The nucleation-collapse mechanism in protein folding: evidence for the non-uniqueness of the folding nucleus

Zhuyan Guo[1] and D Thirumalai[2,3]

Background: Recent experimental and theoretical studies have shown that several small proteins reach the native state by a nucleation-collapse mechanism. Studies based on lattice models have been used to suggest that the critical nucleus is specific, leading to the notion that the transition state may be unique. On the other hand, results of studies using off-lattice models show that the critical nuclei should be viewed as fluctuating mobile structures, thus implying non-unique transition states.

Results: The microscopic underpinnings of the nucleation-collapse mechanism in protein folding are probed using minimal off-lattice models and Langevin dynamics. We consider a 46-mer continuum model which has a native β-barrel-like structure. The fast-folding trajectories reach the native state by a nucleation-collapse process. An algorithm based on the self-organized neural nets is used to identify the critical nuclei for a large number of rapidly folding trajectories. This method, which reduces the determination of the critical nucleus to one of 'pattern recognition', unambiguously shows that the folding nucleus is not unique. The only common characteristics of the mobile critical nuclei are that they are small (containing on average 15–22 residues) and are largely composed of residues near the loop regions of the molecule. The structures of the transition states, corresponding to the critical nuclei, show the existence of spatially localized ordered regions that are largely made up of residues that are close to each other. These structures are stabilized by a few long-range contacts. The structures in the ensemble of transition states exhibit a rather diverse degree of similarity to the native conformation.

Conclusions: The multiplicity of delocalized nucleation regions can explain the two-state folding by a nucleation-collapse mechanism for small single-domain proteins (such as chymotrypsin inhibitor 2) and their mutants. Because there are many distinct critical nuclei, we predict that the folding kinetics of fast-folding proteins will not be drastically changed even if some of the residues in a 'typical' nucleus are altered.

Addresses: [1]Department of Molecular Biology (MB19), The Scripps Research Institute, 10666 North Torrey Pines Road, La Jolla, CA 92037-1027, USA. [2]Institute for Physical Science and Technology and [3]Department of Chemistry and Biochemistry, University of Maryland, College Park, MD 20742, USA.

Correspondence: D Thirumalai
E-mail: thirum@glue.umd.edu

Key words: fluctuating nucleation structures, multiple transition states, non-unique nucleus, nucleation-collapse process

Received: **08 August 1997**
Revisions requested: **28 August 1997**
Revisions received: **16 September 1997**
Accepted: **26 September 1997**

Published: **10 November 1997**
http://biomednet.com/elecref/1359027800200377

Folding & Design 10 November 1997, **2**:377–391

Introduction

It has often been suggested that proteins reach their native conformation from a denatured state by a nucleation mechanism [1,2]. This proposal, based on phenomenological grounds, has been used to rationalize the rapid folding of proteins. Only recently, however, have studies using minimal protein models begun to clarify the microscopic origins of the nucleation mechanism [3–6]. Evidence for the nucleation-collapse processes in protein folding has been given for chymotrypsin inhibitor 2 (CI2), cytochrome *c* and lysozyme [7–10].

In two recent papers [4,11], using off-lattice models and Langevin dynamics, we have described nucleation collapse as one of the possible pathways that leads the protein molecule to its native conformation very rapidly. From consideration of topological frustration [12] it follows that generically there are off-pathway processes that slow down the rates of folding. The multitude of paths leading to the native conformation can be summarized by the kinetic partition mechanism [4,13]. According to the KPM a fraction Φ (the partition factor) of molecules reaches the native conformation by a direct native-conformation nucleation-collapse mechanism. The remaining fraction of molecules, 1–Φ, reaches the native state by a more complex three-stage multiple-pathway mechanism [14]. Based on theoretical arguments it has been suggested [15] that, in situations when the polypeptide chain reaches the native conformation by the nucleation-collapse mechanism, the acquisition of the entire native-state topology is almost synchronous with the collapse process itself.

The nucleation mechanism has also been suggested using lattice Monte Carlo simulations [3,6]. In the original study

Abkevich *et al.* [3] proposed that the 'nucleation-growth' mechanism applies only to the formation of the molten-globule conformation and not to the native state. Apparently these authors now assert (cited as footnote on page 284 of [8]) that the nucleation mechanism is more general, in accord with our earlier findings [13,15] and those of Fersht and others [7,9]. Because the models and the dynamics used in the studies of Abkevich *et al.* [3] are very different from those used in our work, it appears that the nucleation-collapse mechanism leading to the rapid formation of the native conformation should be a general feature of protein folding [1,2,16].

Studies on a variety of minimal models suggest that the transition from an ensemble of collapsed compact structures to the native conformation at the folding transition temperature T_f is a first-order phase transition suitably modified by finite size effects [14,17,18]. Due to the first-order nature of the folding transition we expect that the nucleation process should be operative in the folding of biomolecules in general [12] and proteins in particular. The natural question that arises is: what is the nature of the critical nucleus in proteins? The simple physical picture of an homogeneous nucleation process that occurs when a liquid is cooled below its melting point is now described [19]. The atoms or molecules start to form clusters of varying sizes and whenever a cluster of some critical size is formed it can either spontaneously grow to the crystalline phase or decay to a smaller size nucleus having liquid-like characteristics. Because the liquid is a macroscopic system, the process of nucleation can be initiated anywhere. In fact one could argue that there could be several microscopic structures for the critical nucleus, but it is possible that only a subset of them are 'committed' to crystallization. Proteins, on the other hand, are finite systems. This fact and the observation that the critical nucleus involves a relatively small number of residues [1,2,4] suggest that there cannot be many microscopic structures for the critical nucleus. There is an additional constraint. Because the critical nucleus is committed to folding, the topology of this substructure has to be similar to the native state. It has been shown [4] that the small number of critical mobile nuclei can originate anywhere in the polypeptide chain. There is a preference, however, for these nuclei to be near the turn regions of the folded structures [4]. The aim of this study is to provide an additional microscopic basis to probe the nature of the nucleating regions using an off-lattice α-carbon representation of a protein. We use self-organized neural nets [20,21] as a computational method to probe the dynamics of the formation of the critical nucleus. This technique has already been used by Karpen *et al.* [22] to analyze the conformational substates using long molecular dynamics trajectories of peptides. This method reduces the problem of identifying the critical nucleus to one of 'pattern' recognition.

Results

Methods

Model and simulation methods

The details of the model and simulation methods have been given in our earlier studies [4,23] and we will, therefore, only provide a brief summary here. The continuum model has some, but not all, of the features found in proteins. The α-carbon representation used in our studies consists of M connected beads (each bead represents the α-carbon of a residue) with M = 46. The features of the model are broadly consistent with the data extracted from the Protein Data Bank [24]. We have used a three-letter code to identify a bead as hydrophobic (B), hydrophilic (L) or neutral (N). A sequence is defined by the specific way in which the various beads are strung together.

The simulations were carried out using 'noisy molecular dynamics' by including a frictional force (proportional to the velocity of the residue) and a Gaussian random force in Newton's equation of motion. The resulting equations of motion [4] were solved using the velocity form of the Verlet algorithm. The dynamics scheme we have used here is the same as the Langevin dynamics in the limit of low friction. We measure temperature in units of ε_h/k_B (ε_h is a measure of the hydrophobic interaction and k_B is the Boltzmann constant) and time is given in units of:

$$\tau = \left(\frac{m\sigma^2}{\varepsilon_h} \right)^{1/2} \qquad (1)$$

where σ is the average distance between the residues. The step size used in the integration of the equations of motion is typically 0.005τ, and the value of ζ is taken to be $0.05\tau^{-1}$.

Statistical clustering techniques to identify nucleation regions

We used a nonhierarchical clustering algorithm based on a self-organizing neural net [20–22] to identify the nucleation sites or regions. This algorithm provides a nonbiased partitioning of the various conformations of a dynamical trajectory. Thus, preferred nucleation sites can be identified without resorting to an arbitrary definition of a nucleus.

Consider a trajectory generated from a given initial condition corresponding to one molecule of the ensemble of denatured chains. If h is the stepsize used in the simulations one generates T_{max}/h number of conformations, where T_{max} is the duration of the trajectory. From the total number of conformations we select a subset containing N_c conformations for analysis. Let us parameterize a given conformation j by a vector with P elements $x_j = [x_{1j}, x_{2j}, \ldots x_{Pj}]$. Thus, any characteristic of the conformation deemed important for a particular application may be used. The basic idea of the neural net algorithm [22] is to partition the various conformations — each specified by the vector j, $x_j (j = 1, 2, \ldots N_c)$ — into distinct clusters. The clusters

are described by the cluster center and the size of the cluster is determined by a radius R. The center of the k^{th} cluster is determined by $C_k = [C_{1k}, C_{2k}, \ldots C_{Pk}]$ with:

$$C_k = \frac{1}{P_k}\sum_{j=1}^{P_k} x_j \qquad (2)$$

where P_k is the total number of conformations in C_k. Conformation j belongs to C_k if the Euclidean distance:

$$d_{jk} = |x_j - C_k| \qquad (3)$$

which specifies that the 'distance' between conformation j and the cluster center k is less than a preassigned value R (i.e. $d_{jk} < R$). In this study we used R = 14.05. Thus, the algorithm compares each conformation characterized by the vector x_j to the set of all cluster centers using Equation 2. The cluster with the minimum value of d_{jk} (i.e. the closest cluster) is determined. If the distance from the conformation to cluster k is greater than R (i.e. $d_{jk}/R > 1$) for all values of k (i.e. there is no cluster to which the conformation j can be assigned) then the conformation is assigned to a new cluster.

There are two phases to the implementation of the algorithm. In practice, the first step in the algorithm consists of grouping the various conformations into clusters. In order to do this the N_c conformations to be analyzed are considered one at a time. The first conformation clearly constitutes a cluster on its own. The Euclidean distance between the second conformation and the first is then computed using Equation 3. If $d_{12} < R$ then the second conformation is assigned to the same cluster as the first and the cluster center is calculated according to Equation 2. If $d_{12} > R$ then the second conformation forms a new cluster. This process is continued until all the N_c conformations are partitioned into distinct clusters. Notice that in this initial 'learning phase', cluster centers are recalculated as each new conformation is added. After assigning all the conformations one has N_{max} number of clusters.

The second 'refining' phase is implemented iteratively as follows. The centers of all clusters are fixed to the values $C_1^0, C_2^0, \ldots C_{N^0_{max}}$ which are the values obtained at the end of the learning phase. Now each of the conformations j (j = 1, 2, ... N_c) in every cluster is reexamined by calculating d_{jk} according to Equation 3. If there is another distinct cluster k′ for which $d_{jk'} < d_{jk}$ then conformation j is reassigned to cluster k′. With this reassignment, new cluster centers $C_1^1, C_2^1, \ldots C_{N^1_{max}}$ are computed. This process of reassignment is continued until the values of the cluster centers do not change. In order to implement the above algorithm we need the vector x_j (j = 1, 2, ... P), which characterizes the conformation j. Because we are interested in finding the potential nucleation sites or regions we use the distances between nonbonded B beads to describe a given conformation. The number of pairs of B beads is far too large to be manageable in the analysis. In order to restrict the number used in the algorithm without losing any information we choose certain pairs of nonbonded B beads that are separated by $< 2\sigma$ in the native state. Thus, we identify pairs of hydrophobic residues that are within 2σ using the native conformation and we describe a given structure using the distances between the identical pairs of B beads. The native conformation of our model would be uniquely specified if the distance between the pairs of hydrophobic residues used in our analysis was specified. Thus, the elements of the vector x_j are the distances between the B beads that in the native conformation are separated by $< 2\sigma$. In our analysis we used P = 49.

The clusters that are obtained at the end of the refining phase are relabeled by calculating the distance of the cluster center C_k to the vector $x_0 = [x_{01}, x_{02}, \ldots x_{0M}]$ that characterizes the native conformation. In other words:

$$r_k = |C_k - x_0| \qquad (4)$$

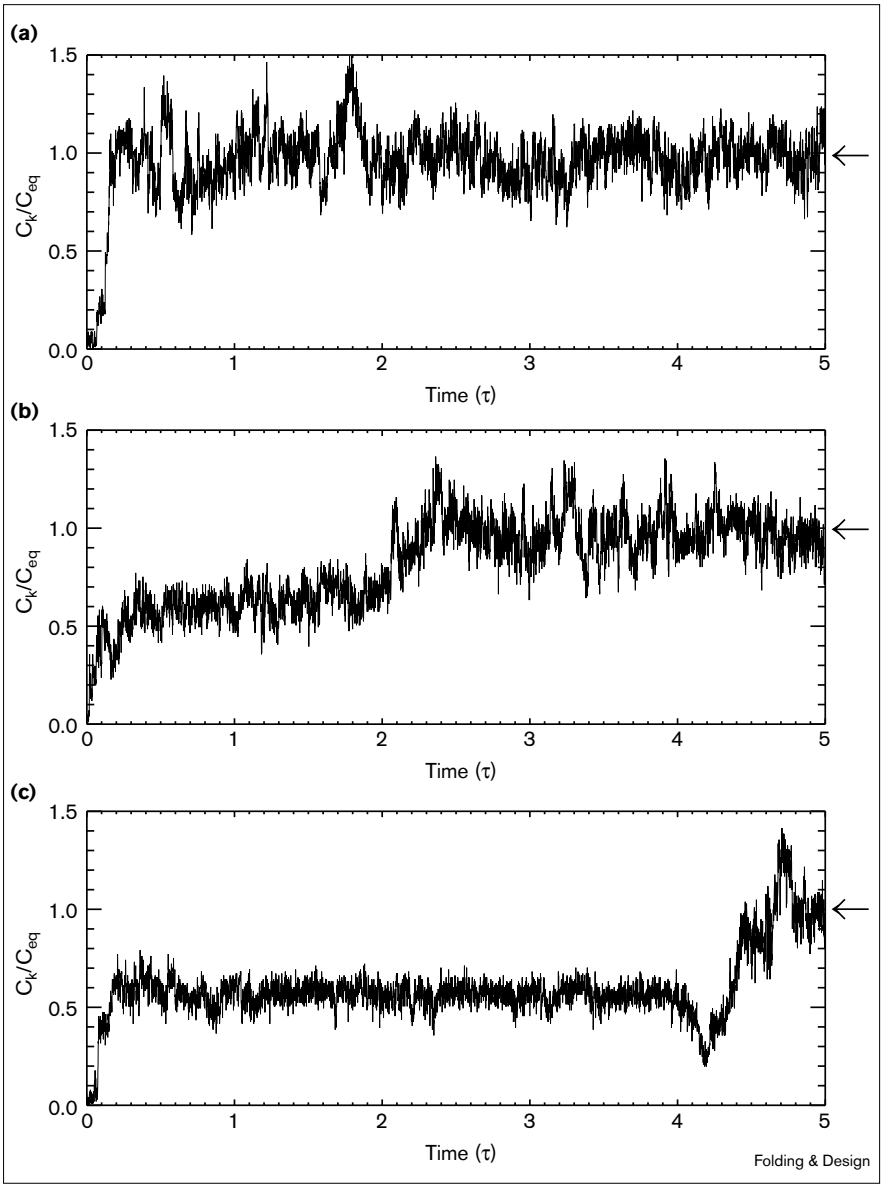
For convenience the clusters are relabeled so that the smaller the value of the cluster number the closer it is to the native conformation. The dynamical evolution of the structure is inferred by examining the number as well as the exact pairs of native-like contacts found in a cluster as a function of time. The cumulative history of this unambiguously gives the dynamics of formation for the specific nonbonded native contacts, and hence potential nucleation region(s). This analysis is repeated for an ensemble of trajectories that reaches the native conformation rapidly without forming any detectable intermediates.

Native conformation nucleation-collapse mechanism

General comments

Microscopic analysis of the dynamics of individual trajectories reveals that the ensemble of denatured molecules can be grouped into two distinct classes [4]. A fraction of initial molecules reaches the native state via a nucleation-collapse mechanism [3,4,6–8,11,25,26], whereas the remainder follows a three-stage multipathway mechanism to reach the native conformation. The dynamical behavior of the two distinct classes is shown in Figure 1, in which the overlap function, which measures the number of contacts in a given conformation that is common to the native state, is plotted as a function of time. Figure 1a shows the typical behavior for nucleation trajectories. This panel shows that once a (critical) number of tertiary contacts are formed (time $t \approx 100\tau$) the chain rapidly reaches the native conformation. In contrast, the behavior of trajectories that lead to the formation of misfolded structures are shown in Figure 1a,b. These two panels show that on a very short time scale ($t \approx 100\tau$) the chain gets trapped in one of the several misfolded structures

Figure 1



The time dependence of C_k/C_{eq} for three trajectories for a minimal model of a protein with M (the number of connected beads) = 46 and a low energy β-barrel structure. Here, C_k measures the number of contacts at time t (in units of τ) that are the same as the contacts in the native conformation. C_{eq} is the equilibrium value of contacts. The definition of C_k is given in Equation 9 of [4]. **(a)** A typical nucleation trajectory showing that after a small number of contacts are established the native state is reached. **(b,c)** Trajectories that rapidly get trapped in misfolded conformations where they stay for a long time. The nature (as measured by C_k) and the dwell time of these misfolded structures vary considerably, reflecting the many possible low-energy states that participate in guiding the folding kinetics. The arrows on the right-hand side indicate the equilibrium values of C_k.

characterized by having the value of overlap function different from the native state. The chain remains trapped as a misfolded conformation for rather long times until fluctuations drive the chain to make a transition to the native state. This process typically involves overcoming an activation barrier separating the native conformation and a multitude of misfolded structures.

What constitutes the critical nucleus?

The statistical clustering technique yields patterns from which the nucleus can be identified. But what constitutes a critical nucleus may still remain a subtle issue. In order to appreciate this it is useful to recount simulations done to identify the characteristics of the critical nucleus in the crystallization of a supercooled atomic liquid. Two important studies [27,28] lucidly point to the difficulties in unambiguously identifying the nature of the nucleation process in the more familiar problem of the crystallization of atomic liquids. These studies showed a marked consistency with the classical homogeneous nucleation theory of crystallization [19], but remained ambiguous in being able to predict in precise terms the characteristics of the critical nucleus. From a purely theoretical point of view, the basic question that is being posed is the following: given a trajectory (generated using some appropriate dynamics) how can the existence (or lack thereof) of a critical nucleus be determined? The identification and the nature of the nucleus (and hopefully the associated

barrier to its formation) should be obtained using only the information in trajectories (given in terms of an appropriate time series) without any additional input.

We follow the earlier treatments [27,28] of crystallization in supercooled liquids to define the critical nucleus. Nucleation theory [4,29] for proteins makes it clear that over some time interval one should observe clusters (constituting some well-defined tertiary contacts) of varying sizes. Some of these will be smaller than the critical size and others larger. The larger ones (with size greater than a critical size) will inevitably (but not always) be committed to folding and hence will be post-critical nucleus. The smaller ones will either shrink or continue to grow with time. We are not concerned with post-critical nucleus as they almost always reach the native conformation rapidly. Furthermore, the post-critical nucleus does not correspond to the saddle point in the free energy (i.e. it is not the transition state). Here, we use the following two criterion to define a nucleus. First, the size of the critical nucleus is taken to be the minimum number of stable tertiary contacts that persists until the native conformation for a given nucleating trajectory is reached for the first time. Second, the tertiary contacts in this nucleus are taken to be stable; in other words, once they are formed they lead to the formation of the native conformation relatively rapidly. By relatively rapidly we mean that the native conformation be reached within $\delta\tau_{lk}$, where τ_{lk} is the first passage time for that trajectory and we take $\delta = 0.2$. (The effects of choosing different values of δ are discussed in the following section.) The second criterion is the 'kinetic criterion' for identifying the critical nucleus. This criterion does not distinguish between the post-critical nucleus and the one at roughly the saddle point in the free energy in the appropriate meanfield description of the nucleation process. The first criterion, however, identifies the smallest set of contacts in the nucleus and hence we hope that this corresponds to the transition state. The stability and the kinetic criterion assure that once these contacts are formed the native conformation is reached rapidly. Notice that even after the critical nucleus is formed there is a probability that it will disintegrate into a small-size nucleus and not reach the native conformation. This, of course, is the hallmark of the classical nucleation mechanism.

The criteria we have used to identify a critical nucleus are similar to the ones adopted in the earlier studies of crystallization of simple liquids. From the technical point of view, our definition of a critical nucleus is nearly the same as the ancestor convention used to identify the nucleating droplet in undercooled liquids [27,28]. In the protein folding problem there is no need to differentiate between the single and multiple ancestor convention because every position along the chain is, in a certain sense, relatively unique in terms of the way in which it participates in determining the overall topology.

Dynamics of tertiary contacts

In order to apply the neural-net algorithm described in the Methods section, it is necessary to define a conformation (i.e. specify the vectors x_j). Because the formation of the hydrophobic core drives the chain to the native conformation (under conditions favorable for folding) we chose the elements of x_j to be the distance between certain pairs of B beads that are in contact in the native conformation. The β-barrel-like [4,30] native state has 107 pairs of B beads that are in contact (separated by $< 2\sigma$); this number is far too large for any meaningful analysis. In order to demonstrate that multiple (> 2) nucleation contacts lead the molecule to a native conformation, a many-point correlation function needs to be determined. This is neither practical nor informative. The algorithm used here reduces the problem of identifying nucleation contacts to one of 'pattern recognition' from which the delocalized nature of the critical nucleus can be immediately inferred independent of the definition used to identify the critical nucleus. We chose a subset from the 107 pairs of B–B beads. The sequence used in our simulation together with the location of the three turns in the β-barrel conformation is given in Figure 2. With reference to this figure, the following subset of B–B beads (that are in contact in the native state) are chosen to specify the conformation (i.e. the vector X_j): all B–B pairs between strands 1 and 2 (turn 1); all B–B pairs between strands 2 and 3 (turn 2); all B–B pairs between strands 3 and 4 (turn 3); and all B–B pairs between strands 1 and 4.

The subset has 49 B–B pairs, all of which are in contact in the native conformation. If these are formed then the native state results. This set of 49 B–B pairs is more than one needs to uniquely specify the native state. The distances between these B–B pairs are used as the elements of vector X in the statistical clustering algorithm.

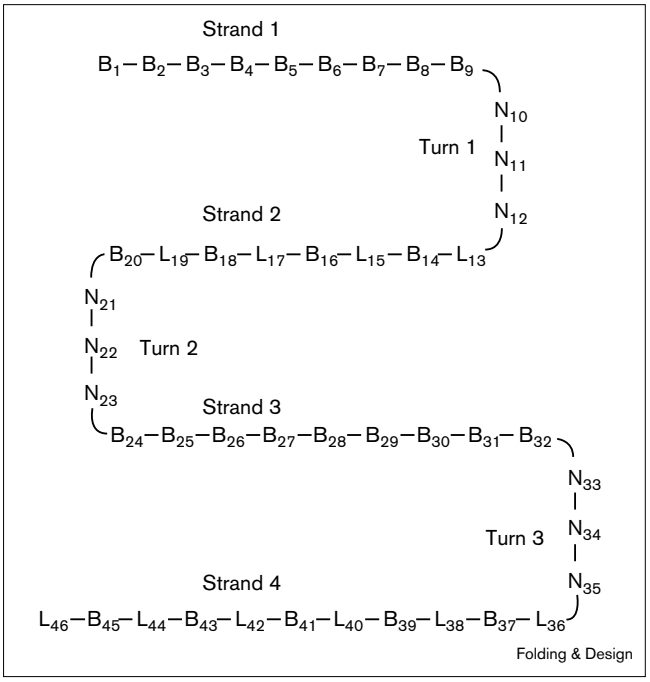
The B beads involved in the 49 pairs are (listed in increasing order of pair number): (6,14), (7,14), (8,14), (4,16), (5,16), (6,16), (2,18), (3,18), (4,18), (1,20), (2,20), (14,29), (14,30), (14,31), (16,27), (16,28), (16,29), (18,25), (18,26), (18,27), (20,24), (20,25), (20,26), (7,37), (8,37), (9,37), (5,39), (6,30), (7,39), (3,41), (4,41), (5,41), (1,43), (2,43), (3,43), (1,45), (30,37), (31,37), (32,37), (28,39), (29,39), (30,39), (26,41), (27,41), (28,41), (24,43), (25,43), (26,43), and (24,45). The ordering of the B beads in the sequence is shown in Figure 2.

Nucleation trajectories

Explanation of the patterns

By looking at the dynamics of formation of nonbonded contacts (and other correlation functions) as a function of time the trajectories that result in the rapid formation of the native state are identified. The typical time dependence of the acquisition of the native conformation for nucleation trajectories is shown in Figure 1a. For each

Figure 2



A schematic linear sequence of beads (α-carbon atoms of each residue) and their locations in our model. The turns and strands involved in the structure are also indicated. The contacts used in the self-organized neural-net analysis are discussed in the text. The code used identifies whether a bead is hydrophobic (B), hydrophilic (L) or neutral (N).

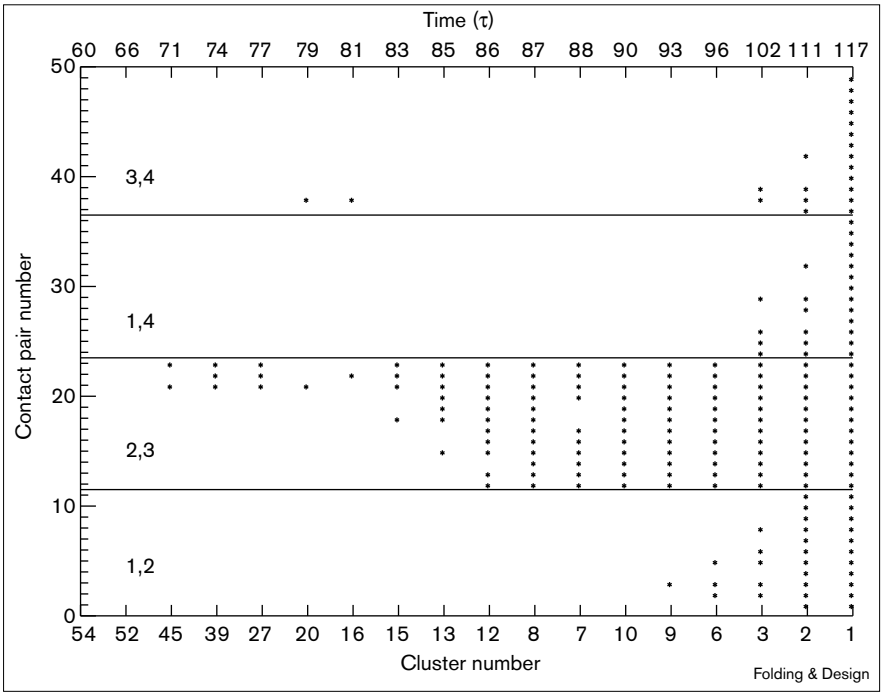
nucleation trajectory we have calculated the patterns of formation of contacts between the B beads used in the statistical clustering algorithms. The result of such an analysis for every trajectory results in contact maps of the sort shown in Figures 3–6. The clusters that arise for each trajectory are rearranged according to the increasing value of r_k (see Equation 4). If the difference in a specific pair of hydrophobic residues between a cluster center and the corresponding pair in the native state is within a tolerance value then we consider that the B–B pair to be a native contact. In other words, if:

$$|C_{jk} - X_{j0}| \leq \varepsilon \qquad (5)$$

where C_{jk} is the j^{th} element in the k^{th} cluster and X_{j0} is the corresponding element in the native conformation, then the pair representing the j^{th} element is considered to be a native contact. In this study we chose $\varepsilon = 1.25$. We found that qualitatively similar results are obtained for $\varepsilon = 1.25$–2.05.

Figures 3–6 show a map of native B–B contacts as a function of time and cluster number. The cluster number is arranged according to proximity to the native conformation so that smaller (larger) values correspond to closer (farther) distances relative to the native state. The lower x axis denotes the cluster number and the upper x axis describes the time a given cluster appears. The y axis identifies the number of the native B–B pairs. An asterisk at position (k,j) denotes that the j^{th} B–B pair in the k^{th}

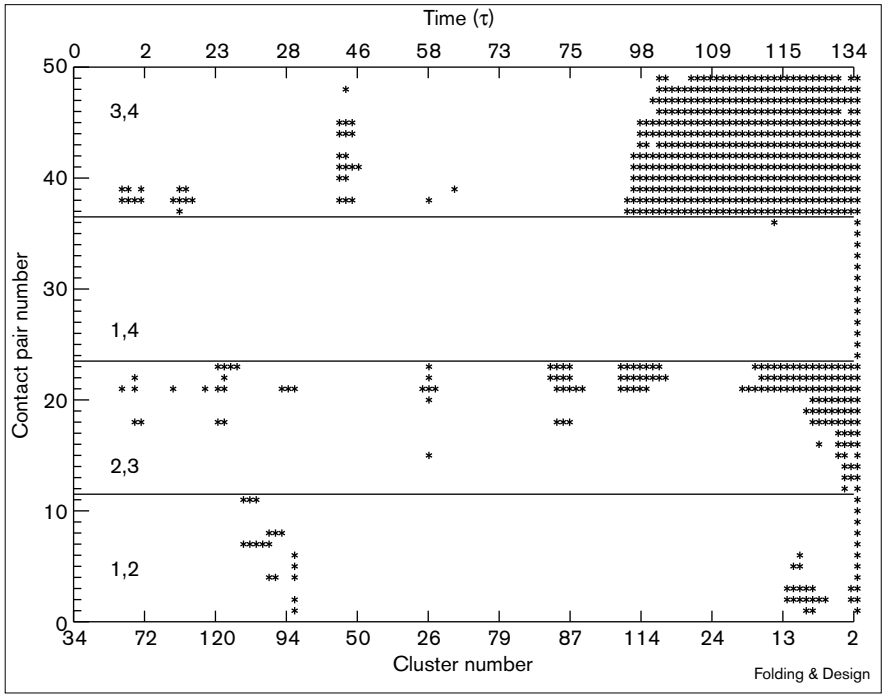
Figure 3



Dynamics of contact formation for one particular nucleation trajectory (i.e. one member of the initially denatured ensemble that reaches the native conformation without getting trapped in any detectable intermediate states). The numbers on the top axis show time t (in units of τ); the bottom axis corresponds to the cluster number, which measures how close the clusters are to the native state; and the vertical axis indicates the contact pair number. An asterisk indicates that a particular native contact belonging to a given cluster number appears at a specified time. For example, the contact numbers 21 and 23 belonging to cluster number 45 appear at $t = 71\tau$. Because the cluster number is quite large, the average conformation is very different from the native state implying that most of the molecule is in the disordered state. Notice that after the formation of the critical nucleus (identified as occurring at $t \approx 93\tau$), the native state is reached rapidly, as indicated by the monotonic decrease in the cluster number, implying increasing similarity with the native conformation.

Figure 4

Dynamics of contact formation as shown in Figure 3 except that it is for a different nucleation trajectory corresponding to a distinct initial condition. In this instance, the critical nucleus is very different from that shown in Figure 3, as can be seen by comparing the patterns of native contact formation. After the nucleus is formed, the cluster number keeps decreasing until the native state is reached. The critical nucleus is located largely in turn 3, stabilized by a few longer range contacts.



cluster is a native contact and appears at a time given by the upper x axis. In this neural-net type method a given cluster could have many conformations at different times. If this is the case then this cluster would appear in the lower x axis at these different times. Because the cluster number is a measure of the similarity to the native state, the order in which the numbers appear in the lower x axis is a marker of the progress (or a collative reaction coordinate)

Figure 5

Dynamics of native-contact formation for a nucleation trajectory, showing that the critical nucleus in this instance involves the beads in turn 1 as well as some in turn 3. This trajectory indicates that the number of beads in the nucleus is larger than the average number, suggesting a distribution of average number as well as a similarity to the native conformation. The probability of occurrence of these highly ordered nuclei is quite small, however. Nevertheless, this example shows a rather broad distribution of similarity from the transition state ensemble to the native conformation.

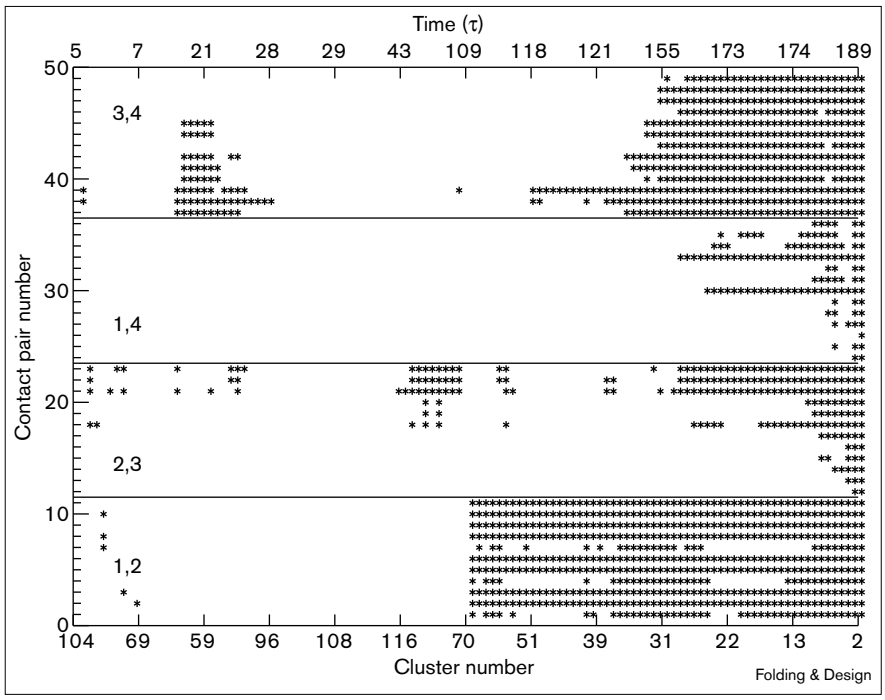
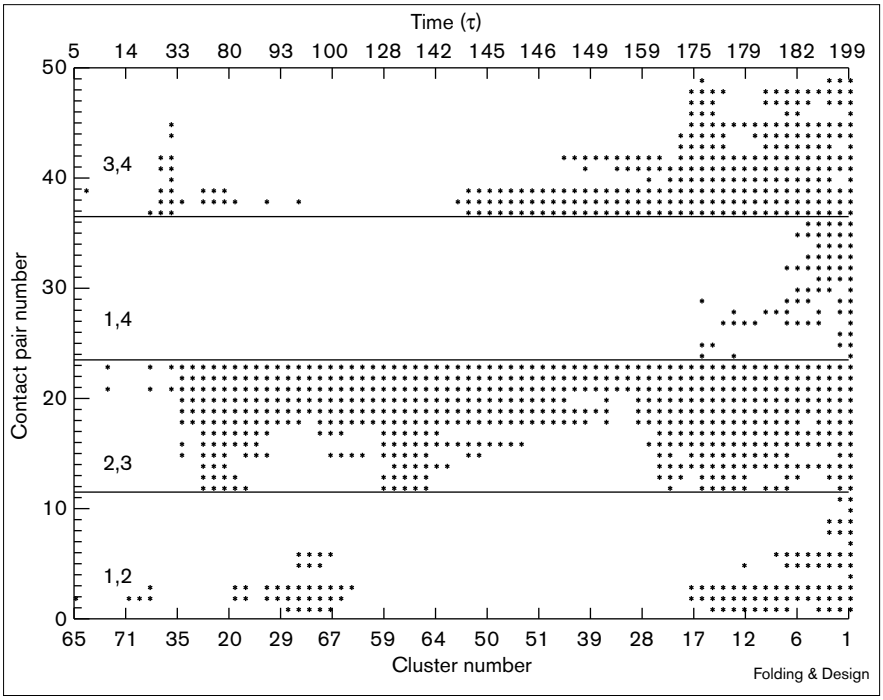


Figure 6



This figure is another example of a nucleation trajectory in which the critical nucleus is nearly the same as for the one shown in Figure 3. Both trajectories involve the almost complete formation of the central turn, but the patterns are quite different and the search time for the critical nucleus is also quite distinct. The distribution of nucleation times can, therefore, be varied even if the nature of the critical nucleus is nearly the same.

towards the native conformation. We expect that at early times there are multiple routes to the native state. As the similarity to the native conformation increases, the number of pathways is drastically reduced [31,32]. This implies that at short times the cluster numbers can appear in any order, whereas after the formation of the critical nucleus we expect that the cluster numbers should decrease nearly monotonically until the native state (cluster number = 1) is reached. This is, in fact, borne out in our computations as shown in Figures 3–6.

It is instructive to discuss these figures in some detail. The pattern of contact formation for one of the nucleation trajectories is shown in Figure 3. This figure shows that contacts between B–B pairs in the central turn region (between strands 2 and 3) form first. At $t = 71\tau$, pairs 21 (beads 18 and 27) and 23 (bead 20 and 26) form native contacts. This is followed by the formation of tertiary contacts between beads 20 and 25, labeled as pair 22 at $t = 74\tau$. Even though these are native contacts they do not last and the contact between some of the beads is broken at subsequent times. At $t = 81\tau$, pair 22 forms a native contact again, which is followed by the establishment of contacts for other pairs, and at $t = 90\tau$, every B–B pair in the central region is in the native conformation. At $t = 93\tau$, native contacts between B beads in strands 1 and 2 (turn 1) begin to be established with the formation of pair 3 (beads 14 and 18). The native contacts between B beads located in strands 3 and 4 (turn 3) are initiated at $t = 102\tau$, involving pairs 38 (beads 31 and 37) and 39 (beads 32 and 37). Notice that these pairs, as well as the one that is initiated at $t = 93\tau$ in turn 1, involve B beads near the turn position of the two strands. This ensures that the loops connecting the strands are in the overall correct position so that the topology of the post-critical nucleus formation is almost native like. For this trajectory, at $t = 117\tau$ every B–B pair is in the native contact thereby implying that the chain is in the native state. Thus, for this trajectory the first passage time is 117τ.

Characteristics of the critical nucleus: the folding nucleus is not unique

The contact maps generated by the neural net algorithm can be used to identify certain characteristics of the nucleus. For example, an estimate of the size range of the critical nucleus can be made. From the pattern in Figure 3 we note that the size of the contacts that are established prior to $t \approx 80\tau$ do not grow. The number of pairs of such contacts is in the range 1–3, involving 2–5 beads. Thus, we can conclude that the size of the critical nucleus should exceed 5. On the other hand, contacts established after ~85τ continue to grow until the native conformation is reached at $t \approx 117\tau$. Thus, we would assess that the critical contacts formed after $\tau \approx 95\tau$ would most definitely correspond to a post-critical nucleus. The search for the critical nucleus is therefore confined to identifying the smallest number of stable contacts (stability criterion) formed between $t \approx (90–95)\tau$ (kinetic criterion). In the present example, the identification of the critical nucleus is straightforward. Our criterion for the critical nucleus is

that once the set of contacts among the beads are formed the native conformation should be reached rapidly (within 20% of the first passage time for a given trajectory, which is the kinetic requirement) and with overwhelming probability. With these observations we notice that subsequent to the formation of all native B–B contacts in turn 2 (at $t = 90\tau$) the chain rapidly (by $t = 117\tau$) reaches the native state. Thus, the critical nucleus involves all the pairs in the central turn region (turn 2) and they correspond to 12 beads: 14, 16, 18, 20, 24, 25, 26, 27, 28, 29, 30, and 31. These beads form the inner hydrophobic core of the native conformation.

The estimate of the size of the critical nucleus made above and the associated beads that participate in it are somewhat dependent on the criterion used. If we had changed the kinetic criterion so that we insist that the native conformation be reached within $0.1\tau_{lk}$ then a somewhat larger size nucleus (containing 21 beads) would result. But in almost all of the nucleation trajectories we analyzed it was found that these larger nuclei grew with almost unit probability until the native state was reached. These were deemed to be post-critical nuclei and hence are not part of the ensemble of transition states. With the kinetic criterion that we have adopted we find that even the larger sized nuclei shrink with some degree of randomness (i.e. they do not lead to the native conformation) in accord with the classical picture of nucleation. Thus, we feel that our identification procedure gives us a picture of the nucleus whose characteristics are closer to the transition state. Nevertheless, this arbitrariness shows that studying the characteristics of the critical nucleus using information from simulations alone is subtle.

If we strictly implemented the ancestor convention [27,28] and identified the smallest stable cluster that reaches the native conformation without regard to the kinetic criterion we would identify the four contacts formed at $(81–83)\tau$ as the critical nucleus. These contain four highly localized beads that have all the native interactions fully satisfied. We classify these as plausible minimum-energy structures that are formed prior to reaching the transition region. Because the cluster number after forming this structure fluctuates strongly (see the lower x axis) we consider these structures to be early events in the search for the critical nucleus. The kinetic criterion excludes these structures as possible critical nucleation events (i.e. as transition states). Nevertheless, these considerations point to the subtle problems in identifying the size of the critical nucleus using simulations.

A natural question is whether, in all instances, the critical nucleus involves exactly the same beads. If this were the case then the contact maps generated by the self-organized neural net, would yield very similar looking patterns independent of which nucleation trajectory was examined. Furthermore, one would find that the rapid folding to the native state should always follow the formation of contacts between strands 2 and 3, and after the creation of turn 2. If all the nucleating trajectories lead to similar patterns, as shown in Figure 3, it would be logical to conclude that the critical nucleus is specific [4,6] and would imply that residues that trigger the nucleation mechanism under folding conditions are preassigned in the primary sequence. The algorithm used here unambiguously allows us to address the issue of a possible unique transition state regardless of the definition of the critical nucleus.

In order to address the issue of a specific critical nucleus, we have examined the dynamics of contact maps obtained using the neural net algorithm for several nucleation trajectories. Consider Figure 4, which shows the evolution of B–B contacts for another nucleation trajectory. The dynamic patterns of forming these contacts are completely different from the pattern shown in Figure 3. It is clear that prior to $t \approx 100\tau$, clusters of various sizes (having 1–7 contacts) form. But the size of this nucleus decreases at subsequent intervals, and hence does not satisfy the stability criterion. At $t \approx 100\tau$, native contacts 37 and 38 involving beads 30, 31, and 37 are established. These contacts persist until the first passage time is reached for this trajectory. This represents the minimum size of the critical nucleus. Such a small nucleus does not satisfy the kinetic criterion, however. If we continue to follow the evolution of the patterns we can identify the set of contacts formed at $t \approx 112\tau$ to satisfy both the stability and the kinetic criterion. This structure, which involves the set of contacts 21 and 37–49 (formed with 15 beads: 20, 24, 25, 26, 27, 28, 29, 30, 31, 32, 37, 39, 41, 43, and 45), gets to the native conformation extremely rapidly (in $< 0.2\tau_{lk}$). The critical nucleus that we have identified for this nucleation trajectory has fully formed contacts between strands 3 and 4, namely, those beads involved in turn 3. The transition state for the nucleation trajectory is topologically distinct from that for the trajectory analyzed in Figure 3. It is interesting to infer the topology corresponding to the identified critical nucleus. In addition to the fully formed turn 3 there is a native contact between beads 20 and 24, which is mediated by the loop connecting strands 2 and 3. This emphasizes the importance of loop flexibility and long-range contacts in initiating the nucleation mechanism. A comparison of Figures 3 and 4 also suggests that, at least for this model, the critical nucleus is not specific.

Another example of a nucleation trajectory in which the critical nucleus involves fully formed turns 1 and 3 is shown in Figure 5. Here, the nucleus is formed at $t \approx 180\tau$ and the native state is reached at $t \approx 200\tau$. The critical nucleus, the one that satisfies the stability and kinetic criteria, involves the pairs 1–11, 21, 22, 23, and 37–49. The 21 beads involved in the nucleus are 1–8, 14, 16, 18, 20, 24–32, 37, 39, 41, 43, and 45. This is one of the very few

examples of nucleation trajectories in which nearly half of the residues participate in the nucleus. The transition state corresponding to this trajectory is considerably more ordered than the others. The occurrence of such transition states, albeit with low probability, is consistent with the notion that various residues participate with differing probability creating an ensemble of such states in the folding process [33].

It is also interesting to analyze the dynamics of contact map formation for another nucleating trajectory shown in Figure 6. This figure shows that the critical nucleus involves turn 2, as does the one shown in Figure 3. The critical nuclei associated with these two nucleating trajectories are roughly similar. But a comparison of the two shows a dramatic difference in the pattern of development of B–B contact maps. For the trajectory shown in Figure 6, the critical nucleus is found much later (at $t \approx 167\tau$) than the one shown in Figure 3, where at $t \approx 90\tau$ the nucleus is found. In the present case the folding is complete at $t \approx 200\tau$. We have analyzed several other nucleation trajectories and find that even if any two trajectories show the same critical nucleus the search times are quite distinct. It is also clear from these figures and others that there is a distribution, albeit not so broad, of fast passage times for finding the native state after forming the critical nucleus.

The lessons from this exercise are the following:

1. The nucleating sites (those beads involved in the critical nucleus) depend on the starting condition, namely, the conformation in the denatured state of the molecule. This means that for an ensemble of molecules, a fraction of which reaches the native conformation via the nucleation mechanism, the precise critical nucleus depends on the initial conformation (the state of the denatured molecules) of the polypeptide chain. The critical nucleus is not preassigned in the primary sequence and is not specific. The critical nuclei are delocalized and the precise location depends on the conformation of the polypeptide chain in the delocalized state.

2. Regardless of the location of the nucleation sites it is clear that the earliest native contacts that subsequently form the critical nucleus are usually between residues near the turn location. This ensures that the critical nucleus has, in this model, nearly the same topology as the native state. This is consistent with the observation of Simmerling and Elber [5] who have shown that the critical nucleus can be viewed as an almost independent entity whose structure would stay intact roughly independent of the rest of the chain conformation. It does not follow, however, that if the residues of the critical nucleus are isolated from the rest of the chain a foldable subunit would result. Our studies and those of Simmerling and Elber [5] do point out that the critical nucleus is spatially localized by residues that are relatively close in sequence space, and the critical nucleus is also stabilized by a few long-range contacts.

3. The almost stochastic assembly of the critical nucleus and the subsequent formation of the native conformation occur at different times for different trajectories. This means that even among nucleation trajectories there is a distribution of first passage times. Thus, the search time for the critical nucleus is dependent on the conformation of the starting molecule, and hence is critically dependent on the ensemble of denatured states.

4. Even in the fast process in the folding of biomolecules, which is presumed to proceed by a suitable nucleation-collapse mechanism, the transition state is not unique [33,34]. There is an ensemble of transition states all with a topology resembling that of the native state and containing a substructure that is fully formed. In asserting this we have assumed that the transition state for a given trajectory coincides with the structure corresponding to the critical nucleus. There is considerable variation of the overlap between these structures and the native state.

5. Analysis of various nucleation trajectories suggests that there is a relatively broad distribution of size of the critical nuclei. This, of course, is consistent with a classical nucleation picture for protein folding [4]. Such a distribution translates into the structures of the transition state having a differing degree of similarity to the native conformation [33].

It should be clear from the preceding analysis of the patterns generated by the neural-net algorithm that there is some arbitrariness in identifying the critical nucleus. Nevertheless they show rather dramatically that, whatever the definition of the critical nucleus is, the patterns are distinct for the nucleation trajectories (Figures 3–6). We have found that the various nucleation trajectories generate different patterns of formation of the native contacts. Even when the critical nucleus is nearly the same (as is the case in Figures 3 and 6) the search time for the nucleus and other characteristics, such as the nature of collapse of the structure after forming the critical nucleus, are different.

Local and non-local contacts in the critical nuclei

Further insight into the structures of the critical nuclei may be obtained by examining the number of local and non-local contacts in the nucleus. If we assume that the critical nucleus is a mobile state of the native conformation we would expect that both local and non-local contacts would be present. Earlier studies have indicated that a considerable fraction of contacts in the nucleus should be local [2,35,36]. The importance of the optimal number of local contacts has also been addressed using lattice models [37,38]. For our purposes we consider a contact to

be local if a tertiary interaction is established between the beads separated by < 7 positions along the sequence. If we alter this criterion by small amounts the conclusions remain the same.

The number of local contacts in the various critical nuclei is found to involve 30–40% of the residues in the total sequence. For example, the critical nucleus in Figure 4 has five (out of a total of 14) local contacts while the critical nucleus in Figure 5 has four (out of a total of 14) local contacts. These numbers are typical for the nucleation trajectories that we have examined. The proportion of local contacts in our model is quite similar to that found in the studies of Abkevich *et al.* [3]. As noted by these authors, it is necessary for the critical nucleus to have a certain fraction of long-range contacts. These contacts stabilize the associated transition states and make the acquisition of the native conformation a rapid process after the critical nucleus is formed.

Off-pathway trajectories and misfolded structures

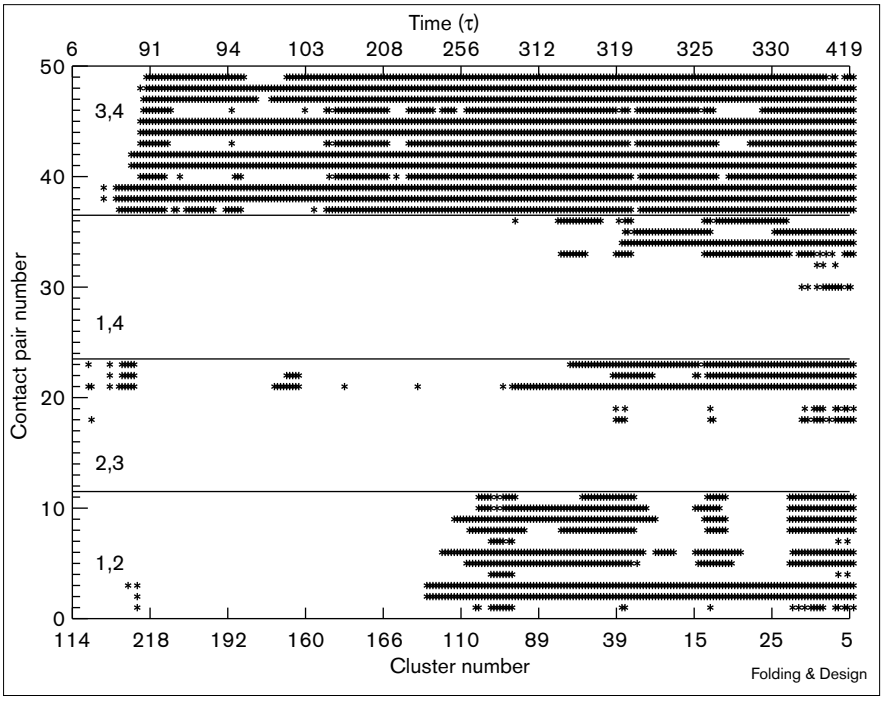
According to the kinetic partition mechanism [12,13], if the parameter $\sigma \equiv (T_\theta - T_f)/T_\theta$, where T_θ is the collapse temperature, is large then a substantial fraction of the initial population of molecules reaches the native conformation by indirect processes, which are conveniently described by a three-stage multipathway mechanism [14]. For the present simulation conditions there is a significant fraction of molecules (≈ 0.4; [4]) that does not reach the native state rapidly by a nucleation-collapse process. It is of interest to subject the trajectories corresponding to these off-pathway processes to an analysis similar to that for the nucleating trajectories. The contact maps generated by the neural net algorithm for two off-pathway trajectories are shown in Figures 7 and 8. The patterns of native B–B contacts seen in these figures are very different from those displayed in Figures 3–6. Even though there are several native contacts in these conformations they are spread throughout the various strands. There is no synchronous formation of contacts in any localized region of the chain. Even though in the off-pathway trajectories a certain number of native contacts are fully established (turn 3 in Figure 7), the native conformation is not reached rapidly. This is illustrated by Figure 7 in which one can identify a critical nucleus, which involves the contacts 37–49 leading to the formation of turn 3 between strands 1 and 4. But the rest of the molecule is in a highly disordered state, so in this case the chain gets trapped in one of the competing basins of attraction in which the conformation is misfolded. Thus, although an identifiable critical nucleus appears to form relatively early, the subsequent events are not favorable for the formation of the native state. It follows that the structure of the critical nucleus should have a mobile native-like topology for the nucleation-collapse process to be productive.

Relation to other theoretical models

Some of the characteristics of the critical nucleus have already been anticipated in an important study by Matheson and Scheraga [1] using a phenomenological model. By assuming that the major driving force for folding is the formation of the hydrophobic core these authors predicted the nucleation regions for several proteins by estimating the free energy of formation of the nucleation sites. Their major conclusions are: first, although there may be one major nucleation site for a given protein there are alternative pockets (a small number of them) whose stabilities are within a few k_BT of the dominant nucleation region. This crucial feature of nucleation regions is supported by the present study. Second, the nucleation sites for several proteins are found to be near the turn regions. This prediction is also supported by our studies, but more studies involving different classes of topologies are required to fully confirm the generality of this observation. Third, Matheson and Scheraga [1] also suggested that the size of the critical nucleus is small and typically consists of ~10–14 residues. This is in accord with the present findings and is consistent with our earlier theoretical estimates [4]. It was also argued that the dominant nucleating pocket is exclusively composed of residues that are contiguous in sequence space. Our simulation and that of Abkevich *et al.* [3] suggest that while the nuclei are predominantly composed of residues that are close together in sequence space, a certain number of them are far apart. In fact the presence of the residues that are distant in sequence space is important in assuring that the process of nucleation-collapse and the acquisition of the native conformation happen almost simultaneously [15]. The presence of a certain fraction of long-range (in sequence space) native contacts ensures that the structures of the transition state are somewhat expanded versions of the native state [7].

The sizes of the critical nuclei in single-domain proteins are predicted to be relatively small [1,4,39,40]. This is in contrast to estimates made by others [29] who suggest that the typical critical droplet size is 100–300 residues and constitutes the entire protein. They further suggest that the "rate limiting step in protein folding involves folding the entire protein, not just some critical nucleus that then nucleates the formation of the rest of the native structure" [29]. This suggestion, which seems to undermine the notion of critical nucleus altogether, is at variance with our present conclusion and earlier studies [1,4]. Furthermore, experiments on CI2 suggest that there is a well-defined critical nucleus that contains an α-helix (12 residues) while "the rest of the protein is in the process of collapsing around the nucleation site as the nucleation site itself is being found" [7]. The more recent study by Onuchic *et al.* [33], which suggests the presence of an ensemble of transition states in which each state exhibits a different similarity with the native conformation, is in accord with our conclusions. A reinterpretation of the

Figure 7



The dynamics of contact formation for a non-nucleation trajectory that leads to a correlation as shown in Figure 1a,b. This example shows patterns of native contacts quite similar to those in Figures 3–6, except that the chain gets trapped in a misfolded conformation for very long times. At $t \approx 419\tau$, there are many similarities between the misfolded structure and the native state, as evidenced by the numerous native contacts. But a large fraction of crucial native contacts are missing. In this case the chain stays misfolded for a long time and only subsequently reaches the native state by a partial unraveling. The presence of such competing basins of attraction slows down the overall folding process.

earlier theory [29] apparently leads to the conclusion that the mean size of the critical nuclei is small [41].

Abkevich *et al.* [3] have suggested that for certain sequences of lattice proteins the critical nucleus, which serves as the transition state, is specific and their formation is a "necessary and sufficient condition" for folding. The analysis presented here, however, suggests that the critical nucleus is not unique and there are several regions in which nucleation sites can exist. Despite the major difference it is worth stating that both studies have shown that the size of the critical nuclei is small (involving perhaps 10–20 residues for single-domain proteins). In addition, the contacts in the nuclei have both local and non-local character and the proportion of each depends on the sequence and external conditions.

The difference in the major conclusion between our studies and those of Abkevich *et al.* [3] may be due to the following reasons. First, in the lattice studies [3,6] the post-critical nucleus was sought and hence it is reasonable that the structure of such an object would contain more order and specificity than structures that are found close to the transition state. Second, because the sorts of topologies found in our model are different from the lattice models, the nature of nucleation events may also be quite distinct. Third, the difference may also arise due to the differences in the potentials and the dynamics used (Langevin versus Monte Carlo). Fourth, the specific critical nucleus identified by Abkevich *et al.* [3] consists exclusively of charged residues, which are seldom formed in the core of proteins. It may be that the proximity of charges in their model initiates the formation of the native structure. These discussions point to the need for further studies to fully understand the nature of the critical nucleus, the formation of which greatly enhances the rate of folding of proteins.

Discussion

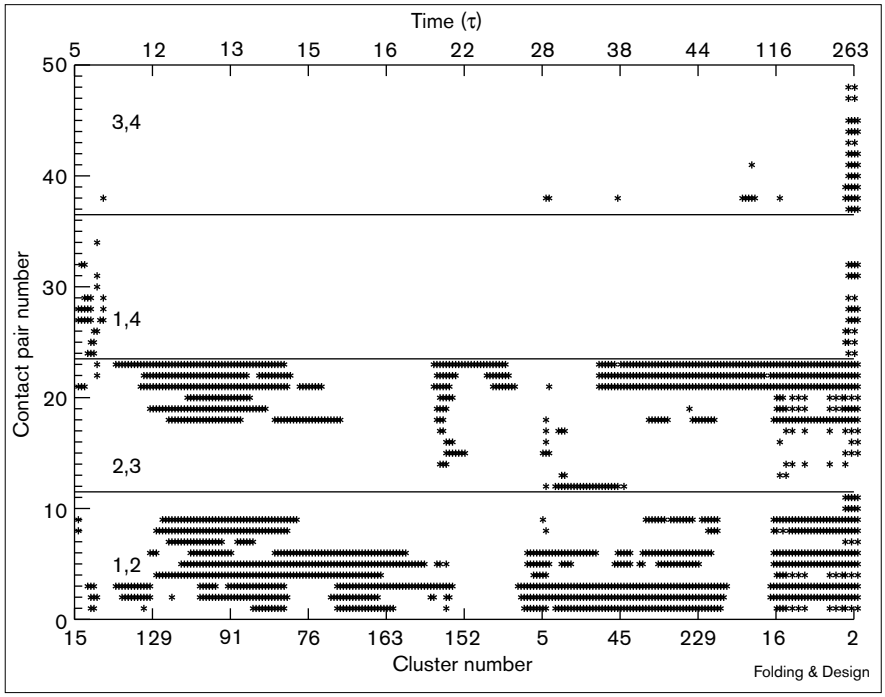
There have been a few recent experimental studies that have argued in favor of the nucleation-collapse mechanism as an efficient way of reaching the folded state. One of them is a comprehensive study of folding of CI2 using protein engineering methods by Itzhaki *et al.* [8]. These authors use the phrase nucleation condensation instead of nucleation collapse. The other study has been on the refolding of cytochrome *c* by Sosnick *et al.* [26] who have used the terminology of molecular collapse to describe the nucleation-collapse mechanism. Because the experimental protocols used in the studies of CI2 and cytochrome *c* are different we discuss each experiment separately.

Chymotrypsin inhibitor 2

CI2 is a 64 residue single-domain protein that apparently folds in a kinetically two-state manner under a range of external conditions [8]. Itzhaki *et al.* [8] have used a protein engineering method to study the folding of CI2 and 100 mutants (all of which apparently also follow two-state kinetics; A.R. Fersht, personal communication) to decipher the structure of the transition state for folding. The major conclusions of these authors are: first, the

Figure 8

This figure is the same as Figure 7 except that it is for another non-nucleation trajectory. In this case, the chain is also trapped in a misfolded conformation, but this misfolded structure is more ordered than the one shown in Figure 7. Even at t ≈ 263τ there are 33 native contacts, whereas at t ≈ 419τ in Figure 7 there are only 31 native contacts. In addition, the cluster number in this figure at the longest time is smaller than that in Figure 7 at t ≈ 419τ, which also shows a greater similarity to the native state.



wealth of data are apparently consistent with a nucleation-collapse mechanism. It is argued that the nucleus "consists primarily of adjacent residues" [8]. For the case of CI2 they suggest that the nucleus is essentially the residues (12–24) that comprise the three-turn α helix. Second, in the transition state, which may be thought of as being the saddle point of the free-energy profile for protein folding, the nucleus is stabilized by weak longer range interactions with other parts of the molecule. Because folding to the native state, following the formation of the nucleus, in conjunction with the overall collapse of the molecule is rapid, Itzhaki *et al.* [8] conclude that the acquisition of the native structure (secondary and tertiary interactions) and the nucleation collapse occur almost simultaneously. Third, from these two observations it follows that in the transition state the structure of CI2 (under the conditions of the experiments) is an expanded state of the native conformations with the crucial condition that the overall topology of the transition state is roughly the same as that of the native conformation.

The present calculations do show that the critical nucleus is indeed composed of several residues (or beads) that are close to each other, which is in accord with the first conclusion stated above. But we find that the regions in which nucleation starts can be in distinct sites, reflecting the heterogeneity of the ensemble of initial conditions. Our findings also suggest that there are a relatively small number of nucleation regions that can be identified as the transition region of the unfolding to folding reaction. The difference found between the theoretical findings and the experimental conclusions may be a result of the intrinsic average performed in the experimental case, and hence only the most stable nucleus is identified. In a recent paper Onuchic *et al.* [33] argue that the experimental data on CI2 suggest that the transition state is not unique, in accord with our findings.

The second major conclusion of Itzhaki *et al.* [8] is completely consistent with our earlier theoretical prediction [15] that in pathways dominated by a nucleation mechanism, the nucleation-collapse process and the acquisition of the native conformation are almost synchronous. In our earlier theoretical work [4] we argued that this process involves long-range contacts between residues that are distant in sequence space. In this study, we find that although the nucleation regions are composed of many contiguous beads there are non-local contacts that impart stability to the critical nucleus. It could in fact be argued that a certain fraction of such long-range contacts are required for triggering the nucleation-collapse process. This observation is also consistent with the studies of Abkevich *et al.* [3].

We find that the critical nucleus is indeed relatively small, containing ~15–18 beads. Shortly after the formation of the critical nucleus the rest of the chain comes into a compact state. Thus, as suggested earlier [4] and described here in more detail, the nucleus is a mobile entity with native-like topology, which once formed gets

to the native conformation (with overwhelming probability) by appropriate arrangements involving the collapse of the polypeptide chain. These findings appear to be in accord with the suggestion of Itzhaki *et al.* [8], in that for CI2 the transition state involving the critical nucleus is similar to the native conformation except in the degree of compactness.

Cytochrome *c*

Sosnick and others [10,25] have found that cytochrome *c* refolds to the native conformation in an almost kinetically two-state manner at 10°C and 0.7 M guanidinium chloride. In general, their experiments support the kinetic partitioning mechanism [12] in which the off-pathway processes represent reconfiguration of the misfolded structures. In a more recent study [26], Sosnick *et al.* have further examined the nature of the fast process (occurring in 15 ms under the conditions of experiments) in cytochrome *c* using a combination of stop-flow spectroscopy and hydrogen-exchange labeling experiments. Following our earlier predictions [4], Sosnick *et al.* [26] interpret the fast process as evidence for a nucleation-collapse mechanism. They call this process molecular collapse to correctly imply that the formation of the critical nucleus in conjunction with the collapse process leads to a structure that is mobile enough (but still contains a native-like topology) so that the native state can be reached quickly. We had suggested in our previous studies [4,11] (and have further elaborated here) that although the nucleus is essentially formed as a spatially localized structure the acquisition of the native conformation involves a nucleation-collapse process in which the collapse and nucleation are coupled in such a way as to lead to native-like transition structures. Such structures reach the native conformation rapidly. The recent interpretation of the fast process by Sosnick *et al.* [26] is in accord with the theoretical ideas.

Conclusions

The analysis presented here suggests that the fluctuating nuclei are small and occur predominantly (but not exclusively) near the loop regions. The small sizes of the folding nuclei can be understood in simple terms. Because the nuclei form relatively early in the folding process we assume that the major driving force is the hydrophobic interaction resulting in the burial of hydrophobic residues. An estimate for the free energy of formation of a compact structure containing N_R residues can be written as ([4]; neglecting logarithmic corrections arising from entropy loss due to loop formation):

$$\Delta F\left(N_R\right) \approx -\frac{\varepsilon_D}{2} f_H^2 N_R^2 + 4\,\pi\gamma a^2 N_R^{2/3} \qquad (6)$$

where ε_D is the driving interaction to the formation of an ordered structure, f_H is the fraction of hydrophobic residues, γ is the average surface tension, and a is the mean size of the residues. The optimal or average size of the critical nucleus is:

$$N_R^* = (8\pi\gamma a^2 / 3\, f_H^2\, \varepsilon_D)^{3/4} \qquad (7)$$

If we use the standard values 25–75 cal/Å^2-mol, 1–2 kcal/mol, and $f_H \approx 0.55$ for γ, ε_D, and f_H, respectively, we obtain N_R^* in the range of 15–22 residues. This estimate is in accord with the present study and earlier works [1,2]. The width of the transition region can be estimated from the curvature of free energy, κ, and is given by:

$$\frac{\kappa}{k_B T} = \left| \frac{2}{3} \frac{\varepsilon_H f_H^2}{k_B T} \right| \qquad (8)$$

A typical estimate for this is in the range 0.3–0.6 using the standard values of ε_H and f_H. These small values imply a very broad transition region for folding. All these estimates should be viewed as tentative because we have ignored polymeric effects due to chain connectivity and surface fluctuations of the folding nucleus, which are significant in altering the thermodynamic and the kinetic processes of nucleation.

Acknowledgements

D.T. is grateful to R.D. Mountain and D.K. Klimov for illuminating discussions. We are grateful to A.R. Fersht and P.G. Wolynes for communicating their recent results prior to publication. This work was supported in part by a grant from the National Science Foundation through grant number NSF CHE96-29845.

References

1. Matheson, R.R. & Scheraga, H.A. (1978). A method for predicting nucleation sites for protein folding based on hydrophobic contacts. *Macromolecules* **11**, 814-829.
2. Wetlaufer, D. (1973). Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc. Natl Acad. Sci. USA* **70**, 697-701.
3. Abkevich, V.I., Gutin, A.M. & Shakhnovich, E.I. (1994). Specific nucleus as the transition state for protein folding. Evidence from the lattice model. *Biochemistry* **2**, 10026-10036.
4. Guo, Z. & Thirumalai, D. (1995). Kinetics of protein folding: nucleation mechanism, time scales, and pathways. *Biopolymers* **36**, 83-103.
5. Simmerling, C. & Elber, R. (1994). Hydrophobic 'collapse' in a cyclic hexapeptide: computer simulations of CHDLFC and CAAAAC in water. *J. Am. Chem. Soc.* **116**, 2534-2547.
6. Shakhnovich, E., Abkevich, V. & Ptitsyn, O. (1996). Conserved residue and the mechanisms of protein folding. *Nature* **379**, 96-98.
7. Fersht, A.R. (1995). Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. *Proc. Natl Acad. Sci. USA* **92**, 10869-10873.
8. Itzhaki, L.S., Otzen, D.E. & Fersht, A.R. (1995). The structure of the transition state for folding of chymotrypsin inhibitor 2 analyzed protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.* **254**, 260-288.
9. Kiefhaber, T., Bachmann, A., Wildegger, G. & Wegner, C. (1997). Direct measurement of nucleation and growth rates in lysozyme folding. *Biochemistry*, in press.
10. Sosnick, T.R., Mayne, L., Hiller, R. & Englander, S.W. (1994). The barriers in protein folding. *Nat. Struct. Biol.* **1**, 149-156.
11. Thirumalai, D & Guo, Z. (1995). Nucleation mechanism and theoretical predictions for hydrogen-exchange labelling experiments. *Biopolymers* **35**, 137-140.
12. Thirumalai, D. & Woodson, S.A. (1996). Kinetics of folding of proteins and RNA. *Acc. Chem. Res.* **29**, 433-439.
13. Thirumalai, D. (1994). In *Theoretical Perspectives on In Vitro and In Vivo Protein Folding in Statistical Mechanics, Protein Structure, and Protein Substrate Interactions.* (Doniach, S., ed.) pp. 115-134, Plenum Press, New York.

14. Camacho, C.J. & Thirumalai, D. (1993). Kinetics and thermodynamics of folding in model proteins. *Proc. Natl Acad. Sci. USA* **90**, 6369-6372.
15. Thirumalai, D. (1995). From minimal models to real proteins: time scales for protein folding kinetics. *J. Physique* I **5**, 1457-1467.
16. Tsong, T.Y., Baldwin, R.L. & McPhee, P. (1972). A sequential model of nucleation-dependent protein folding: kinetic studies of ribonuclease A. *J. Mol. Biol.* **63**, 453-457.
17. Fukugita, M., Lancaster, D. & Mitchard, N.G. (1993). Kinematics and thermodynamics of a folding heteropolymer. *Proc. Natl Acad. Sci. USA* **90**, 6365-6368.
18. Guo, Z. & Brooks, C.L., III (1997). Thermodynamics of protein folding: a statistical mechanical study of a small all β protein. *Biopolymers,* in press.
19. Oxtoby, D.W. (1988). Nucleation of crystals from the melt. *Adv. Chem. Phys.* **70**, 263-296.
20. Carpenter, G.A. & Grossberg, S. (1987). ART 2: self-organization of stable category recognition codes for analog input patterns. *Appl. Optics* **26**, 4919-4930.
21. Pao, Y.-H. (1989). *Adaptive Pattern Recognition and Neural Networks.* Addison-Welsley, New York.
22. Karpen, M.E., Tobins, D.J. & Brooks, C.L., III. (1993). Statistical clustering techniques for the analysis of long molecular dynamics trajectories: analysis of 2-2 in trajectories of YPGDY. *Biochemistry* **32**, 412-420.
23. Honeycutt, J.D. & Thirumalai, D. (1992). The nature of folded states of globular proteins. *Biopolymers* **32**, 695-709.
24. Oldfield, T.J. & Hubbard, R.E. (1994). Analysis of C_α geometry in protein structure. *Proteins* **18**, 324-337.
25. Daggett, V., Li, A., Itzhaki, L.S., Otzen, D.E. & Fersht. A.R. (1996). Structure of the transition state for folding of a protein derived from experiment and simulation. *J. Mol. Biol.* **257**, 430-440.
26. Sosnick, T.R., Mayne, L. & Englander, S.W. (1996). Molecular collapse: the rate-limiting step in two-state cytochrome c folding. *Proteins* **24**, 413-426.
27. Swope, W.C. & Andersen, H.C. (1990). 10^6-particle molecular-dynamics study of homogeneous nucleation of crystals in a supercooled atomic liquid. *Phys. Rev. B* **41**, 7042-7053.
28. Yang, J., Gould, H., Klein, W. & Mountain, R.D. (1990). Molecular dynamics investigation of deeply quenched liquids. *J. Chem. Phys.* **93**, 711-723.
29. Bryngelson, J.D. & Wolynes, P.G. (1990). A simple statistical field theory of heteropolymer collapse with application to protein folding. *Biopolymers* **30**, 177-188.
30. Honeycutt, J.D. & Thirumalai, D. (1990). The metastability of folded states of globular proteins. *Proc. Natl Acad. Sci. USA* **87**, 3526-3529.
31. Bryngelson, J.D., Onuchic, J.N., Socci, N.D. & Wolynes, P.G. (1995). Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* **21**, 167-195.
32. Camacho, C.J. & Thirumalai, D. (1995). Modeling the role of disulfide bonds in protein folding. *Proteins* **22**, 28-40.
33. Onuchic, J.N., Socci, N.D., Luthey-Schulten, Z. & Wolynes P.G. (1996). Protein folding funnels: the nature of the transition states ensemble. *Fold. Des.* **1**, 441-450.
34. Boczko, E.M. & Brooks, C.L. III. (1995). First principle calculation of the folding free energy of a three helix bundle protein. *Science* **269**, 393-396.
35. Rooman, M.J., Kocher, J-P. & Wodak, S.J. (1992). Extracting information on folding from the amino acid sequence: accurate predictions for protein regions with preferred conformation in the absence of tertiary interactions. *Biochemistry* **31**, 10226-10238.
36. Rooman, M.J. & Wodak, S.J. (1992). Extracting information on folding from the amino acid sequence: consensus regions with preferred conformation in homologous proteins. *Biochemistry* **31**, 10239-10249.
37. Gorindarajan, S. & Goldstein, R.A. (1995). Optimal local propensities for model proteins. *Proteins* **22**, 413-418.
38. Unger, R. & Moult, J. (1996). Local interactions dominate folding in a simple protein model. *J. Mol. Biol.* **259**, 988-994.
39. Moult, J. & Unger, R. (1991). An analysis of protein folding pathways. *Biochemistry* **30**, 3816-3824.
40. Dill, K.A., Fiebig, K.M. & Chan, H.S. (1993). Cooperativity in protein-folding kinetics. *Proc. Natl Acad. Sci. USA* **90**, 1942-1946.
41. Wolynes, P.G. (1997). Folding funnels and energy landscapes of larger proteins within the capillarity approximation. *Proc. Natl Acad. Sci. USA* **94**, 6170-6175.

Because *Folding & Design* operates a 'Continuous Publication System' for Research Papers, this paper has been published via the internet before being printed. The paper can be accessed from http://biomednet.com/cbiology/fad.htm – for further information, see the explanation on the contents pages.