, 77–214, deviates significantly in its shape as compared to the corresponding native form. In our simulations, too, the ADKΔCBB fragment adopts a compact non-native conformation, even though it is ~40 residues longer.

Using site-directed mutagenesis, almost all the residues in the P-loop of adenylate kinase have been mutated and their structural and catalytic properties characterized. These mutants have been shown to have altered structures, susceptible to thermal denaturation and proteolysis (Reinstein et al., 1988; Muller and Schulz, 1993; Yoneya et al., 1989). The P-loop lies in the middle of the identified critical building block.

Despite these experimental studies, direct experimental evidence for our proposal for adenylate kinase is lacking. Nevertheless, the model and cutting algorithm have already been shown to correspond nicely with two experimental fragment complementation studies, the first on *E. coli* dihydrofolate reductase (Gegg et al. (1997) on the experimental as compared with Sham et al. (2001) on the computational side) and α-lactalbumin (Polverino de Laureto et al. (1999) on the experimental and Tsai et al. (2000) on the computational side).

Protein folding and protein function

The critical role of certain building block fragments in the folding of their corresponding proteins suggests that mutations in these regions will be disfavored. Consistently, we observe that the structures and sequences of the critical building blocks are more conserved than other fragments in the proteins. The fragment we have identified as critically important for folding is essential for function as well. Here we have shown it for ADK. An analogous case has also been observed for the N-terminus building block in dihydrofolate reductase (Ma et al., 2000; Sham et al., 2001). In this case, there is experimental evidence for the N-terminal building block (residues 1–36) being critical for correct protein folding (Gegg et al., 1997). In our procedure, this N-terminal fragment has the highest CIndex value. However, its CIndex value is significant by only one standard deviation (Z-score = 1.13). The CIndex value for a given protein fragment also depends upon the number of fragments in the protein. Hence, a critical fragment in a small protein, such as dihydrofolate reductase, may not have a large enough CIndex value to satisfy a statistical criterion for significance. Nevertheless, its importance toward protein structure is reflected in its high CIndex value.

The intramolecular chaperones constitute a third example. The proregion intramolecular chaperones are both critical for attaining the native fold, and fulfill an important biological function, by acting as inhibitors for their corresponding proteases, as in the cases of, e.g., α-lytic protease or subtilisin. Hence, critical building blocks may play a dual role in folding and in function.

These observations are not surprising. Although many more cases need to be examined to see to what extent this is a general phenomenon, the linkage among folding, binding, and function has been noted in numerous cases before. There are many examples of domains that are unstructured, i.e., existing in a range of conformations, most of which are non-native in solution. However, upon binding to a cofactor, ligand, inhibitor, or an ion, they reach the native state. Well-studied examples include the Ca^{2+} binding to the α-lactalbumin, nucleotide binding to the adenine binding domain in the dihydrofolate reductase, DNA binding to the GCN4, CRB binding to the kinase-induced activation domain of the CREB transcription factor, MDM2 binding to the acidic activation domain of p53, and RNA binding to RNA binding proteins. In all of these, the bound conformations are more stable than the unbound, driving the binding reaction. At the same time, they are critically essential for function. Hence, the linkage between being critically important for folding and for function, illustrated here for a fragment of the protein structure, is consistent with this simple evolutionary principle. For the protein, it implies guarding against mutational events largely in a single, given building block fragment, and thereby attaining both goals.

A search for critical building blocks in 930 nonhomologous protein chains whose crystal structures are available in the PDB (Bernstein et al., 1977) identifies at least one critical building block in each of the 225 dissimilar protein chains. Most of these protein fold in complex nonsequential

manner (Tsai et al., 1999b). An examination of the results of this large analysis shows that relatively large proteins tend to contain several building blocks with significant CIndex values. Hence, these proteins may have more than one critical building block. In several such cases, different building blocks lie in different protein domains. In many other cases, a single domain may contain more than one critical building block (Kumar and Nussinov, unpublished results). Removal of critical building block(s) from large proteins may result in larger structural changes. However, we have not yet explored these aspects. Sequence, structural conservation, and functional importance of the critical building blocks in large protein remain to be analyzed.

It may seem obvious that removal of a large enough fragment from a protein core would trigger collapse of the rest of the protein to fill in the "hole" created due to this removal. However, if the location in the protein core were a sufficient condition for a building block to be critical, the extent of the structural perturbation caused in the rest of the protein would be directly proportional to the size of the fragment removed from the core. Our results suggest otherwise. We have performed molecular dynamics simulations after removing two fragments, in addition to the N-terminus critical building block, from the core region of adenylate kinase. These fragments are Leu-62–Leu-112 and Gly-109–Thr-130. One of them is smaller than the critical building block (residues 1–36) and the other is larger. The structural perturbations produced by the removal of both these building blocks, one at a time, from the core region of adenylate kinase are smaller than that due to the removal of the N-terminal critical building block. Hence, the location in the protein core is a necessary but insufficient condition for a building block to be critical. In order to be critical for correct protein folding, the building block in question must interact with all (or most) of the other building blocks in the structure and mediate interactions among them, even if the other building blocks are sequentially connected. All of these conditions are satisfied by only the N-terminal fragment of adenylate kinase (Fig. 1 and Tables 3–5).

CONCLUSIONS

According to the hierarchical model of protein folding, folding initiates with local elements, which gradually assemble to yield the final native fold. Within this general model, the building block folding model considers these elements to be local minima along the sequence of the protein. Building blocks may be stable or unstable. However, even if unstable, the population times of the native conformations are still likely to be higher than of all alternates. The underlying premise in the model is that native contacts prevail during folding. This limits the conformational space search of the polypeptide chain. Building block fragments attain their preferred conformations, and via hierarchical combinatorial assembly, reach the final native fold.

However, not all building blocks play an equally important role in the folding process. Using a simple surface area-based approach, here we identify building blocks that may be critical for the protein structure. A critical building block should fulfill three criteria: 1) it should be buried in the protein core, with extensive contacts with most other building blocks in the structure, 2) it should preferably mediate interactions between sequentially connected building blocks, and most importantly, 3) in its absence the conformations of the individual building blocks are still preserved; however, they mis-associate.

We have devised an automated procedure to locate critical building blocks. Here we focus on adenylate kinase, a well-characterized protein. An ~30-residue-long critical building block is observed at its N-terminus, containing the ancient P-loop motif. This motif is conserved not only in adenylate kinases, but also in all related nucleoside monophosphate kinases. Molecular dynamics simulations of yeast ADK with the N-terminus building block removed show that the rest of the protein acquires a non-native stable conformation, the outcome of a mis-association of the rest of the building blocks in the ADK structure. However, the conformations of the building blocks themselves remain native-like. The shrunk conformation is likely to be inactive. A similar observation has also been made on the dihydrofolate reductase and on the proregion. However, the extent of the generality of this concept needs further examination.

We thank Dr. Buyong Ma for helpful discussions. In particular, we thank Dr. Jacob Maizel for encouragement throughout this project. The personnel at FCRDC are thanked for their assistance. The research of R. Nussinov in Israel has been supported in part by Grant 95-00208 from BSF, Israel, by a grant from the Ministry of Science, by the Center of Excellence, administered by the Israel Academy of Sciences, by the Magnet grant, and by the Tel Aviv University Basic Research and Adams Brain Center grants. This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under contract NO1-CO-56000.

REFERENCES

Abele, U., and G. E. Schulz. 1995. High resolution structures of adenylate kinase from yeast ligated with inhibitor Ap_5A, showing the pathway of phosphoryl transfer. *Protein Sci.* 4:1262–1271.

Baldwin, R. L., and G. D. Rose. 1999. Is protein folding hierarchic? I. Local structure and peptide folding. *Trends Biochem. Sci.* 24:26–33.

Bernstein, F. C., T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. 1977. The protein data bank: a computer based archival file for macromolecular structures. *J. Mol. Biol.* 112:535–542.

Berry, M. B., B. Meador, T. Bilderback, P. Liang, M. Glaser, and G. N. Phillips, Jr. 1994. The closed conformation of a highly flexible protein: the structure of *E. coli* adenylate kinase with bound AMP and AMPPNP. *Proteins.* 19:183–198.

Berry, M. B., and G. N. Phillips, Jr. 1998. Crystal structures of *Bacillus stearothermophilus* adenylate kinase with bound Ap_5A, Mg^{2+} Ap_5A and

Mn^{2+} Ap_5A reveal an intermediate lid position and six coordinate octahedral geometry for bound Mg^{2+} and Mn^{2+}. *Proteins.* 32:276–288.

Brooks, B. R., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. 1983. CHARMM: a program for macromolecular energy minimization and dynamic calculations. *J. Comput. Chem.* 4:187–217.

Brunger, A. T., and M. Karplus. 1988. Polar hydrogen positions in proteins: empirical energy placement and neutron diffraction comparison. *Proteins.* 4:148–156.

Byeon, I. J. L., Z. Shi, and M. D. Tsai. 1995. Mechanism of adenylate kinase. The "essential lysine" helps to orient the phosphates and the active site residues to proper conformations. *Biochemistry.* 34: 3172–3182.

Chan, H. S., and K. A. Dill. 1990. Origins of structure in globular proteins. *Proc. Natl. Acad. Sci. U.S.A.* 87:6388–6392.

Dreusicke, D., P. A. Karplus, and G. E. Schulz. 1988. Refined structure of porcine cytosolic adenylate kinase at 2.1 Å resolution. *J. Mol. Biol.* 199:359–371.

Dreusicke, D., and G. E. Schulz. 1986. The glycine-rich loop of adenylate kinase forms a giant anion hole. *FEBS Lett.* 208:301–304.

Elamrani, S., M. B. Berry, G. N. Phillips, Jr., and J. A. McCammon. 1996. Study of global motions in proteins by weighted masses molecular dynamics: Adenylate kinase as a test case. *Proteins.* 25:79–88.

Gegg, C. V., K. E. Bowers, and C. R. Matthews. 1997. Probing minimal independent folding units in dihydrofolate reductase by molecular dissection. *Protein Sci.* 6:1885–1892.

Gerstein, M., G. Schulz, and C. Chothia. 1993. Domain closure in adenylate kinase. Joints on either side of two helices close like neighboring fingers. *J. Mol. Biol.* 229:494–501.

Haney, P., J. Konisky, K. K. Koretke, Z. Luthey-Schulten, and P. G. Wolynes. 1997. Structural basis for thermostability and identification of potential active site residues for adenylate kinases from the archaeal genus *Methanococcus. Proteins.* 28:117–130.

Kabsch, W., and C. Sander. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers.* 22:2577–2637.

Kern, P., R. M. Brunne, and G. Folkers. 1994. Nucleotide binding properties of adenylate kinase from *Escherichia coli*: a molecular dynamics study in aqueous and vacuum environments. *J. Comp. Aid. Mol. Des.* 8:367–388.

Lazaridis, T., and M. Karplus. 1998. Discrimination of the native and misfolded protein models with an energy function including implicit solvation. *J. Mol. Biol.* 288:477–487.

Lazaridis, T., and M. Karplus. 1999. Effective energy function for proteins in solution. *Proteins.* 35:133–152.

Lim, W. A., and R. T. Sauer. 1991. The role of internal packing interactions in determining the structure and stability of a protein. *J. Mol. Biol.* 219:359–376.

Lim, W. A., D. C. Farrugio, and R. T. Sauer. 1992. Structural and energetic consequences of disruptive mutations in a protein core. *Biochemistry.* 31:4324–4333.

Ma, B., C. J. Tsai, and R. Nussinov. 2000. Binding and folding: in search of intra-molecular chaperone-like building block fragments. *Protein Eng.* 13:617–627.

Matte, A., L. W. Tari, and L. T. J. Delbaere. 1998. How do kinases transfer phosphoryl groups? *Structure.* 6:413–419.

Matthews, B. W. 1993. Structural and genetic analysis of protein stability. *Annu. Rev. Biochem.* 62:139–160.

Muller, C. W., G. J. Schlauderer, J. Reinstein, and G. E. Schulz. 1996. Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. *Structure.* 4:147–156.

Muller, C. W., and G. E. Schulz. 1992. Structure of the complex between adenylate kinase from *Escherichia coli* and the inhibitor Ap_5A refined at 1.9 Å resolution. A model for a catalytic transition state. *J. Mol. Biol.* 224:159–177.

Muller, C. W., and G. E. Schulz. 1993. Crystal structures of two mutants of adenylate kinase from *Escherichia coli* that modify the Gly-loop. *Proteins.* 15:42–49.

Polverino de Laureto, P., E. Scaramella, M. Frigo, F. G. Wondrich, V. De Filippis, M. Zambonin, and A. Fontana. 1999. Limited proteolysis of bovine alpha-lactalbumin: isolation and characterization of protein domains. *Protein Sci.* 8:2290–2303.

Reinstein, J., M. Brune, and A. Wittinghofer. 1988. Mutations in the nucleotide binding loop of adenylate kinase of *Escherichia coli. Biochemistry.* 27:4712–4720.

Reinstein, J., I. Schlichting, and A. Wittinghofer. 1990. Structurally and catalytically important residues in the phosphate binding loop of adenylate kinase from *Escherichia coli. Biochemistry.* 29:7451–7459.

Rose, T., M. Brune, A. Wittinghofer, K. L. Blay, W. K. Surewicz, H. H. Mantsch, O. Barzu, and A. M. Gilles. 1991. Structural and catalytic properties of a deletion derivative $\Delta_{133-157}$ of *Escherichia coli* adenylate kinase. *J. Biol. Chem.* 17:10781–10786.

Saint Girons, I., A. M. Gilles, D. Margarita, S. Michelson, M. Monnot, S. Fermandjian, A. Danchin, and O. Barzu. 1987. Structural and catalytic characteristics of *Escherichia coli* adenylate kinase. *J. Biol. Chem.* 262:622–629.

Saraste, M., P. R. Sibbald, and A. Wittinghofer. 1990. The P-loop—a common motif in ATP- and GTP-binding proteins. *Trends Biochem. Sci.* 15:430–434.

Schlauderer, G. J., K. Proba, and G. E. Schulz. 1996. Structure of a mutant adenylate kinase ligated with an ATP-analogue showing domain closure over ATP. *J. Mol. Biol.* 256:223–227.

Schulz, G. E. 1992. Binding of nucleotides by proteins. *Curr. Opin. Struct. Biol.* 2:61–67.

Schulz, G. E., C. W. Muller, and K. Diederichs. 1990. Induced-fit movements in adenylate kinase. *J. Mol. Biol.* 213:627–630.

Sham, Y. Y., B. Ma, C. J. Tsai, and R. Nussinov. 2001. Molecular dynamics simulation of *Escherichia coli* dihydrofolate reductase and its protein fragments: relative stabilities of the protein fragments in experiment and simulations. *Protein Sci.* 10:135–148.

Tsai, C. J., S. Kumar, B. Ma, and R. Nussinov. 1999a. Folding funnels, binding funnels and protein function. *Protein Sci.* 8:1181–1190.

Tsai, C. J., J. V. Maizel, and R. Nussinov. 1999b. Distinguishing between sequential and nonsequentially folded proteins: implications for folding and misfolding. *Protein Sci.* 8:1591–1604.

Tsai, C. J., J. V. Maizel, and R. Nussinov. 2000. Anatomy of protein structures: visualizing how a 1-D protein chain folds into a 3-D shape. *Proc. Natl. Acad. Sci. U.S.A.* 97:12038–12043.

Tsai, C. J., and R. Nussinov. 1997a. Hydrophobic folding units derived from dissimilar monomer structures and their interactions. *Protein Sci.* 6:24–42.

Tsai, C. J., and R. Nussinov. 1997b. Hydrophobic folding units at protein-protein interfaces: implications to protein folding and protein-protein association. *Protein Sci.* 6:1426–1437.

Tsai, C. J., D. Xu, and R. Nussinov. 1998. Protein folding via binding and vice versa. *Folding and Design.* 3:R71–R80.

Vonrhein, C., H. Bonisch, G. Schafer, and G. E. Schultz. 1998. The structure of a trimeric archaeal adenylate kinase. *J. Mol. Biol.* 282: 167–179.

Wu, L. C., R. Grandori, and J. Carey. 1994. Autonomous subdomains in protein folding. *Protein Sci.* 3:359–371.

Yan, H., and M. D. Tsai. 1999. Nucleoside monophosphate kinases: structure, mechanism, and substrate specificity. *Adv. Enzymol. Relat. Areas Mol. Biol.* 73:103–134.

Yoneya, T., M. Tagaya, F. Kishi, A. Nakazawa, and T. Fukui. 1989. Site-directed mutagenesis of Gly-1 and Gly-20 in the glycine-rich region of adenylate kinase. *J. Biochem. (Tokyo).* 105:158–160.