Articles

An Analysis of Protein Folding Pathways†

John Moult*,‡ and Ron Unger‡,§,∥

Center for Advanced Research in Biotechnology, Maryland Biotechnology Institute, University of Maryland, 9600 Gudelsky Drive, Rockville, Maryland 20850, Department of Applied Mathematics and Computer Science, Weizmann Institute of Science, Rehovot 76100, Israel, and Institute for Advanced Computer Studies, University of Maryland, College Park, Maryland 20742

Received June 11, 1990; Revised Manuscript Received October 8, 1990

ABSTRACT: We have developed a model of the protein folding process based on three primary assumptions: that burying of hydrophobic area is the dominant contribution to the relative free energy of a conformation, that a record of the folding process is largely preserved in the final structure, and that the denatured state is a random coil. Detailed folding pathways are identified for 19 protein structures. The picture of the folding process that emerges from this analysis is one of nucleation by regions of 8–16 residues. Nucleation sites then lead to larger structures by two mechanisms: propagation and diffusion/collision. A Monte Carlo simulation is used to follow the folding pathway when propagation is the dominant mechanism. Because detailed pathways are derived for each protein, the models are susceptible to experimental verification.

Soluble globular proteins fold into their functional structure on a time scale of typically seconds and, at least for small monomeric proteins, without the aid of other biochemical machinery (Anfinsen, 1970). In principle it is therefore possible to deduce the three-dimensional structure from the amino acid sequence. The large number of conformations a protein polypeptide chain can adopt prevents a survey of all the possibilities and poses a problem in the natural folding process too: If a protein is to find its functional conformation by wandering randomly through conformational space, in excess of 10^{50} years would be required for folding (Levinthal, 1968). Proteins solve this problem, apparently, by utilizing some form of folding pathway encoded in the sequence.

Short segments of chain [up to about 20 residues long (Wetlaufer, 1973)] can search through all possible conformations in less than a second, hence the idea of nucleation or initiation sites for folding: short regions of chain that will tend to adopt their final conformation early in the folding process. Following nucleation, folding has generally been supposed to proceed by one of two mechanisms—diffusion of semistable nucleation sites until two or more happen to collide in the folded state, associating to form the next level of structure [the diffusion/collision model (Karplus & Weaver, 1976; Bashford et al., 1988)], or growth of structure out from the nucleation sites [propagation (Wetlaufer, 1973)]. Experimental evidence for molten globule denatured states and intermediates (Kunihiro, 1989) provide a different view of the folding process, not considered further here.

A variety of models based on these ideas have been proposed. The diffusion/collision process has been elegantly characterized (Karplus & Weaver, 1976) and used to suggest detailed pathways (Bashford et al., 1988). The tendency of neighboring β strands to be near in the sequence has been explained in terms of a propagation mechanism (Richardson, 1981). Local compactness (Lesk & Rose, 1981) and contact density (Montelione & Scheraga, 1989) have been used to define folding pathways. Lattice models in which the possible conformations are restricted by limiting the volume enclosing the chain (Covell & Jernigan, 1990) or by tuning interresidue interactions and turn propensities (Skolnick & Kolinski, 1989) have recently begun to provide valuable insight.

Here we present a detailed pathway model, together with the folding mechanisms it implies for a set of different proteins. The model rests on three primary assumptions: that burying hydrophobic area is the dominant contribution to the free energy stabilizing the folded state, that a record of the folding process is preserved in the final structure, and that the denatured state is a random coil. These assumptions have been used by others, and we do not justify them in detail here. Rather, we show that the observed structures determined by X-ray crystallography provide evidence to support the model, and we suggest further experimental tests.

Calculation of Free Energy Contributions to Folding. The quantitative contributions of different free energy terms to protein stability remains a controversial issue (Dill, 1990; Murphy et al., 1990). For the present model, we take the position that changes in enthalpy in going from the unfolded to the folded state are of relatively minor importance. That is, the energy of new intramolecular protein interactions is assumed to be approximately compensated for by lost protein–solvent ones (Roseman, 1988; Finkelstein & Shakhnovich, 1989). Contributions to the entropy of a conformational state are considered to arise from two main sources: the burying of nonpolar surface (the hydrophobic effect) and the configurational entropy.

Protein stability has been shown to correlate with the surface area changes accompanying folding (Rashin, 1984; Chiche et al., 1990). We thus express this contribution to the free energy, ΔG_{HB}, in terms of the change in exposed nonpolar area:

$$\Delta G_{HB} = K_H(A_U - A_F) \quad (1)$$

where A_U is the average exposed nonpolar area in the unfolded state of the structure or substructure, A_F is the corresponding area in the folded state, and K_H is a constant. Estimates of K_H based on transfer data for model compounds range from

† This work was supported in part by Grant RO1 GM41034 from the National Institutes of Health.

‡ Center for Advanced Research in Biotechnology, Maryland Biotechnology Institute, University of Maryland.

§ Weizmann Institute of Science.

∥ Institute for Advanced Computer Studies, University of Maryland.

0006-2960/91/0430-3816$02.50/0

a low value of −0.016 kcal mol^{-1} Å^{-2}, when only nonpolar atoms are included (Eisenberg & McLachlan, 1986), to a high of −0.024 kcal mol^{-1} Å^{-2}, including all atoms (Chothia, 1984).

Surface area is calculated with the Lee and Richards algorithm (Lee & Richards, 1971). Atoms are considered hydrophobic if they carry a zero partial charge in an empirical polar hydrogen force field (Dauber-Osguthorpe et al., 1988). The accessible surface areas in the folded state are calculated from the final atomic coordinates. Those for the unfolded state are obtained by the method of Shrake and Rupley (1973), where it is assumed that the frequency with which a particular conformation is observed in a set of structures reflects the population of that conformation in the denatured state. A_U values for each residue type are therefore calculated from the average accessibility of all occurrences of that residue in a set of proteins, including only the residue itself and the main chain of the neighboring residues on either side in the accessibility calculation.

Offsetting the favorable contribution to folding from the burial of hydrophobic area, the selection of essentially only one conformation represents an increase in order compared with the denatured state. This contribution to the free energy (ΔG_{CF}) may be expressed as

$$\Delta G_{CF} = n_{on}\Delta g \quad (2)$$

where n_{on} is the number of residues changing from the unfolded to the folded state and Δg is the free energy change per residue. We make the approximation that Δg is a constant independent of residue type and that it is not sensitive to changes in excluded volume during folding. Calorimetric data suggest a value for Δg of approximately 1.3 kcal mol^{-1} residue^{-1} (Murphy et al., 1990). We shall see that folding behavior depends primarily on the ratio $-\Delta g/K_H$ and is not extremely sensitive to precise parameter values.

Combining eqs 1 and 2, the free energy of any state of subfolding relative to the fully denatured state is

$$\Delta G_{fold} = K_H(A_U - A_F) + n_{on}\Delta g \quad (3)$$

where n_{on} is the number of residues folded.

Models of Nucleation Propensity. A simple and useful approximation, particularly for nucleation processes, is to assume that any stretch of chain under consideration will populate only two states: fully folded or fully unfolded, i.e., $n_{on} = 0$ or $n_{on} = N$, where N is the number of residues in the stretch. It is also useful to introduce the average change in hydrophobic burial per residue on folding:

$$a_{HB} = (A_U - A_F)/N \quad (4)$$

Then, considering the condition of $\Delta G_{fold} = 0$ in eq 3, that is, 50% of the sample folded and 50% fully unfolded, we have for the folded molecules ($n_{on} = N$)

$$a_{HB} = -\Delta g/K_H \quad (5)$$

This implies that a constant burial per residue is needed to maintain 50% folding, independent of the length of chain considered.

In a more rigorous model, we consider all possible substates of folding of a piece of chain. Then, the fraction of the population in some substate of folding, k, at equilibrium will be

$$f_k = \exp(-\Delta G_k/RT)/\sum_i \exp(-\Delta G_i/RT) \quad (6)$$

where the ΔG values for each subfolding state are given by eq 3 and the sum is over all the possible subfolding states, i.e., all possible choices of a set of n_{on} folded residues, for all values of n_{on}. The fully folded populations may be calculated for all short fragments of chain (up to about 24 residues long). For longer peptides, the number of terms to be summed becomes prohibitively large. We show later that the detailed and simple models have similar behavior, and we will use the simpler model for analysis of nucleation propensity.

Identification of Initiation Sites. We now make use of the assumption that the conformation of any significant intermediate in the folding process, such as a nucleation site, will usually be preserved in the final fully folded structure. Then, we may analyze fragments of the final structures to see if they conform to any particular folding model. To do this, the variation in the hydrophobic burial per residue in parts of protein structures is examined. A chain length, or window, is selected, of length N_W residues. The window is slid along the chain, and the hydrophobic burial relative to the denatured state is calculated for the residues currently in the window. The final structure coordinates are used, and the rest of the structure outside the window is ignored. Thus the calculated burial corresponds to that present when only the residues in the window are folded. Nineteen high-resolution well-refined globular protein structures (listed in Table I) were treated in this manner, with varying window lengths. The result is a series of hydrophobic burial profiles. Figure 1 (top panel) shows an example of such profiles for window lengths of 8, 12, and 20 residues for the lysozyme molecule from T4 bacteriophage [21zm in the Brookhaven Data Bank (Bernstein et al., 1977)]. Two features are immediately apparent. First, for the shortest 8-residue windows, the profile exceeds 50 Å^2 of hydrophobic burial per residue only once, at the segment starting at residue 119. Second, as the window length increases, the number of high burial regions (HBRs) also increases. Since the area is normalized by the number of residues, this observation suggests a possible propagation of short nucleation sites to longer, more stable, structures.

Table I summarizes the occurrence of HBRs of length 12 in the set of proteins studied. Where there are overlapping HBRs, the highest burial value position is listed. There is a consistent pattern of approximately an average of one HBR of greater then 45 Å^2/residue burial per every 50 residues. An exception to this behavior is phospholipase A2 (lbp2), which has no HBRs. The burial per residue for the complete protein structures varies from 72 to 88 Å^2/residue, excluding phospholipase A2, significantly lower with only 65 Å^2/residue.

Given these burial data, an approximate relationship between Δg and K_H can be established from a knowledge of the typical stabilities of proteins and shorter peptides (Rashin, 1984). The free energy of of a fully folded protein, ΔG_{fold}, is small and negative, in the range −5 to −15 kcal/mol. For most small fragments of proteins, ΔG_{fold} is zero or positive (i.e., they do not have a stable conformation). We can use eq 5 to estimate the ratio $-\Delta g/K_H$ in terms of a_{HB}. To ensure folding, this value must be higher than that usually observed for short peptides (i.e., more than 50 Å^2). For folded proteins, a_{HB} is approximately 80 Å^2. Thus the ratio of parameters must be

$$50 < -\Delta g/K_H < 80 \quad (7)$$

Since intermediate states of folding (eq 6) are ignored, $-\Delta g/K_H$ will be overestimated by this approximation. We shall see that the lower limit, $\Delta g = 1$ kcal mol^{-1} residue^{-1} for $K_H = -0.02$ kcal mol^{-1} Å^{-2}, produces the most reasonable estimates of folded populations for fragments and propagation behavior.

We next compare the results obtained with the simple two-state model with those of the detailed multistate model. The simple model has the advantage of directly relating hydrophobic burial per residue to nucleation propensity but ig-

Table I: Folding Pathway Characteristics for 19 Proteins[a]

	WINDOW	BURIAL	SEQUENCE	MOTIF	GROWTH	ASSOCIATION
1bp2 phospholipase A2 (bovine pancreas)						11%
A	2 13	43.0	LWQFNGMIKCKI	H	2 22	C
B	53 64	39.9	KQAKKLDSCKVL	O	52 68	
C	99 110	37.2	DRNAAICFSKVP	H	99 110	A
1lz1 lysozyme (human)						91%
A	106 117	59.6	IRAWVAWRNRCQ	K	11 117	B C
B	17 28	55.1	MDGYRGISLANW	O	11 28	A C
C	54 65	47.4	YGIFQINSRYWC	H	11 65	A B
1tim triose phosphate isomerase (chicken)						84%
A	190 201	58.8	WLKTHVSDAVAV	O	39 245	
B	90 101	57.0	VILGHSERRHVF	H	39 245	C
C	164 175	52.2	EPVWAIGTGKTA	H	39 245	B
D	132 143	49.6	EREAGITEKVVF	O	39 245	
E	42 53	48.7	APSIYLDFARQK	H	39 245	
2act actinidin (kiwi fruit)						48%
A	124 135	54.8	LQTAVTYQPVSV	H	112 217	B
B	169 180	45.8	YGTEGGVDYWIV	P	112 217	A C E
C	184 195	44.6	WDTTWGEEGYMR	O	112 217	B E
D	37 48	44.5	INKITSGSLISL	O	37 48	
E	143 154	42.8	AFKQYASGIFTG	H	112 217	B C
2sni subtilisin novo (part of complex with CI2)						52%
A	205 216	54.8	YSTYPTNTYATL	P	103 273	
B	4 15	54.6	VPYGIPLIKADK	H	4 15	E
C	164 175	52.6	IGYPAKYDSVIA	H	103 273	
D	112 123	52.6	WATTNGMDVINM	O	103 273	
E	255 266	51.2	YLGSSFYYGKGL	O	103 273	B
F	234 245	49.6	LSKHPNLSASQV	H	103 273	
2app penicillopepsin (penicillium janthinellum)						28%
A	19 30	55.4	TPVTIGGTTLNL	P	9 59	F
B	141 152	51.4	FDTVKSSLAQPL	H	122 192	F
C	155 166	49.9	VALKHQQPGVYD	P	122 192	D
D	298 309	48.8	FGDIFLKSQYVV	H	296 319	C E G
E	256 267	47.9	FSVSISGYTATV	P	217 273	D G
F	89 100	46.5	SVTVGGVTAHGQ	P	89 100	A B
G	219 230	46.0	LLLDDSVVSQYY	K	217 273	D E
2cga chymotrypsinogen A (bovine pancreas)						73%
A	199 210	51.9	LVCKKNGAWTLV	P	121 238	
B	165 176	49.0	NTNCKKYWGTKI	H	121 238	
C	44 55	47.4	GSLINENWVVTA	P	31 68	D
D	231 242	47.0	VTALVNWVQQTL	H	121 242	C
E	32 43	45.9	SLQDKTGFHFCG	P	31 68	
2lzm lysozyme (bacteriophage T4)						90%
A	115 126	57.9	TNSLRMLQQKRW	H	14 161	D
B	24 35	56.8	YYTIGIGHLLTK	P	14 161	
C	138 149	55.9	WYNQTPNRAKRV	H	14 161	
D	77 88	54.3	GILRNAKLKPVY	H	14 161	A
2mhr myohemerythrin (sipunculan worm)						79%
A	66 77	60.3	KYSEVVPHKKMH	H	25 77	B D
B	102 113	54.3	WLVNHIKGTDFK	H	25 113	A C
C	4 15	53.9	IPEPYVWDESFR	O	4 17	B
D	51 62	52.0	TTNHFTHEEAMM	H	25 62	A
2sga proteinase A (streptomyces griseus)						29%
A	70 81	63.5	RVYLYNGSYQDI	P	70 81	
B	112 123	51.6	ATVNYGSSGIVY	P	111 126	
C	18 29	46.0	FNVSVNGVAHAL	P	18 61	D
D	169 180	43.5	VTEALSAYGATV	H	169 180	C
2sns nuclease (staph. aureus)						74%
A	23 34	60.3	VKLMYKGQPMTF	P	14 140	
B	129 140	54.4	EAQAKKEKLNIW	H	14 140	C
C	98 109	50.3	MVNEALVRQGLA	H	14 140	B
3blm beta-lactamase (staph.aureus)						59%
A	47 58	55.1	AILLEQVPYNKL	H	47 58	B
B	104 115	54.3	KIIKEIGGIKKV	H	47 115	A F
C	216 227	53.0	FVYPKGQSEPIV	P	47 227	
D	73 84	52.1	SPILEKYVGKDI	H	47 84	
E	3 14	49.8	LNDLEKKYNAHI	H	3 14	
F	131 142	47.1	RYEIELNYYSPK	O	47 142	B
3rp2 serine protease (rat atypical mast cells)						66%
A	211 222	54.5	RVSTYVPWINAV	H	116 222	B C
B	32 43	53.0	GFLISRQFVLTA	P	17 110	A D
C	77 88	52.1	IHESYNSAPNLH	H	17 110	A
D	18 29	48.8	LDIVTEKGLRVI	P	17 110	B
E	151 162	47.4	EKACVDYGYYEY	H	116 222	
3tln thermolysin (B. thermoproteolyticus)						62%
A	80 91	64.9	TYDYYKNVHNRL	H	68 91	E G
B	20 31	54.6	INTTYSTYYYLQ	P	20 31	
C	112 123	53.6	NAFWNGSEMVYG	P	68 123	
D	239 250	53.1	KAAYLISQGGTH	H	68 250	
E	167 178	49.6	AISDIFGTLVEF	H	68 178	A G
F	155 166	47.7	LIYQNESGAINE	O	68 166	G
G	268 279	47.2	YRALTQYLTPTS	H	68 279	A E F
4dfr dihydrofolate reductase (E. coli)						53%
A	42 53	56.2	MGRHTWESIGRP	H	30 112	B
B	92 103	50.4	IMVIGGGRVYEQ	O	30 112	A D
C	20 31	46.5	MPWNLPADLAWF	H	20 112	D
D	104 115	44.8	FLPKAQKLYLTH	O	30 115	B C
4pti trypsin inhibitor (bovine pancreas)						65%
A	21 32	55.7	YFYNAKAGLCQT	P	18 55	B
B	1 12	41.0	RPDFCLEPPYTG	O	1 12	A
5cpa carboxypeptidase A (bovine pancreas)						89%
A	115 126	54.7	GFAFTHSENRLW	H	7 305	G J
B	89 100	54.4	NYGQNPSFTAIL	H	7 305	D E F
C	227 238	53.7	VAALKSLYGTSY	H	7 305	E H
D	25 36	53.3	LVAQHPELVSKL	H	7 305	B F I
E	294 305	51.7	WLGVLTIMEHTV	H	7 305	B C F
F	77 88	51.6	ATGVWFAKKFTE	H	7 305	B D E G J
G	276 287	51.3	RYGFLLPASQII	O	7 305	A F J
H	202 213	51.3	LLYPYGYTTQSI	O	7 305	C
I	37 48	51.1	QIGRSYEGRPIY	P	7 305	D
J	7 18	50.4	FNYATYHTLDEI	O	7 305	A F G
6ldh lactate dehydrogenase (dogfish)						87%
A	148 159	61.4	WKLSGLPMHRII	H	24 325	E F G
B	204 215	60.9	MNVASIKLHPLD	O	24 325	
C	38 49	55.2	SILMKDLADEVA	O	24 325	
D	235 246	53.4	AYEVIKLKGYTS	H	24 325	
E	114 125	53.3	VNIFKFIIPNIV	H	24 325	A G
F	274 285	53.3	VKDFYGIKDNVF	O	24 325	A
G	136 147	51.7	VSNPVDVLTYVA	H	24 325	A E
7rsa ribonuclease A (bovine pancreas)						80%
A	25 36	59.9	YCNQMMKSRNLT	H	25 36	B C
B	87 98	53.3	TGSSKYPNCAYK	P	25 98	A
C	2 13	52.9	ETAAAKFERQHM	H	2 13	A D
D	108 119	51.3	VACEGNPYVPVH	P	25 119	C

[a] For each protein, data for the 12-residue nonoverlapping windows with the highest hydrophobic burial per residue are given. The number of sites included for each protein is proportional to its length (about one per 50 residues). There are no other windows with more than 50 Å^2/residues of burial. The windows are referenced A, B, etc. For each window, the sequence position, the average burial per residue (in Å^2), the sequence, and the type of motif are listed. Sequential residue numbering is used throughout. The approximate motif types are P for hairpin loops, H for helices, K for kinked helices, and O for unclassified structures. The last two entries for each window give the propagation and association properties of the windows. "Growth" gives the limits of propagation around the window, taken from a set of 12 500 000-step Monte Carlo simulations of the folding of each protein. The title line gives the percent folding propagation for the complete protein. The association column lists the other windows in the protein in contact such that the hydrophobic burial per residue increases by 5 Å^2 or more when the contacting window is included in the burial calculation.

nores the intermediate states of folding. On a short (milliseconds to seconds) time scale, we can assume that the population of all possible states of folding in windows up to at least 16 residues long is at equilibrium (Wetlaufer, 1973). Then, the fraction of molecules fully folded can be calculated from eq 6. Figure 1 (bottom panel) shows that distribution for all the 12-residue-long windows in T4 lysozyme. The maximum population of the fully folded state found is about 10%, that is, a factor of 5 down on the 50% population implied by the two-state model for 50 Å^2 burial per residue. However, the positions of the strongest nucleating sites agree closely with those obtained by using the two-state approximation, so that burial per residue is a useful and simple indicator of nucleating propensity. The multistate model indicates a variation of 4 orders of magnitude between the most- and least-folded fractions. Thus, there is an enormous variation in nucleating potential along the chain. Although the absolute folded fraction does depend on the parameter values used, the ratio of folding fractions for any two windows is almost independent of the values chosen.

Types of Nucleating Structure. We next consider what types of structural motifs are involved in nucleation. To in-

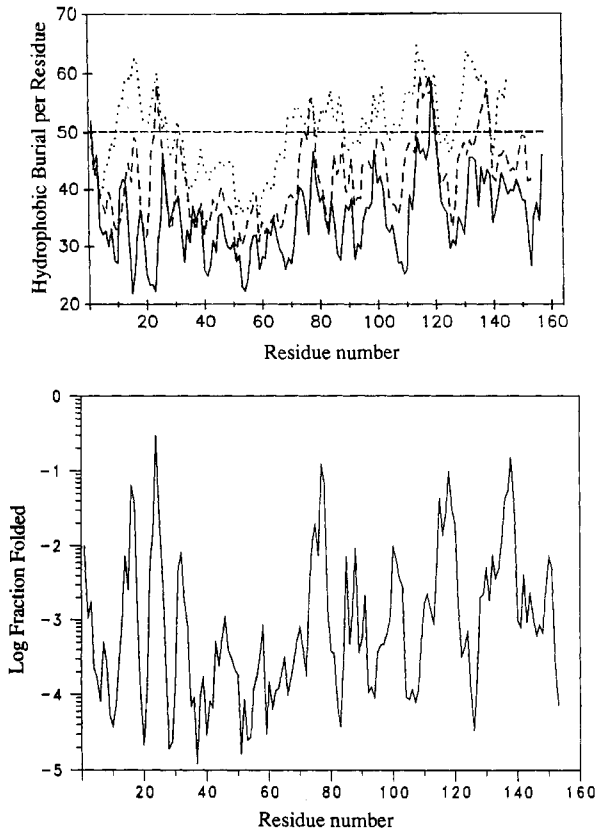

FIGURE 1: (Top panel) Variation in the average hydrophobic burial per residue in T4 lysozyme. Data are given for windows 8 (solid line), 12 (dashed line), and 20 (dotted line) residues in length. The total hydrophobic burial on going from the unfolded to folded state is calculated for each possible window in the sequence. Burials greater then 50 Å^2/residue are considered to represent strong initiation sites for folding. (Bottom panel) Population of the folded state for 12-residue segments of T4 lysozyme. The vertical axis gives the $\log_{10}$ of the fraction of the time a segment would be fully folded, according to the multistate model. The free energy contributions from hydrophobic burial (K_H = −0.02 kcal mol^{-1} Å^{-2} of hydrophobic area) within the segment and the configurational free energy (Δg = 1.0 kcal mol^{-1} residue^{-1} folded) are used to obtain a relative free energy for each possible folding substate within a window. A standard partition function is then used to obtain the population of the fully folded state.

vestigate this, a cluster analysis was performed on the high burial region (HBR) structures. A simple neighborhood membership clustering algorithm (Fu & Lu, 1978; Unger et al., 1989) was used, with the minimum root mean square (rms) distance between the sets of α-carbon atoms of pairs of HBRs (McLachlan, 1979) as a metric. The first cluster is formed by finding all the HBRs within a specified threshold rms distance of the first HBR and then adding all the HBRs within the threshold rms of any of those neighbors, and so on until no new members are found. The next cluster is then formed in the same way, starting with the first remaining unclustered HBR. The process is continued until all HBRs have been assigned to a cluster. The method produces an order-independent set of clusters but will work only if the objects are well separated in the clustering space. Such turns out to be the case with the HBRs. The number of clusters generated depends on the rms threshold used in defining neighbors. At low rms distances (2.0 Å for windows of 8 residues, 2.5 Å for 12 and 16 residues) three main clusters form. These main clusters can be classified as β hairpin loops, helices, and kinked helices. The clearest distinction is between the hairpin loops and the other two predominantly helical clusters. Structures classified as helices usually include one or two residues beyond

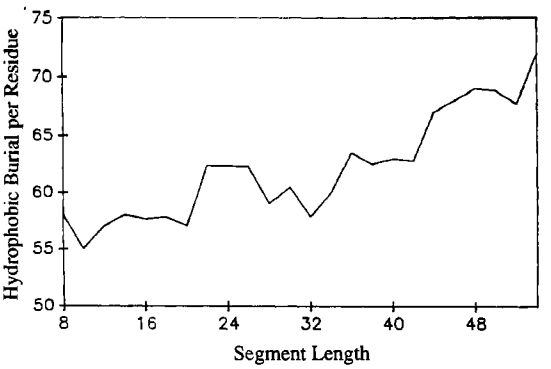

FIGURE 2: Growth of a folding initiation site by propagation. The average hydrophobic burial per residue is plotted as a function of the number of residues in the segment. The first point is for the burial of an 8-residue-long initiation site beginning at residue 119 in the C-terminal domain of T4 lysozyme. Pairs of residues are added to each end of the segment in turn, and the burial is recalculated. As the segment grows, the burial generally increases, indicating a greater stability of the folded structure. Thus, once the initiation site has formed, structure will develop around it. Parameter values are as in Figure 1.

the helical region and are often imperfect. Kinked helices are formed by two helical segments making an acute angle. HBR structures that do not cluster into one of the main groups are sometimes distorted versions of those motifs. Occasionally ω loop structures (Leszczynski & Rose, 1986) are also found. Table I lists the structural type of each HBR.

Pathway after Nucleation. The final structures also provide a means of reconstructing events after folding initiation. Tests can be made of the relative importance of diffusion/collision (Karplus & Weaver, 1976; Bashford et al., 1988) and propagation (Wetlaufer, 1973) mechanisms.

When association is involved we expect to find some initiation sites packed against each other in the final structure. The stabilization provided by such packing can be assessed by calculating the increased average hydrophobic burial per residue, Δb, when a pair of nucleation sites are considered, compared with the two sites isolated from each other:

$$\Delta b = (\Delta B_{W1+W2} - \Delta B_{W1} - \Delta B_{W2})/(N_{W1} + N_{W2}) \quad (8)$$

where W1 and W2 are two initiation sites of lengths N_{W1} and N_{W2}. ΔB_{W1+W2} is the total hydrophobic burial of the associated sites, and ΔB_{W1} and ΔB_{W2} are the total burials for the folded, isolated sites. We consider sites to be significantly stabilized by each other if Δb is greater than 5 Å^2/residue, that is, an increase of approximately 10% in the burial per residue compared with that of the isolated sites. Table I shows which 12-residue-long high burial regions (HBRs) associate by this criterion. Data for larger windows show increasing association. Thus the association mechanism becomes more important in the later stages of folding.

Where propagation is the mechanism, the stability of the folded portion should increase as more residues are added to it. According to the two-state model, a simple indicator of this is the change in the burial per residue as the window is lengthened. Figure 2 shows a plot of the behavior of this quantity for the growth of one of the putative initiation sites in T4 lysozyme. Burial tends to increase up to the limit of the calculation, indicating a strong propagation behavior. The existence of such increasing stability is an approximate indicator of propagation propensity. We will see in the next section that a Monte Carlo simulation of the kinetic process of propagation provides a more rigorous model.

Examination of the parts of structures that are able to propagate shows that this type of behavior is closely correlated

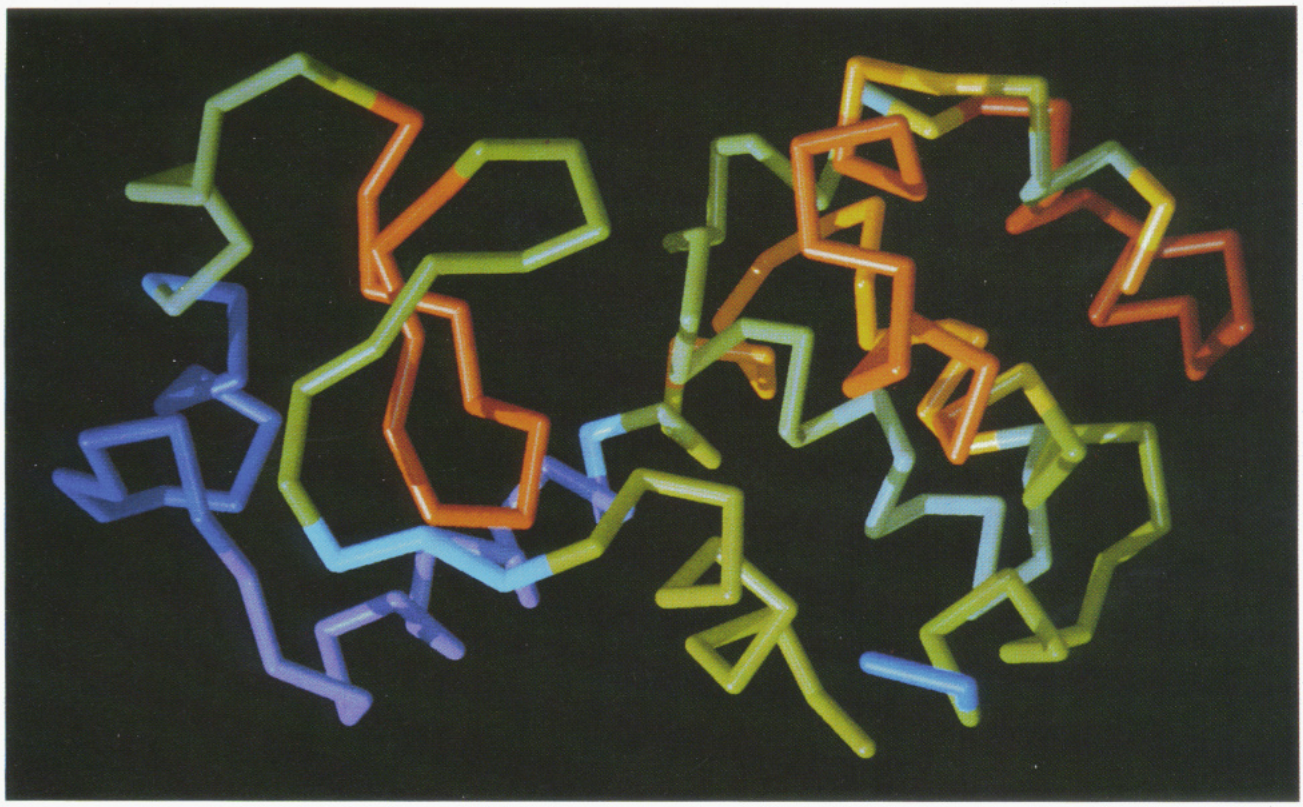

FIGURE 3: Nucleation propensity for T4 lysozyme. The backbone of the protein is colored according to the average hydrophobic burial per residue within each 12-residue segment. The higher the burial, the redder the color; the lower, the bluer. High hydrophobic burial per residue is assumed to represent a relatively high level of folding early in the pathway. In the mainly helical C-terminal domain at right there are three principal initiation sites. Two of these sites are the ends of helices and the adjacent turns. The third is centered on a kink in a helix. In the folding simulations, structure propagates out from these sites by building helical segments against the template of those already formed. For the N-terminal domain, initiation of folding takes place mainly in a single 12-residue segment formed by a hairpin loop. Growth from this is through bends on each end of that region, which lay extended pieces of chain on top of the nucleation site. Because adjacent pieces of structure are formed from adjacent pieces of chain for almost the whole structure, the folding can proceed predominantly by a propagation mechanism.

with the folding of adjacent regions of the linear chain against one another. Thus, for example, β-structures in which strands adjacent in the sequence are adjacent in the structure are found to propagate. Conversely, when there is a break in the folding pattern, such as jumping over one or more strands, there is a breakdown in the ability to propagate. This type of folding behavior has been suggested before as an explanation of why adjacent strand and Greek key topologies are so common in protein structures (Richardson, 1981). Figure 3 shows how the topological continuity of T4 lysozyme facilitates folding by a propagation mechanism.

Monte Carlo Simulation of Folding Behavior. The nucleation models consider semiequilibrium conditions early in folding, but folding of a complete protein is a kinetic process. We have therefore developed a simulation algorithm that emulates kinetic behavior for the cases where propagation is the dominant folding mechanism. The folding state of the polypeptide chain is defined in terms of the state of the peptide link between each pair of residues. Links are considered to be in one of two states, folded or unfolded. For a chain N residues long there are thus 2^{N-1} possible folding states in the model, ranging from all links "off" (fully denatured) to all links "on" (fully folded). A set of adjacent folded or on links form a window of $n_{on} + 1$ folded residues, terminated at each end by an unfolded (off) link or a chain end.

In order to follow the kinetic behavior implied by the model, we define the relative on and off rates for each link for each possible state of folding. The ratio of the on to off rates for a link is obtained by an extension of eq 3 as

$$K_{eq} = \exp[-(K_H \Delta A + \Delta g)/RT] = k_{on}/k_{off} \qquad (9)$$

where ΔA is the increase in hydrophobic burial on altering the condition of the link from off to on. To extract the individual rate constants we make use of the fact that both $K_H \Delta A$ and Δg represent entropic contributions to the free energy of the folding state. Then, there is no mechanical force driving the transition in either direction. This is easiest to appreciate for the off to on transition, where the peptide link is envisioned to be randomly fluctuating in conformation (i.e., changes in the ψ, ϕ angles) under the influence of thermal energy. The frequency with which the native conformation is encountered will depend on the size of conformational space, represented by Δg. More quantitatively

$$k_{on} = \tau \exp(-\Delta g/RT) \qquad k_{off} = \tau \exp(K_H \Delta A/RT) \qquad (10)$$

Here τ strictly depends on the residue type. In particular, where cis/trans isomerizations are involved (Garel & Baldwin, 1973) and perhaps also for population of the left handed α helix region (Shäfer et al., 1984), a relatively large value of τ is appropriate. We treat τ as a constant, independent of residue type (1 for convenience in the simulation algorithm). A similar expression for k_{off} has recently been used in a diffusion/collision model (Bashford et al., 1988).

For a strict propagation process, the change in hydrophobic burial, ΔA, on changing a link from off to on depends only on the number of consecutive links immediately to the left and right (n_L and n_R) of it that are already in the on state:

$$\Delta A = A_{(n_L+n_R)} - A_{n_L} - A_{n_R} \qquad (11)$$

that is, the area buried when all the links in the sequence through the $n_L + n_R$ residues are in the native conformation,

less the area buried when the two separate stretches n_R and n_L are in the native conformation, with a break in folding between them. The required set of hydrophobic burials for all substretches of chain is precomputed from the knowledge of the final structure [$(N^2 - N)/2$ values for N residues]. Some properties of this model, particularly the two states for every residue and the concept of noninteracting locally folded regions, have been used before in pathway analysis (Miyazawa & Jernigan, 1982).

Given the set of off and on rates, we can use a Monte Carlo procedure to follow folding. We proceed as follows: (a) A random order of links is generated. (b) Each link is considered in turn, in the established order. (c) For each link considered, a random number r, $0 < r < 1$, is selected. (d) (i) If the link is currently in the on state and if $r < \exp(K_H \Delta A/RT)$, change the state to off; otherwise, leave the state on. (ii) If the link is currently in the off state and if $r < \exp(-\Delta g/RT)$, change the state to on; otherwise, leave the state off. (e) Repeat steps b–d until an equilibrium distribution of states is obtained.

This procedure generates the proper time-dependent behavior implied by the model (Metropolis et al., 1953; Taketomi et al., 1975). For a particular choice of the ratio $\Delta g/K_H$, the folding of a protein by propagation can thus be followed. Initially all links are set to off. If propagation is sufficient to produce folding, the simulation will proceed to a state in which nearly all the links are on nearly all of the time. Otherwise, links continue to fluctuate between on and off indefinitely. One of the most strongly propagating structures is T4 lysozyme. Almost all of this protein will fold by propagation within the range of $\Delta g = 1.5$ kcal mol^{-1} residue^{-1} and $K_H = -0.020$ kcal mol^{-1} Å^{-2} to $\Delta g = 1.0$ kcal mol^{-1} residue^{-1} and $K_H = -0.015$ kcal mol^{-1} Å^{-2}. For larger values of $-\Delta g/K_H$, the structure will oscillate and never fully fold. The lower the value of $-\Delta g/K_H$, the fewer fluctuations are observed before folding is attained. To assess folding behavior for a set of proteins, we have used the previously discussed parameter values of $K_H = -0.02$ kcal mol^{-1} Å^{-2} and $\Delta g = 1$ kcal mol^{-1} residue^{-1}.

The top and middle panels of Figure 4 show two examples of the propagation behavior of T4 lysozyme. Equilibrium distributions of folding were reached in less than 200 000 steps (about 1200 transition trials/link). At the first sampling of the folding state, after 4000 steps, most of the chain is still unfolded, with randomly distributed folded interresidue links. In the example of Figure 4 (top panel), the region around residue 80 becomes folded by the next sample. The rest of the chain continues to fluctuate, but after 30 000 steps, regions around residues 20, 100, 120, and 140 have folded. By 40 000 steps, the three different folding initiations in the C-terminal domain have merged, and all but the C-terminus of that domain is stably folded. The single nucleation event in the N-terminal domain expands more slowly but eventually encompasses the whole of the domain except the N-terminus. The N-terminal region continues to fluctuate and does not fold stably, in spite of repeated local folding around residue 5.

Repeat runs using different random numbers result in different detailed behavior. Figure 4 (middle panel) shows another run. Here the earliest region to fold is around residue 120, and a rather different pattern of folding is seen in the N-terminal domain, with a minor denaturation after about 70 000 steps. Averaging over a number of simulation runs provides the average folding time for each residue. We define the folding time of a link as the point after which it is 90% or more in the folded state. Figure 4 (bottom panel) shows the average folding time curve derived from 12 runs. The

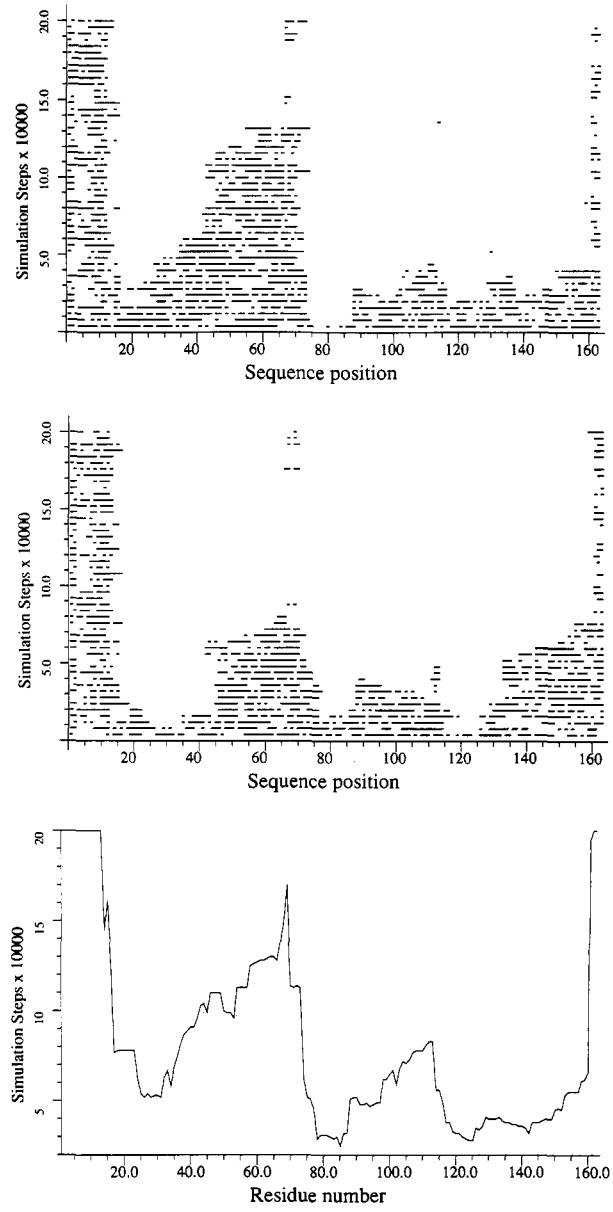

FIGURE 4: Monte Carlo simulations of the folding of T4 lysozyme by propagation. The top and middle panels show the progress of folding in two separate simulation runs. Folding time starts at zero at the bottom of the vertical axis. The state of folding is shown for 50 equally spaced samples out of the 200 000 folding steps in the simulations. Each sample is a horizontal line on the plot. Where a dash appears, the corresponding interresidue link was unfolded at that point. A blank indicates a folded link. At the first sample, after 4000 steps, most links are still unfolded, but by the end of the run all but the N-terminus is stably folded. The difference between the folding behavior in the two runs indicates the amount of variability in folding pathways. The bottom panel shows the average folding time for each residue in the sequence, obtained by averaging the results from 12 such simulation runs. On average, the initiation sites identified from the plot in Figure 1 fold first. Alternative initiation sites in the C-terminal domain mean that a plurality of pathways occur there. The single site in the N-terminal domain restricts folding to essentially one pathway.

earliest regions to fold are the three high burial regions (HBRs) in the C-terminal domain. Because of the presence of three different initiation sites for folding, the whole domain folds rapidly and with a small variation in average folding time. Folding in the N-terminal domain begins on average at the single HBR there and takes longer. Thus, the simulation supports the conclusion that the HBRs will act as initiation regions for folding.

The average folding time curve for T4 lysozyme shows that the first 15 residues in the chain are not stably folded and that the helix linking the N- and C-terminal domains is slow to fold. Examination of the structure shows that there are three interdomain salt bridges, one of which has recently been shown by mutagenesis to contribute approximately 3 kcal/mol to the stability of the folded structure (Anderson et al., 1990). Inclusion of salt bridge effects in the simulation is a desirable improvement.

We have investigated the propagation behavior of the set of proteins. The degree to which propagation takes place around each initiation site and the fraction of the structure folded by the propagation mechanism are given in Table I. Equilibrium distributions of folding were reached in less than 500 000 steps for all the structures. Thus, the results are insensitive to the number of steps in the simulation—tests with 10 times as many steps gave similar results—but do vary with the ratio of $-\Delta g/K_H$ used. The earliest residues to fold are the most insensitive to the parameters, while the latest to fold will not do so for higher values of $-\Delta g/K_H$. At one extreme, phospholipase A2 exhibits almost no propagation properties. The two strongest propagators are the two lysozymes (11zl and 21zm). The behavior of these two structures is more similar than their weak structural similarity (Matthews et al., 1981) implies, suggesting a high degree of conservation of the folding mechanism. On the other hand, although the two mammalian serine proteases included [chymotrypsinogen (2cga) and rat mast cell protease (3rp2)] show similar strong propagation patterns, the bacterial member of the family, *Streptomyces griseus* protease A (2sga), is a rather poor propagator. Some proteins, such as subtilisin (2sni) and actinidin (2act), propagate well for only part of the structure. In general the N- and C-terminal regions of the chains propagate poorly. This is partly explained by the limited amount of adjacent structure contributing to burial. That does not appear to be the whole story, though, and propagation is markedly weaker at the N termini than the C termini. The terminal regions do contain a large number of high burial regions (Table I), and these partly compensate for the lack of longer range interactions. For T4 lysozyme, other interactions, particularly salt bridges, may play a role in stabilizing these parts of the chains.

Comparison of Results with Available Experimental Data. Experimental data support the view that some regions of proteins have significant populations of the final structure before folding is complete, i.e., there are initiation sites for folding (Wright et al., 1988). The two cases for which coordinates are available, helices in ribonuclease S (Shoemaker et al., 1985) and myohemerythrin (Dyson et al., 1988), are both classified as nucleation sites in Table I. The N-terminal portion of the myohemerythrin helix is found to be the most strongly nucleating in our analysis, whereas the NMR data indicate the C-terminal portion is the more folded for the isolated peptide in water. The recent finding of a hairpin loop conformation for part of that helix region in association with an antibody fragment (Stanfield et al., 1990) is consistent with the sorts of population expected to be helical both from experiment and from the partition function calculations. These latter give a population of only 4% for the fully folded conformation in the initiation site, the highest in the protein. It would thus be expected that some antibodies raised to the fragment would bind another conformation. Synthesis and characterization of the other peptides predicted to be nucleation sites should provide extensive tests of the model.

The multistate model data for the initiation of folding of phospholipase A2 show that the whole curve is depressed by a factor of about 100 compared with the other proteins. In most other respects the structure appears normal (Drenth et al., 1987), with a usual composition, secondary structure, and charge–charge interactions but with a large number of disulfide bridges. There is more exposed hydrophobic surface than usual, presumably to bind to membrane surfaces, where the protein operates. Also, the amount of hydrophobic burial per residue in the final structure is significantly less than for any other protein examined (65 Å^2/residue as opposed to 72 Å^2/residue for the least buried of the rest). Binding to a lipid surface is required to obtain full activity of the enzyme, in addition to the removal of a 6-residue pro piece from the N-terminus (Volwerk & de Haas, 1982) (several other proteins included here have pro forms). It is thus likely that the full folding of this protein utilizes the membrane environment in some way, with the approximate structure maintained by the disulfide bridges.

Average folding times from the Monte Carlo simulation can be compared with the individual residue folding rates obtained from quenched proton exchange NMR data. Data are so far available on only two proteins, ribonuclease A (Udgonkar & Baldwin, 1988) and cytochrome *c* (Roder et al., 1988). The example of ribonuclease is complicated by the presence of intact disulfide bridges during the experiment, and propagation behavior is prevented in cytochrome *c* by a rate-limiting proline isomerization at position 76. Data should shortly be available for T4 lysozyme, where no such complications are known, so that the results should be directly comparable with figure 4 (bottom panel).

Mutagenesis together with fluorescence measurements has established that Trp 74 is one of the first residues to be buried during the folding of dihydrofolate reductase (4dfr) (Garvey et al., 1989). The first region to fold in the simulations contains residues 38–60, around a predicted initiation site (Table I). The region 60–110, containing Trp 74, is the next to fold. The rest of the protein does not fold by a propagation mechanism, so that we conclude that the early folding intermediates would include Trp 74. However, the very earliest portion to fold includes Trp 47, a residue whose fluorescence does not appear to be affected by early folding. Our data suggest that region of chain folds very rapidly. We speculate that it may fold too rapidly to be observed experimentally or that it may not fully denature. Mutagenesis work on this and other proteins offers good prospects for further testing the model. For instance, it should be possible to slow folding of the N-terminal domain of T4 lysozyme by the use of mutants that destabilize the single nucleation site predicted there but leave the final stability unaltered.

Discussion. The model leads to the following conclusions about the nature of folding pathways: (1) The level and frequency of the regions of highest hydrophobic burial in these proteins (except for phospholipase A2) is consistent with their suggested role in early folding. (2) High levels of hydrophobic burial are associated largely with a few types of structure. These structures are not confined to regular secondary structures such as helices. (3) The variation in the expected folded population for different regions of chain early in folding is very large, showing a strong early pathway signal (Figure 1, bottom panel). (4) The increase in the average hydrophobic burial per residue with increasing fragment size is a characteristic property of parts of these structures (Figure 2) and implies an ability to grow by a propagation process early in folding. For some proteins, propagation alone appears to be

an adequate mechanism to complete folding. The simulation probably underestimates the role of propagation in folding, since relatively short loop regions between well-packed units are sufficient to interrupt it. On the other hand, the extensive association of predicted nucleation sites in the final structure shows a clear role for the diffusion/collision mechanism, particularly in the later stages of the folding process. (5) In a propagation simulation, the stability of each link in the chain depends on the folding state of its neighbors. In this respect, propagation may be classified as cellular automata behavior. (6) Propagation restricts the possible pathways, providing a means of reducing confusion by not allowing wrong regions to associate. The two largely α/β proteins examined, triose phosphate isomerase (1tim) and lactate dehydrogenase (61dh), are both strong propagators, so that the frequent use of the α/β motif in large domains may partly reflect the tidiness of the resulting folding pathways. (7) The different detailed pathways found in the repeat simulations are suggestive of the "jigsaw" model of folding (Harrison & Durbin, 1985), with many different ways of putting the pieces together. This is a misleading analogy, however. There are often very strong preferences for the order in which assembly of the structure occurs. In the case of the N-terminal domain of T4 lysozyme, a single initiation site controls the folding, while in the multinucleating C-terminal domain, many detailed pathways are possible. (8) The conundrum of how such rapid folding is achieved in such a large space of possible conformations is solved in this model by the importance of contacts between residues close together in the sequence versus longer range interactions. The significance of this effect has been emphasized before by other explorers of pathways (Wetalaufer, 1973; Montelione & Scheraga, 1989; Lesk and Rose, 1981; Miyazawa & Jernigan, 1982). The pathways found are very short compared with a random search through the possible subfolding states: For T4 lysozyme, complete folding takes place in under 10^6 steps, although the model (for 163 links in the main chain of this protein) has $2^{163}(10^{49})$ possible states. We may define the acceleration of folding due to following the pathway as the ratio of these two quantities, giving an acceleration factor of 10^{43} for T4 lysozyme.

The model provides the basis for development of algorithms for the determination of structure from sequence, by mimicking the natural folding process. Such an approach starts with the detection of nucleation sites and their structural type. In general, it is clear that the conformation of short regions of chain are only partly determined by the local sequence, hence the very limited success of secondary structure predictions (Kabsch & Sander, 1983) and the finding that pentapeptides of identical sequence have radically different conformations in about half the instances known (Kabsch & Sander, 1984). However, initiation sites are by definition regions of chain where local sequence *does* determine conformation. Inspection of the sequence data for the different classes of initiation site data (Table I) suggests a sequence signal that may be adequate for site identification. More detailed initiation site structure should be obtainable by using systematic conformational search (Moult & James, 1986) or related procedures. These regions may then be extended a few residues at a time, determining their conformation by the systematic conformational search type procedures, refining the structure, and then adding more residues. Related "build up" procedures (Vásquez & Scheraga, 1985) have been tried with encouraging results. An algorithm that builds the most stable regions of chain first will come closer to following a realistic pathway.

Acknowledgments

We thank Osnat Herzberg and Phil Bryan for helpful and stimulating discussions.

References

Anderson, D. E., Becktel, W. J., & Dahlquist, F. W. (1990) *Biochemistry 29*, 2403–2408.
Anfinsen, C. B. (1970) *Science 181*, 323–330.
Bashford, D., Cohen, F. E., Karplus, M., Kuntz, I. D., & Weaver, D. L. (1988) *Proteins: Struct., Funct., Genet. 4*, 211–227.
Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Simanouchi, T., & Tasumi, M. (1979) *J. Mol. Biol. 112*, 441–484.
Chiche, L., Gregoret, L. M., Cohen, F. E., & Kollman, P. A. (1990) *Proc. Natl. Acad. Sci. U.S.A. 87*, 3240–3243.
Chothia, C. (1984) *Annu. Rev. Biochem. 55*, 537–572.
Covell, D. G., & Jernigan, R. L. (1990) *Biochemistry 29*, 3287–3294.
Dauber-Osguthorpe, P., Roberts, V. A., Osguthorpe, D. J., Wolff, J., Genest, M., & Hagler, A. T. (1988) *Proteins: Struct., Funct., Genet. 4*, 31–47.
Dill, K. A. (1990) *Biochemistry 29*, 7133–7155.
Drenth, J., Dijkstra, B. W., & Renetsedar, R., (1987) in *Biological Macromolecules and Assemblies* (Jurnak, F. A., & McPherson, A., Eds.) Vol. 3, pp 287–312, Wiley, New York.
Dyson, H. J., Rance, M., Houghton, R. A., Wright, P. E., & Lerner, R. A. (1988) *J. Mol. Biol. 201*, 201–217.
Eisenberg, D., & McLachlan, A. D. (1986) *Nature 319*, 199–203.
Finkelstein, A. V., & Shakhnovich, E. I. (1989) *Biopolymers 28*, 1681–1694.
Fu, K. S., & Lu, S. Y. (1978) *IEEE Trans. Syst. Man. Cybern. SMC 8*, 381–389.
Garel, J. R., & Baldwin, R. L. (1973) *Proc. Natl. Acad. Sci. U.S.A. 70*, 3347–3351.
Garvey, E. P., Swank, J., & Matthews, C. R. (1989) *Proteins Struct., Funct., Genet. 6*, 259–266.
Harrison, S. C., & Durbin, R. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 4028–4030.
Karplus, M., & Weaver, D. L. (1976) *Nature 260*, 404–406.
Kunihiro, K. (1989) *Proteins: Struct., Funct., Genet. 6*, 87–103.
Lee, B., & Richards, F. M. (1971) *J. Mol. Biol. 55*, 379–400.
Lesk, A. M., & Rose, G. D. (1981) *Proc. Natl. Acad. Sci. U.S.A. 78*, 4304–4308.
Leszczynski, J. F., & Rose, G. D. (1986) *Science 234*, 849–855.
Levinthal, C. (1968) *J. Chim. Phys. 65*, 44–45.
Kabsch, W., & Sander, C. (1983) *FEBS Lett. 155*, 179–182.
Kabsch, W., & Sander, C. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 1075–1078.
Matthews, B. W., Remington, S. J., Gruetter, M. G., & Anderson, W. F. (1981) *J. Mol. Biol. 147*, 545–558.
McLachlan, A. D. (1979) *J. Mol. Biol. 128*, 49–79.
Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953) *J. Chem. Phys. 21*, 1087–1092.
Miyazawa, S., & Jernigan, R. L. (1982) *Biochemistry 21*, 5203–5213.
Montelione, G. T., & Scheraga, H. A. (1989) *Acc. Chem. Res. 22*, 70–76.
Moult, J., & James, M. N. G. (1986) *Proteins: Struct., Funct., Genet. 1*, 146–163.

Murphy, K. P., Privalov, P. L., & Gill, S. J. (1990) *Science 247*, 559-561.
Rashin, A. A. (1984) *Biopolymers 23*, 1605-1620.
Richardson, J. S. (1981) *Adv. Protein Chem. 34*, 167-339.
Roder, H., Elove, G. A., & Englander, S. W. (1988) *Nature 335*, 700-704.
Shäfer, L., Klimkowski, V. J., Momany, F. A., & Chuman, H. (1984) *Biopolymers 23*, 2335-2347.
Shoemaker, K. R., Kim, P. S., Brems, D. N., Marqusee, S., York, E. J., Chaikin, I. M., & Baldwin, R. L. (1985) *Proc. Natl. Acad. Sci. U.S.A 82*, 2349-2353.
Shrake, A., & Rupley, J. A. (1973) *J. Mol. Biol. 79*, 351-371.
Skolnick, J., & Kolinski, A. (1989) *J. Mol. Biol. 212*, 787-817.
Stanfield, R. L., Fieser, T. M., Lerner, R. A., & Wilson, I. A. (1990) *Science 248*, 712-719.
Taketomi, H., Ueda, Y., & Gō, N. (1975) *Int. J. Pept. Protein Res. 7*, 445-459.
Udgonkar, J. B., & Baldwin, R. L. (1988) *Nature 335*, 694-699.
Unger, R., Harel, D., Wherland, S., & Sussman, J. L. (1989) *Proteins: Struct., Funct., Genet. 5*, 355-373.
Vâsquez, M., & Scheraga, H. A. (1985) *Biopolymers 24*, 1437-1447.
Volwerk, J. J., & de Haas, G. H. (1982) in *Lipid-Protein Interactions* (Jost, P. C., & Griffith, O. H., Eds.) Vol. 1, pp 69-149, Wiley, New York.
Wetlaufer, D. B. (1973) *Proc. Natl. Acad. Sci. U.S.A. 70*, 697-701.
Wright, P. E., Dyson, H. J., & Lerner, R. A. (1988) *Biochemistry 27*, 7167-7175.

Site-Specific Mutagenesis of Conserved Residues within Walker A and B Sequences of *Escherichia coli* UvrA Protein†

Gary M. Myles,‡,§ John E. Hearst,‖ and Aziz Sancar*,‡

Department of Biochemistry and Biophysics, University of North Carolina School of Medicine, Chapel Hill, North Carolina 27599-7260, and Department of Chemistry and Division of Chemical Biodynamics, Lawrence Berkeley Laboratory, University of California, Berkeley, Berkeley, California 94720

Received November 8, 1990; Revised Manuscript Received January 15, 1991

ABSTRACT: UvrA is the ATPase subunit of the DNA repair enzyme (A)BC excinuclease. The amino acid sequence of this protein has revealed, in addition to two zinc fingers, three pairs of nucleotide binding motifs each consisting of a Walker A and B sequence. We have conducted site-specific mutagenesis, ATPase kinetic analyses, and nucleotide binding equilibrium measurements to correlate these sequence motifs with activity. Replacement of the invariant Lys by Ala in the putative A sequences indicated that K37 and K646 but not K353 are involved in ATP hydrolysis. In contrast, substitution of the invariant Asp by Asn in the B sequences at positions D238, D513, or D857 had little effect on the in vivo activity of the protein. Nucleotide binding studies revealed a stoichiometry of 0.5 ADP/UvrA monomer while kinetic measurements on wild-type and mutant proteins showed that the active form of UvrA is a dimer with 2 catalytic sites which interact in a positive cooperative manner in the presence of ADP; mutagenesis of K37 but not of K646 attenuated this cooperativity. Loss of ATPase activity was about 75% in the K37A, 86% in the K646A mutant, and 95% in the K37A-K646A double mutant. These amino acid substitutions had only a marginal effect on the specific binding of UvrA to damaged DNA but drastically reduced its ability to deliver UvrB to the damage site. We find that the deficient UvrB loading activity of these mutant UvrA proteins results from their inability to associate with UvrB in the form of $(UvrA)_2(UvrB)_1$ complexes. We conclude that UvrA forms a dimer with two ATPase domains involving K37 and K646 and that the work performed by ATP hydrolysis is the delivery of UvrB to the damage site on DNA.

(A)BC excinuclease is the enzymatic activity resulting from the coordinated action of UvrA, UvrB, and UvrC proteins which excise a wide variety of modified nucleotides from DNA by hydrolyzing the eighth phosphodiester bond 5′ and the fifth phosphodiester bond 3′ to the damaged base (Sancar & Sancar, 1988; Grossman & Yeung, 1990; Van Houten, 1990). A current model for the reaction mechanism is as follows (Orren & Sancar, 1989, 1990). UvrA, which is an ATPase and a DNA binding protein with higher affinity for damaged DNA than for undamaged DNA (Seeberg & Steinum, 1982), interacts with UvrB to form a $(UvrA)_2(UvrB)_1$ complex. UvrA delivers UvrB (which has a cryptic ATPase activity and no affinity for DNA) to the damage site and dissociates from the UvrB-damaged DNA complex. UvrC binds to this complex and either directly or indirectly mediates the dual single-strand DNA incisions. ATP binding and hydrolysis by UvrA and UvrB (which becomes activated in the ternary UvrA-UvrB-DNA complex; Seeley & Grossman, 1989, 1990) are required during several stages of the overall reaction. This study is aimed at characterizing the ATPase activity of UvrA and its role in specific steps of the excision nuclease activity.

† This work was supported by Grants GM32833 from the National Institutes of Health and PCM351212 from the National Science Foundation and in part by a grant CTR1872 from the Council for Tobacco Research Inc. and by U.S. Department of Energy Grant DE-AC03-76SF-00098.

* Correspondence should be addressed to this author.
‡ University of North Carolina School of Medicine.
§ Present address: Fred Hutchinson Cancer Research Center, 1124 Columbia St., Seattle, WA 98104.
‖ University of California, Berkeley.

0006-2960/91/0430-3824$02.50/0