Identification of kinetically hot residues in proteins

MELIK C. DEMIREL,1 ALI RANA ATILGAN,1 ROBERT L. JERNIGAN,2 BURAK ERMAN,3 AND IVET BAHAR1,2
1Polymer Research Center, Bogazici University, and TUBITAK Advanced Polymeric Materials Research Center, Bebek 80815, Istanbul, Turkey
2Molecular Structure Section, Laboratory of Experimental and Computational Biology, Division of Basic Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892-5677
3Sabanci University, Sabanci Center 80745, and TUBITAK Advanced Polymeric Materials Research Center, Bebek 80815, Istanbul, Turkey

(RECEIVED February 20, 1998; ACCEPTED June 26, 1998)

Abstract

A number of recent studies called attention to the presence of kinetically important residues underlying the formation and stabilization of folding nuclei in proteins, and to the possible existence of a correlation between conserved residues and those participating in the folding nuclei. Here, we use the Gaussian network model (GNM), which recently proved useful in describing the dynamic characteristics of proteins for identifying the kinetically hot residues in folded structures. These are the residues involved in the highest frequency fluctuations near the native state coordinates. Their high frequency is a manifestation of the steepness of the energy landscape near their native state positions. The theory is applied to a series of proteins whose kinetically important residues have been extensively explored: chymotrypsin inhibitor 2, cytochrome c, and related C2 proteins. Most of the residues previously pointed out to underlie the folding process of these proteins, and to be critically important for the stabilization of the tertiary fold, are correctly identified, indicating a correlation between the kinetic hot spots and the early forming structural elements in proteins. Additionally, a strong correlation between kinetically hot residues and loci of conserved residues is observed. Finally, residues that may be important for the stability of the tertiary structure of CheY are proposed.

Keywords: CheY; chymotrypsin inhibitor; conserved residues; cytochrome c; folding pathway; hot residues; thermal fluctuations; vibrational dynamics

A number of single domain, small proteins are known to fold and unfold according to a two-state kinetics (Schmid, 1992; Fersht et al., 1994; Karplus et al., 1995; Fersht, 1997; Zwanitzig, 1997), while kinetic intermediates, or molten globule states generally form under mild denaturing conditions (Kim & Baldwin, 1990; Pitsyn, 1995). A typical example observed, both experimentally (Jackson et al., 1993; Otzen et al., 1994; Izhaki et al., 1995) and from molecular dynamics simulations (Li & Daggett, 1996; Daggett et al., 1996), to obey a two-state kinetics, is the chymotrypsin inhibitor 2 (CI2). The latter folds without accumulation of an intermediate or stepwise formation of secondary and tertiary structure. The two-state folding of small proteins has been attributed to a nucleation-condensation mechanism (Abkevich et al., 1994; Fersht, 1997). According to this mechanism, there is a weak local nucleus which may be represented by a few residues. The latter is stabilized by nonlocal, tertiary contacts. Also it has been shown that the embryonic secondary structure and stabilizing long-range contacts concurrently form the tertiary structure. This description departs from both the framework model of folding (Pitsyn & Rashin, 1975; Kim & Baldwin, 1990), which proposes the sequential formation of secondary and tertiary structures, and the hydrophobic collapse model (Dill et al., 1995) in which the secondary structure appears as a consequence of the overall condensation of the structure.

Although the nucleation-condensation mechanism generally pertains to the two-state kinetics of small proteins, it is conceivable that the same type of cooperativity between local propensities and tertiary contacts may be operative in the folding of independent structural units, or domains, in larger proteins. Theoretical and experimental studies further suggest that there may be a correlation between residues that are positionally conserved, and those revealed through protein engineering to play a key role in stabilizing tertiary folds. A method for predicting the kinetically important residues and supporting the existence of such a correlation was proposed by Shakhnovich et al. (1996). Therein, the most conserved residues of CI2 were found to be A35, I39, L68, V70, and I76, in agreement with experiments (Fersht, 1997): A35 participates in a nucleus on an α-helix, stabilized by contacts with L68 and I76.

Reprint requests to: Robert L. Jernigan, Molecular Structure Section, Laboratory of Mathematical Biology, Division of Basic Sciences, National Cancer Institute, National Institutes of Health, MSC 5677, Room B-116, Bldg. 12B, Bethesda, Maryland 20892-5677; e-mail: jernigan@structure.nci.nih.gov.
There is apparently a mechanism by which a nucleus, whose vital contacts are instituted upon passage over a transition state, preserves its structure throughout its path to the native state, possibly through a funnel-like energy landscape. The energy localization at residues forming the nucleus may be a reliable instrument preventing the transfer of energy from these to other residues. In fact, energy transfer to neighbors may be strongly hindered in the presence of nonperiodic, or disordered interactions, as first pointed out by Anderson (1958, 1978). Depending on the three-dimensional configuration, especially for medium-to-high-frequency modes (Dean & Bacon, 1963; Sayar et al., 1997), the mode shapes localize at certain loci, which may be termed “kinetically hot spots.”

Recently, we developed an analytical method for determining the dynamic characteristics of protein fluctuations in the folded state (Bahar et al., 1997, 1998a, 1998b; Haliloglu et al., 1997). The method, referred to as the Gaussian network model (GNM), is based on contact distributions in the folded state. The spectrum of vibrational modes ranging from fast fluctuations of isolated, individual residues, to cooperative, global motions of large domains, is obtained by eigenvalue analysis of the so-called Kirchhoff matrix of contacts, characteristic of each tertiary fold, as illustrated in the Appendix for a simple model chain. Our approach bears a close resemblance to an earlier study of Holm and Sander (1994). Therein, the slowest modes of motion were considered for identifying structural domains of proteins. Here, we concentrate on the fastest modes.

The fast modes of motion provide information about the kinetically hot residues. These are tightly packed residues, in general. They are trapped in severely constrained minima on the conformational energy landscape. We note that the term “hot spots” was used by Shoemaker et al. (1997) for describing the highly ordered protein whose dynamic characteristics have been studied in some detail, together with its application to chymotrypsin inhibitor 2 (CI2), a protein whose dynamic characteristics have been studied both experimentally (Jackson et al., 1993; Fersht et al., 1994; Otzen et al., 1994; Ithaki et al., 1995; Neira et al., 1997) and theoretically (Daggett et al., 1996; Li & Daggett, 1996; Shakhnovich et al., 1996). A perfect agreement between the kinetic hot residues identified by the GNM, and those previously pointed out to underlie the nucleation-condensation mechanism will be observed.

Calculations will be repeated for a series of cytochrome c’s and related C2 proteins, and for the chemotactic protein from Escherichia coli, CheY (Table 1). In contrast to CI2, which folds via a two-state mechanism, the folding/unfolding of cytochrome c involves intermediate states that have been observed by hydrogen exchange labeling coupled with NMR (Roder et al., 1988; Jeng & Englander, 1991; Bai et al., 1995; Kuroda et al., 1995; Sosnick et al., 1996), small angle X-ray scattering (Katoaka et al., 1993), time-resolved circular–dichroism, and fluorescence (Elowe et al., 1992). Here several proteins from the cytochrome c family will be considered, to further investigate the possible relationship between conserved residues and folding nuclei. Overall, a correlation will be pointed out between the kinetically hot residues, on the one hand, and formation and stabilization of folding nuclei, and beyond this, the conservation of residues.

Methods

Theory and illustration for chymotrypsin inhibitor 2

General description of the GNM and assumptions

The GNM is based on the following postulate (Bahar et al., 1997, 1998a, 1998b; Haliloglu et al., 1997): In folded proteins, residues undergo Gaussianly distributed fluctuations around their mean positions, due to harmonic potentials between all “contacting” residues. No residue specificity need be invoked as a first order approximation. Instead, the inter-residue potentials are all represented by the same single-parameter ($\gamma$) harmonic potential, as first proposed by Tirion (1996). Accordingly, the fluctuations $\Delta r_{ij}$ in distances between $C_\alpha$ and $C_\alpha'$ are controlled by the potential $V = \frac{1}{2} \gamma (\Delta r_{ij})^2$, $\alpha$-carbons being used as representative sites for residues. The dynamic characteristics of the molecule are fully described in this model by the so-called Kirchhoff matrix $\Gamma$ of contacts, the elements of which are

$$\Gamma_{ij} = \begin{cases} H(r_c - r_0) & i \neq j \\ -\sum_{i \neq j} \Gamma_{ij} & i = j \end{cases}.$$ (1)

Here $N$ is the total number of residues, $r_0$ denotes the distance between the $i^{th}$ and $j^{th}$ $C_\alpha$-atoms, and $H(x)$ is the Heaviside step function given as $H(x) = -1$ for $x > 0$ and $H(x) = 0$ for $x \leq 0$.

<table>
<thead>
<tr>
<th>Table 1. Proteins in the present study</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB code</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>2ci2</td>
</tr>
<tr>
<td>1hrc</td>
</tr>
<tr>
<td>1cer</td>
</tr>
<tr>
<td>1cot</td>
</tr>
<tr>
<td>1cxa</td>
</tr>
<tr>
<td>1cry</td>
</tr>
<tr>
<td>5cýt</td>
</tr>
<tr>
<td>3c2c</td>
</tr>
<tr>
<td>3chy</td>
</tr>
</tbody>
</table>
Defined in this way, $r_c$ is the cut-off distance or upper limit for the separation between two "contacting" residues.

The value $r_c = 7 \text{Å}$ has been adopted in previous applications of the GNM to proteins. This distance includes neighbors within the first coordination shell around a central residue, as revealed from analyses of known structures (Miyazawa & Jernigan, 1996; Bahar & Jernigan, 1997). Alternatively, one could select a broader distance, particularly in proteins containing $\beta$-sheets, so as to include the hydrogen-bond forming residue pairs of adjacent strands. The contact map for C12 is shown in Figure 1. C12 consists of a four-stranded $\beta$-sheet packed against an $\alpha$-helix, with a wide loop between strands 2 and 3, which contains the reactive site. The secondary structure elements are shown along the axes of Figure 1. The horizontal and vertical bars in the map indicate the nonzero elements of $\mathbf{r}$ obtained from the crystal structure (McPhalen & James, 1987) using $r_c = 7$ and 10 Å, respectively. Residue indices are in the range $19 \leq i \leq 83 = N$, corresponding to the numbers assigned in the crystal structure.

The $i$th diagonal element of $\mathbf{r}$ provides a measure of the local packing density near residue $i$, in terms of its coordination number, and the trace of $\mathbf{r}$ divided by $N$ gives the mean coordination number of the residues. Its inverse, $r_i^{-1}$, yields the correlations between the thermal fluctuations of the $\text{C}^\alpha$ atoms near the native state (Bahar et al., 1997); the diagonal elements of $r^{-1}$ scale with the mean-square (MS) fluctuations $\langle \Delta r_i^2 \rangle$, while the off-diagonal elements $(r^{-1})_{ij}$ refer to the cross-correlations $(\Delta r_i, \Delta r_j)$. The proportionality constant between $(r^{-1})_{ij}$ and $(\Delta r_i, \Delta r_j)$ is simply $3k_BT/\gamma$, where $k_B$ is the Boltzmann constant and $T$ is the temperature.

Recent applications of the GNM approach to a series of proteins showed that the predicted $(\Delta r_i^2)$ are in excellent agreement with the X-ray crystallographic temperature factors $B_i = 8\pi^2\langle \Delta r_i^2 \rangle / 3$ (Bahar et al., 1997). For illustrative purposes, the comparison of the $B$ factors predicted by the GNM model (thick solid line) and experiment (thin solid line) is presented in Figure 2 for C12, C2 protein isolated from Paracoccus denitrificans, and CheY. Furthermore, a strong correlation between hydrogen exchange free energy change measurements and GNM predictions was recently obtained for a series of proteins in their native states or under mild denaturing conditions (Bahar et al., 1998b). These comparisons support the use of GNM as a simple, yet physically reliable, means for characterizing the collective dynamics of proteins.

In the hydrogen exchange study, the free energy costs of exchange with the solvent for individual residues was shown to be inversely proportional to their MS fluctuations $\langle \Delta r_i^2 \rangle$. The $(\Delta r_i^2)$ values calculated therein result from the superposition of all $N$ modes. Here, the dynamics is decomposed into $N$ normal modes, and the shapes of the different modes are examined.

![Fig. 1. Contact map for chymotrypsin inhibitor 2 (C12). Secondary structure elements are shown along the axes. Horizontal and vertical bars on the map indicate the non-zero elements of $\mathbf{r}$ obtained from the crystal structure (McPhalen & James, 1987) (PDB code 2ci2) using the cutoff distances $r_c = 7$ and 10 Å, respectively. Residue indices along the two axes vary in the range $19 \leq i \leq 83 = N$.](image)

![Fig. 2. Comparison of theoretical (thick curve) and experimental (thin curve) $B$ factors for 2ci2, 1cot, and 3chy (see Table 1). The curves are normalized, i.e., the areas enclosed by both curves are equal. The inset displays the residue MS fluctuations driven by the slowest two modes. Smoother curves with peaks and minima corresponding to the most flexible and most restricted regions, respectively, in the global motions of the proteins are obtained by examining the slowest modes.](image)
Kinetically hot residues in proteins

**Dynamic fluctuation and correlation characteristics from GNM**

The dynamic characteristics of a given structure are described in terms of (1) its natural frequencies and (2) the shapes of the corresponding modes of motion. In the GNM, the former is given by the eigenvalues $\lambda_k$, $2 \leq k \leq N$ of $\Gamma$, excluding the zero eigenvalue $\lambda_1$, and the latter by the eigenvectors $u_k$, $2 \leq k \leq N$. Thus, it suffices for elucidating the dynamics of a protein in the folded state, to decompose its $\Gamma$ matrix into its eigenvalues and eigenvectors. The cross-correlation $\langle \Delta R_i \cdot \Delta R_j \rangle_{\lambda_k}$ associated with the $k^{\text{th}}$ mode of motion is (Bahar et al., 1997; Halliloglu et al., 1997)

$$\langle \Delta R_i \cdot \Delta R_j \rangle_{\lambda_k} = (3k_BT/\gamma)[\lambda_k u_i u_j^T]_\theta = (3k_BT/\gamma)\lambda_k^{-1}[u_i][u_j].$$

(2)

Here the subscripts designate the elements of the matrices (or vectors) enclosed in brackets. In the following, we will be interested in the MS fluctuations of residues ($i = j$) in a given mode $k$. The problem thus reduces to the calculation of $[u_i]$, as a function of $i$, for a given $k$, all other quantities in Equation 2 being invariant with respect to residue. The change in $[u_i]$, with residue $i$ referred to as the $k^{\text{th}}$ mode shape.

In general, the slowest modes give information on the global dynamics of the molecule (Bahar et al., 1999b). Domain movements or flexible loop motions, characterized by relatively longer correlation times, are identified by focusing on these modes (Holm & Sander, 1994), similarly to the common approach in normal mode analysis. Calculations for CI2, C2 protein from *P. denitrificans* and CheY yield the curves shown in the insets of Figure 2 for the MS fluctuations $\langle \Delta R_i \rangle^2_{\lambda_k}$ of residues in the two slowest modes of motion. For example, a smooth curve with a single peak including residues 54–62 on the active (binding) loop (McPhalen & James, 1987) emerges in CI2, revealing the expected correspondence between the global motion and function of the protein.

The fastest modes, on the other hand, reveal the residues subject to rapid, small amplitude fluctuations. Their high frequencies result from the steepness of the potential energy landscape surrounding their local minima. In the present study, we will concentrate on these modes. In contrast to slow modes, which occur cooperatively and cover segments of several contiguous residues along the backbone, fast modes affect individual residues, or residue pairs, which may be viewed as “hot spots,” crucial for the stability of the structure, as explained below.

**Chain connectivity vs. nonbonded contacts**

In the extreme case of a highly extended chain conformation, no contacts apart from those between the nearest neighbors along the chain occur, and the Kirchhoff matrix $\Gamma$ reduces to the classical tridiagonal Rouse matrix (Rouse, 1953), widely used for exploring the collective dynamics of polymers in the random coil state. The off-diagonal and the diagonal terms are $-1$ and $2$, respectively, in this case. The only exception is the first and last diagonal elements, which are equal to one. This form of $\Gamma$ represents the extreme case including solely the chain connectivity effect. The chain is conceived in this model as a linear sequence of springs with identical force constants $\gamma$, as illustrated in Figure 3A.

In globular structures, the one-dimensional geometric periodicity due to chain connectivity is broken by three-dimensional topological contacts. The matrix $\Gamma$ displays a unique, irregular form for each protein, depending on the distribution of nonbonded contacts. The coupling of residue $i$ to its nonbonded neighbors within a distance $r_i$ may be represented by the spring constant $\kappa_i$, as illustrated in Figure 3B.

An equivalent tridiagonal contact matrix may be created from the one that pertains to the native state of a protein. This is achieved by a Householder transformation, which reduces $\Gamma$ to a tridiagonal form after a sequence of $N - 2$ orthogonal transformations. To quantify the extent of irregularity introduced by nonbonded contacts, we may examine for each residue the ratio of the diagonal to off-diagonal terms in the equivalent tridiagonal contact matrix, and compare it to the value $-2$ of the Rouse matrix. For the one-dimensional system shown in Figure 3B, this ratio is $-(2\gamma + \kappa_i)/\gamma = -2 - \kappa_i/\gamma$. Thus, the difference $-\kappa_i/\gamma$ for each residue with respect to the value $-2$ describes the disorder introduced by nonbonded contacts.

Figure 4 illustrates the highest frequency mode shapes evaluated for CI2. Curve A (thin line) displays the regular shape of the highest frequency mode obtained for the chain devoid of nonbonded interactions. This is obtained simply from the $N^{\text{th}}$ eigenvector of the Rouse matrix of order $N$. Curve B (thick line), on the other hand, includes all nonbonded contacts between residue pairs separated by a distance $r_i = 10$ Å. Here, the X-ray $\alpha$-carbon coordinates of CI2 are used to construct $\Gamma$, and the elements of the $N^{\text{th}}$ eigenvector $[u_N]$ of $\Gamma$ are plotted. An extremely sharp peak at residue 68 is observed in curve B. This is a kinetically hot residue. This residue, together with Ile76, has been precisely pointed out to make tertiary contacts with the $\alpha$-helical residue Ala35 buried in the core, and thus stabilize the folding nucleus of CI2 (Itzhaki et al., 1995; Fersht, 1997). Curve A, on the other hand, is devoid of such hot spots, due to the absence of tertiary contacts. It covers the whole spectrum obeying a short-wavelength periodicity.

The highest frequency mode shapes

The analysis performed above for the fastest mode may be repeated for a series of high frequency modes. This permits identi-
protein engineering studies of Fersht and collaborators to initiate folding and stabilize the folding nucleus (Itzhaki et al., 1995; Fersht, 1997). Wolynes and coworkers estimated the hot residues of CI2 from a correlation analysis of a theoretical transition state ensemble (Shoemaker et al., 1997). In addition to Ala35, Leu68, and Ile76, they calculated Ile39 to be highly sensitive to mutations, similarly to our results. Ile39 and Ala35 are on adjacent turns of the \( \alpha \)-helix, which could explain the appearance of Ile39 among the set of hot residues. Likewise, the participation of Val66 and Phe69 among hot residues could result from the joint effect of chain connectivity (these being located on the same \( \beta \)-strand as Leu68 and Val70) and their tendency to be buried in the hydrophobic core.

In addition to these residues, there are weaker but noticeable peaks at residues Leu27 and Val32 participating in the type II reverse turn at the N-cap of the \( \alpha \)-helix, and at Pro80, the first residue of the terminal \( \beta \)-strand, which raises the question of the possible role of kinetically hot residues for initiating or stopping secondary structural elements in native folds.

We note that the residues presently identified as the kinetically hot spots in CI2 are strongly correlated with those found by Shakhnovich et al. (1996) to be conserved among sequences designed to fold into the same conformation. The sequence variation entropy values were also calculated at these sites using the HSSP database sequence alignments (Sander & Schneider, 1991). Relatively low entropy values (±0.30 after normalization to 1) are found for Ala35, Ile39, Val66, Val70, and Ile76. A correlation between residue conservation and kinetically hot spots is, therefore, clearly observed.

**Results and discussion**

**Results and discussion for cytochrome and C2 proteins, and for CheY**

An alternative way of examining the cumulative effect of the highest frequency modes is to combine the individual curves into a weighted average. From Equation 2, the MS fluctuations contributed by a given subset of modes \( k_1 \leq k \leq k_2 \) are

\[
(\langle \Delta R_i^2 \rangle)_{k_1 - k_2} = (k_B T/\gamma) \sum_{k_1}^{k_2} \lambda_k^{-1} \left[ \mathbf{u}_k \right]^2 / \sum_{k_1}^{k_2} \lambda_k^{-1}.
\]  

In the following we will examine the MS fluctuations \( (\Delta R_i^2)_{k_1 - k_2} \) induced by the modes \( N - 4 \leq k \leq N \) in the proteins listed in Table 1. Representative results obtained using \( r_c = 7.0 \) Å are displayed in Figure 6. Figure 6A displays the results for CI2, Figure 6B for horse cytochrome c, Figure 6C for protein C2 isolated from *P. denitrificans*, and Figure 6D for CheY. The fluctuation distributions are normalized, i.e., the MS values are divided by \( 3k_B T/\gamma \) so that the area enclosed by each curve is equal to unity. Peaks refer to kinetically hot residues. Usually one sharp peak is associated with each mode, as may be seen in Figure 5 for CI2, although two weaker peaks also appear in a few cases. This leads to about six peaks per protein in the cumulative distributions obtained for the five fastest modes, as listed in Table 2.

Ribbon diagrams of some representative proteins are shown in Figure 7. Therein, the residues identified as the kinetically hot spots, and hence implicated in possible nucleation-condensation mechanisms, are shown in red. Most of these residues make pair-
Kinetically hot residues in proteins

which residues having the same local packing density exhibited distinct fluctuation behavior. In the C2 protein (1cot) for example, residues 20 and 71 are not observed among the hot residues, although their coordination number (12) is equal to that of hot residues 110, 112–114.

Cytochrome c and related C2 proteins

Calculations performed for cytochrome c and related C2 proteins listed in Table 1 show that the two terminal helices, or more precisely a few specific residues in these helices, are distinguished by sharp peaks in the high frequency mode distributions (Fig. 6). The normalized MS fluctuations for horse cytochrome c (1hrc) residues are displayed in Figure 6B, as representative of other cytochrome c's (e.g., tuna, bonito, yeast) that are closely superimposable both spatially and sequentially. The curve for the C2 protein (1cot) is displayed in Figure 6C. The results for the other proteins in this group (1cct, 1cry, 5cry, and 3c2c) are summarized in Table 2.

In all cases, the set of hot residues comprise a glycine (G6 or G8) in the N-terminal α-helix, four or five residues in the C-terminal helix, together with a histidine (or a tryptophan in 1cry, and a methionine in 3c2c), which ligate the iron atom at the center of the heme group in the central portion of the molecule. The C-terminal helix is the most stable region in native cytochrome c (Bai et al., 1995) possessing helical structure even in isolated form, which is consistent with the appearance of several residues belonging to this helix in Table 2. The key residues of this helix generally lie between an aspartic acid (or asparagine in 1cxa and 3c2c) at one end, and an aromatic residue, Phe or Tyr, at the other. These might form the embryonic secondary structure, or nucleus, to be stabilized by tertiary contacts. In fact, close contacts with the hot residues G6 or G10 occur in the folded state, forming the clusters listed in the last column of Table 2. The histidines do not participate in these clusters. Although the residue indices differ therein, sequence alignments using matrices of contacts demonstrate (Ptitsyn, 1998) that the residues participating in these clusters are equivalent, i.e., have the same geometry of contacts in all five cases.

The packing of residues in the C- and N-terminal helices is known to play a critical role in the kinetics of folding (Ellobe et al., 1992; Colón et al., 1996). These helices acquire about 60% protection within the 20 ms burst phase of refolding (Roder et al., 1988), and also control the global unfolding of the molecule (Bai et al., 1995; Englander et al., 1997). In particular, the L94A mutation is observed to block all folding steps after the initial collapse (Colón & Roder, 1996), in conformity with the present identification of L94 as a kinetically important residue. The acid-denatured form (A-form) of cytochrome c is stabilized by the interactions between the conserved residues at the interface of these helices (e.g., Gly6-Leu94) (Marmorino & Pielak, 1995). Our recent GNM calculations of hydrogen exchange free energy changes for cytochrome c also revealed the dominant role of the presently identified hot residues (Bahar et al., 1998b).

Apart from a few residues at the terminal helices, our calculations indicate the important role of a histidine, or tryptophan, buried near the heme-binding site. The His residue at position 19 (or 18 in 1hrc) has a very low sequence variation entropy (≤0.06 after normalization to 1) as determined from HSSP database sequence alignments (Sander & Schneider, 1991), which indicates strong conservation at this position. We note that Raman scattering measurements coupled with submillisecond mixing techniques also
Table 2. Kinetically hot residues and their clusters/contacts in the folded state

<table>
<thead>
<tr>
<th>PDB code</th>
<th>Kinetically hot residues&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Clusters/contacts&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>2ci2</td>
<td>A35*, V66, L68*, V70*, I76*</td>
<td>{68, 70, 76}, {66,68}</td>
</tr>
<tr>
<td>lhrc</td>
<td>G6*, H18*, D93, L94*, A96, Y97*</td>
<td>{6, 93, 94, 96, 97}</td>
</tr>
<tr>
<td>lcot</td>
<td>G8*, H19*, D110, V111*, V112, F114*</td>
<td>{8, 110, 111, 112, 114}</td>
</tr>
<tr>
<td>lcry</td>
<td>G6*, W58, D92, L93*, A95, F100, N101</td>
<td>{6, 92, 93}, {100,101}</td>
</tr>
<tr>
<td>lct</td>
<td>G6*, H18, D93, L94*, A96, Y97*</td>
<td>{6,93,94,96,97}</td>
</tr>
<tr>
<td>lcyc</td>
<td>G7*, H18, M55, N103,V104*, A106, T107*</td>
<td>{7,103,104,106,107}</td>
</tr>
<tr>
<td>lcyt</td>
<td>V11*, R18, D38, D57*, D64, G65*, M85*</td>
<td>(11,18), (11,38), (11,57), (57,65), (57,85)</td>
</tr>
</tbody>
</table>

<sup>a</sup>Residues whose normalized MS fluctuations in the high frequency modes are above 6N<sup>−1</sup>. See Figure 6. Highly conserved residues are indicated by an asterisk.

<sup>b</sup>Clusters are shown in braces. Each pair of residues in a cluster is separated by ≤7 Å. Contacts, shown in parentheses, refer to pairs within 7 Å. Distances refer to α-carbon separations.

showed the important role of histidine heme-ligand interactions, and the possible trapping in a misligated form determining folding kinetics (Takahashi et al., 1997; Yeh et al., 1997), in parallel with earlier observations (Elöve et al., 1992).

As a final remark, we note that the kinetically hot residues listed in the last column of Table 2 exhibit a close correlation with the most conserved residues recently identified for the same proteins (Piitsyn, 1998). Residues G6, F10, L94, and Y97 in lhrc, G14,

Fig. 7. Ribbon diagrams of 2ci2, lhrc, lcot, and 3chy. The residues presently identified as the kinetically hot residues are shown in red. Contacts between hot residues are indicated by yellow dashed lines.
F18, L102, and Y105 in 1c5r, G8, F12, V111, F114 in 1c2t, G6, F10, L93, Y96 in 1cry, and G8, F12, I113, and Y116 in 1cxa were indicated by the asterisks in the second column of Table 2. The ally three are identified by GNM as kinetically hot. These are the highest peak succeeding D110 when the force constant y associated with inter-residue interactions was rescaled with the number of atom-atom contacts within 6.0 Å. This observation suggests a possible direction for improving the GNM model in future studies.

Results for CheY

CheY is an α/β bacterial signal transduction protein of 129 amino acids from E. coli (Stock et al., 1990). It contains two subdomains: the first includes the first two β-strands (residues 7–11 and 32–36) and two α-helices (15–26 and 39–46), and the second the remaining three β-strands (53–57, 82–87, and 106–108) and three α-helices (65–73, 92–99, and 113–125). It is reported that the structure of the first subdomain in the transition state resembles that observed for C12 (López-Hernández & Serrano, 1996) and the second subdomain is unstable until the first has folded (Muñoz et al., 1994; López-Hernández & Serrano, 1996).

The folding pathway of CheY exhibits a kinetic intermediate as does cytochrome c. However, the results of the present method as presented in Figures 6D and 7D indicate that the structure of the core formed by the hot residues is different from that of cytochrome c. Essentially, the hot residues are distributed over a broader distance range, compared to those in cytochrome c that are precisely located at the terminal helices. To clarify the distinction, we show in Figure 8 a ribbon diagram of CheY, in which the kinetically hot residues are explicitly labeled, and the two subdomains, the first shown in blue and the second in gray, are distinguished. The kinetically hot residues consists of two subgroups, each composed of three residues. The former three, V11, R18, and D38, located in the first subdomain (lower pari), are shown in red, and those in the second subdomain, D57, G65, and M85, are shown in green. The nonlocal interaction between V11 and D57 establishes the communication between the hot residues of the two subdomains.

López-Hernández and Serrano (1996) found that the packing of the first α-helix against the first two β-strands forms the nucleus around which CheY folds. The nucleus mentioned in that study may be associated with the first group of hot residues presently found. However, a second group of residues, located in the second subdomain, is identified in our approach. It is interesting to observe that the key residue D57, which is presently shown to be active in the communication of the two subgroups of hot residues in the different domains, is found by Stock et al. (1990) to be highly conserved among the family of bacterial signal transduction proteins including CheY. Two other highly conserved residues are shown therein to be D13 and K109; also, stretches of residues forming a conserved hydrophobic core are reported. The latter includes the kinetically hot residues V11 and M85 presently identified. Besides, we note that G65 is also conserved among all of the 20 reported homologous regions of response regulators aligned against CheY, except one in which it is replaced by Ala, and that residue 38 exhibits a strong preference to be aspartic acid or asparagine. We also note that the sequence variation entropy values for D57 and G65 calculated from the HSSP database (Sander & Schneider, 1991) are identically zero, the residue identity at these two positions being completely conserved in all aligned sequences.

Conclusion

In the present study, a systematic method for identifying the kinetically hot residues in folded proteins, is presented, based on the GNM of proteins. These are generally tightly packed residues. Thus, the coordination number of individual residues is an important property dominating the fluctuation behavior. In addition to their relatively high local packing density, residues presently identified as hot spots experience an extremely strong coupling to all other residues within the particular network topology of native nonbonded contacts. They are distinguished after decomposition of the protein dynamics into a set of collective modes, and examination of the residue fluctuations driven by the highest frequency/smallest amplitude modes.

The occurrence of high frequency modes is associated with the steepness of the energy landscape in the neighborhood of the local minima corresponding to the equilibrium (native) positions of residues. Any departure from these mean positions is strongly opposed due to the steepness of the surrounding energy walls. Such centers cannot efficiently exchange energy with their surroundings, and thus preserve their state despite changes in other parts of the structure, which suggests their possible involvement at the nucleation-condensation stage of folding. Our recent study of hydrogen exchange behavior also showed that residues subject to
small amplitude/high frequency fluctuations cannot efficiently exchange with the solvent (Bahar et al., 1998b).

The residues identified in Table 2 and displayed in Figure 7 are from application of the GNM method to C12, cytochrome c, and related C2 proteins. Most of these residues were pointed out in previous theoretical and experimental studies to be crucial for the formation and/or stability of the tertiary fold. They either participate in the formation and stabilization of a folding nucleus, as in C12 obeying a two-state folding kinetics, or exhibit native-like contacts in the intermediate or molten globular states formed on the folding pathway. Examples of the latter case are cytochrome c and CheY. Thus, irrespective of the folding mechanism, it is possible to describe a nucleus or core that consists of relatively small numbers of residues forming native-like contacts. As recently pointed out by Pitsyn (1996), the tightly packed residues in intermediates may include the folding nucleus, or coalesce with it. Furthermore, even proteins unfolding via a noncooperative mechanism may exhibit a core, or a nucleus, preserving its native structure, as observed in recent NMR experiments of α-lactalbumin unfolding by Schulman et al. (1997).

Therein a core region that remains collapsed under extreme denaturating conditions is detected, which supports a model in which the core provides a template for the correct assembly of the rearranged protein.

To conclude, the present method identifies the residues, pairs or clusters, which exhibit highly stabilized, key interactions, in ter-

Acknowledgments

Partial support from Bogazici University Research Funds A970401 and NATO Collaborative Research Grant Project #CRG951240 is gratefully acknowledged.

References


Otzen DE, Izhaki LS, ELMasry NF, Jackson S, Fersht AR. 1994. Structure of
Kinetically hot residues in proteins


Appendix

The GNM calculations are illustrated here for a simple model of eight residues. We consider the following model protein on a square lattice

A quadratic form of energy is assumed for the fluctuation of all residues, in which the interactions between all “contacting” sites are represented with the same single parameter. Contacting residues are those located on adjacent sites, bonded, or nonbonded. The corresponding Kirchhoff matrix \( \mathbf{\Gamma} \), defined by Equation 1, is

\[
\mathbf{\Gamma} = \begin{bmatrix}
3 & -1 & 0 & -1 & 0 & 0 & 0 & -1 \\
-1 & 2 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & -1 & 2 & -1 & 0 & 0 & 0 & 0 \\
-1 & 0 & -1 & 4 & -1 & 0 & -1 & 0 \\
0 & 0 & 0 & -1 & 2 & -1 & 0 & 0 \\
0 & 0 & 0 & 0 & -1 & 2 & -1 & 0 \\
-1 & 0 & 0 & 0 & 0 & -1 & 3 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & -1 & 2
\end{bmatrix}
\]

The eigenvalues \( \lambda_i \) and eigenvectors \( \mathbf{u}_i \) of \( \mathbf{\Gamma} \), \( 1 \leq i \leq 8 \), are computed using a standard eigenvalue decomposition subroutine (Press et al., 1992). The eigenvalues are found to be 0, 0.75, 1, 2, 4, 2.52, 2.52, 3.44, 3.44, 3.80, 5.76, in ascending order. The latter refers to the highest frequency mode, and \( \lambda_1 = 0.75 \) is associated with the slowest mode. The zero eigenvalue is due to the translational invariance of the system. In other words, the elements of \( \mathbf{\Gamma} \) depend on the distances between the residues, only; they are not altered during rigid body motion. In the evaluation of the \( B \) factors, the inverse of \( \mathbf{\Gamma} \) is obtained after eliminating the zero eigenvalue, i.e., \( B_i = \mathbf{u}_i^T (\mathbf{\Delta R}) \mathbf{u}_i \), where the summation is performed over all nonzero eigenvalues \( 2 \leq i \leq 8 \), and the contribution of the \( k^{th} \) mode is given by Equation 2.

The eigenvectors corresponding to this simple model are plotted in Figure 9. The ordinate values are not explicitly shown, as the shapes, rather than the absolute values, are of interest. The absolute values may be readily found using \( \sum_i |\mathbf{u}_i|^2 = 1 \) for each eigenvector \( \mathbf{u}_i \). Here \( |\mathbf{u}_i| \) is the \( i^{th} \) element (associated with the \( i^{th} \) residue) of \( \mathbf{u}_i \). The eigenvector associated

Fig. 9. Mode shape for the simple on-lattice structure presented in the Appendix. The curves display the eigenvectors \( \mathbf{u}_k \), \( 2 \leq k \leq 8 \), as a function of residue index. Each element \( \mathbf{u}_k \) corresponds to a given residue \( i \) along the abscissa.
with the zero eigenvalue ($\lambda_1$), which is a straight line, is not shown. The
eigenvector for the second smallest eigenvalue ($k = 2$) reflects the coop-
erative motion of the molecule. The fastest mode shape ($k = 8$), on the
other hand, exhibits a pronounced peak at the fourth residue, the kinetically
hot residue in the present simple model. This is the most tightly packed
residue in the present simple model. For molecules with a larger number of
residues, the mode shapes associated with the higher frequencies display

sharper peaks at hot loci, as shown for C12, for instance, in Figure 4, and
are affected by the specific tertiary contacts, in addition to packing density.
The general shapes closely resemble the modes of a vibrating string, in
having approximately $k - 1$ axis crossings for mode $k$.

The FORTRAN code for calculating the $B$ factors of PDB structures
using GNM, and for evaluating the shapes of all modes is available on the
internet (http://klee.bme.boun.edu.tr/code.f).