Chem., **79**, 2361 (1975).

(11) A computer program named ECEPP (Empirical Conformational Energy Program for Peptides) was used.[10] The Fortran computer program for ECEPP, its description, and the associated structural and energy parameters are available on magnetic tape from the Quantum Chemistry Program Exchange, as program No. QCPE 286. See footnote 60 of ref 10 for the procedure to obtain this material.

(12) P. N. Lewis, F. A. Momany, and H. A. Scheraga, *Biochim. Biophys. Acta*, **303**, 211 (1973).

(13) S. S. Zimmerman and H. A. Scheraga, *Macromolecules*, **9**, 408 (1976).

(14) A. E. Tonelli, *J. Mol. Biol.*, **86**, 627 (1974).

(15) J. E. Dennis and H. H. W. Mei, Technical Report No. 75-246 (1975), Department of Computer Science, Cornell University, Ithaca, N.Y. 14853.

(16) M. J. D. Powell, *Comput. J.*, **7**, 155 (1964).

(17) S. S. Zimmerman, M. S. Pottle, G. Némethy, and H. A. Scheraga, *Macromolecules*, **10**, 1 (1977).

(18) M. S. Pottle, unpublished results.

(19) Z. I. Hodes, unpublished results, prior to the work of ref 5.

(20) P. N. Lewis, F. A. Momany, and H. A. Scheraga, *Isr. J. Chem.*, **11**, 121 (1973).

(21) Actually the type I′, II′, and III′ bends could not be formed in TPRK because ϕ of Pro is restricted to −75°. These bend conformations were calculated with $\phi_{\mathrm{Pro}} = -75°$, so that they are really type IV bends. Nevertheless, these bends are reported in the results as type I′, II′, and III′ bends since they were derived from the same dihedral angles for ψ_{Pro}, ϕ_{Arg}, and ψ_{Arg}.

(22) ϕ_{Pro} was again restricted to −75°.

(23) Y. Isogai, G. Némethy, and H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.A.*, **74**, 414 (1977).

(24) P N Lewis, F. A. Momany, and H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.A.*, **68**, 2293 (1971).

(25) J. L. Crawford, W. N. Lipscomb, and C. G. Schellman, *Proc. Natl. Acad. Sci. U.S.A.*, **70**, 538 (1973).

(26) K. D. Kopple and A. Go, *Biopolymers*, **15**, 1701 (1976).

(27) K. G. R. Pachler, *Spectrochim. Acta*, **20**, 581 (1964).

(28) M. T. Cung, M. Marraud, and J. Neel, *Macromolecules*, **7**, 606 (1974).

(29) I. D. Rae, S. J. Leach, and H. A. Scheraga, work in progress.

(30) H. A. Kent, *Gen. Comp. Endocrinol.*, **29**, 253 (1976).

A Method for Predicting Nucleation Sites for Protein Folding Based on Hydrophobic Contacts[1]

R. R. Matheson, Jr.,[2] and H. A. Scheraga*

Department of Chemistry, Cornell University, Ithaca, New York, 14853.

Received December 27, 1977

ABSTRACT: A method, based on hydrophobic bonding, is introduced for predicting nucleation sites for protein folding. The amino acid sequence of the protein is searched for pockets of nonpolar residues whose (negative) free energy of interaction compensates for the increase in free energy that is required to bring them into contact to form hydrophobic pockets. The predicted nucleation sites (and their associated electrostatic properties) are used to rationalize some equilibrium and kinetic results on protein folding, including the relative amplitudes of absorption by transient species observed in kinetic studies. Finally, the method proposed here is compared to the proline-isomerization hypothesis for protein folding proposed by Brandts and co-workers.

I. Introduction

Several recent reviews on the nature of protein folding have been published.[3-5] There is convincing evidence from many systems that the process is not a two-state one[6-11] although it is highly, cooperative.[12,13] However, the number, order, and conformational characteristics of the stages through which a protein passes during folding are all topics of speculation. For the case of the thermally induced unfolding of ribonuclease A (i.e., passage from the native to more disordered conformations, with the disulfide bonds intact), a large body of experimental evidence is consistent with the hypothesis that a rather well-defined sequential pathway exists.[14] Yet, even for this well-studied system, it is not established whether or not this pathway is the exact reverse of the refolding process (i.e., from the disordered to the more native structure), whether it bears any relationship to the folding when disulfide bonds are broken, or whether it is applicable to the cooperative conformational changes induced by other denaturants.

A variety of computational procedures have been proposed for the determination of the structures of native proteins,[15-24] and one of these[18] has led to the hypothesis that the folding process can be considered to take place in three steps.[25] However, most of these consider only the energetic aspects of the problem without taking the conformational entropy into account explicitly. A number proportional to the conformational entropy can be obtained in principle by the procedure of Crippen,[22] but in practice it is practical to compute only an estimate.[22] The entropy does appear implicitly (in terms of a partition function) in statistical mechanical treatments of protein folding[23,24,26,27] and is treated indirectly by at least one computational method[18] which selects conformations from a distribution, and includes the effect of solvation in a contact free energy. In the treatment presented here, we shall account explicitly (albeit approximately) for the conformational entropy and for the free energy of formation of hydrophobic bonds.

In this paper, we shall consider the initial aspects of folding, viz., the nucleation process. More specifically, we shall examine the question of the existence of nucleation sites, and the molecular structural basis for their formation. By the term "nucleation site" we mean a specific conformation of a limited section of the polypeptide chain whose existence can significantly increase the rate of formation of the native structure of the protein from its unfolded state(s). The existence of the nucleation site is a necessary but not a sufficient condition for the formation of the native structure of a protein from the nascent polypeptide chain; i.e., it may be possible for the nucleating site to form and not produce proper folding if the additional, cooperative interactions which follow it can be prevented from occurring, e.g., by the presence of a denaturant or if the pH is unsuitable. Some general classes of nucleation sites have been described by Tsong et al.[28] and by Tanaka and Scheraga.[18,25] Brandts et al.[29] have proposed that the cis-trans isomerization of proline can act as a nucleation step for folding.

In considering the initial aspects of folding, we shall examine the possibility that nucleation can be accomplished by formation of a specific pocket in the polypeptide chain, stabilized by hydrophobic contacts. A method for identifying the residues involved in this specific pocket from a knowledge of the amino acid sequence of a protein will be discussed. We shall show that application of this method yields results in

0024-9297/78/2211-0819$01.00/0

qualitative agreement with the experimental facts on the folding of several proteins.

II. Model for Nucleation by Formation of Pockets Involving Hydrophobic Bonds

For some time, it has been recognized that the tendency of nonpolar groups to minimize their contact with water, i.e., to form "hydrophobic bonds", is an important factor contributing to the formation of the native structure of a protein.[30-33] Even at temperatures above the midpoint of the thermal unfolding transition, T_m, this tendency toward hydrophobic contact will be important. Indeed, the endothermic nature of the formation of hydrophobic bonds makes them more stable at moderately elevated temperatures. Thus, one possible way of characterizing the *unfolded* state of a protein is as an ensemble of interconvertible conformations differing in their arrangements of hydrophobic contacts. Some of these hydrophobic contacts (not necessarily native ones) will be sufficiently strong so that they will persist in many of the accessible conformations of the unfolded protein. Others will be relatively weak, and not make any significant contribution to the ensemble of conformations under a given set of conditions (temperature, pressure, and solvent composition). The question arises as to whether all of these unfolded forms with different arrangements of hydrophobic bonds are equivalent conformations from which the folding process may start. The basis assumption of this paper is that they are not, and that the conformation (reminiscent of the hydrophobic pockets discussed earlier[34]) which acquires the most stability from hydrophobic bonding is the nucleation site. Since the number of nonpolar groups is large, roughly 40% of the residues in proteins being nonpolar,[35] the total number of contacts between nonpolar groups can reasonably be supposed to remain about the same in all of the species of the ensemble. Thus, the various conformations which together comprise the unfolded form of a protein will have similar thermodynamic and presumably spectroscopic properties (but, see discussion in section IVB). The nucleating pocket, when formed (under conditions which permit folding of the remainder of the chain), allows rapid, directed formation of the complete native conformation; its absence (even under conditions where folding may occur) prevents folding, independent of the state of the remaining nonpolar groups.

To this point, we have ignored all nonbonded interactions which occur in proteins except the hydrophobic contact. We now consider these other types of interactions. Presumably, all of the ionizable groups of an unfolded protein will be exposed to the solvent. Because of the high dielectric constant of aqueous salt solutions, and the relatively flexible nature of the unfolded protein, it seems to be a rather good first approximation to ignore any influence of electrostatic interactions unless charged groups are constrained to be near one another by hydrophobic interactions. Thus, we shall consider electrostatic interactions only when the formation of a hydrophobic pocket introduces a loss of flexibility and potentially leads to large electrostatic interactions between charged groups constrained to be near one another. We also shall ignore intramolecular hydrogen bonds. The rationale for this is that, in the unfolded form, most potential hydrogen-bonding groups will interact with the excellent hydrogen-bonding solvent, water. Of course, hydrogen bonding will no longer contribute negligibly to the conformational energy once folding has started, since solvent will be progressively excluded. Another category of nonbonded interactions known to be significant for protein structure is the intraresidue (here, called short range) interactions.[5] In what is certainly a crude approximation, but one leading to great simplification of the calculations, we assume that, for all *unfolded* states, all amino acids have the same effective number, Ω, of accessible, isoenergetic positions of internal rotation. The validity of this assumption increases as the temperature increases. Although this assumption is known to be incorrect,[36,37] and this neglect of the short-range interactions will lead to an overestimate of the entropy loss in creating a hydrophobic interaction, we shall try to compensate for this shortcoming by overestimating the effect of nonpolar associations in stabilizing nucleation sites. We have taken Ω as equal to 7.[34]

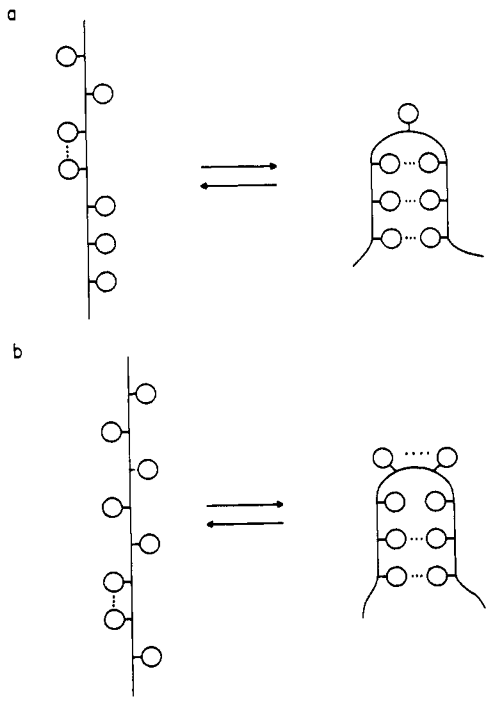


Figure 1. Schematic representation of nucleation step in which a hydrophobic pocket is formed from an ensemble of unfolded species [which *may* contain neighbor–neighbor hydrophobic bonds[34] as indicated by the species on the left side of the equilibria in (a) and (b)]. In (a), only one side chain is involved in the "turn", and it does not participate in hydrophobic bonds with nearby residues. In (b), two side chains (which themselves can form a neighbor–neighbor hydrophobic bond) are involved in the "turn", and the pocket is shown as an imperfect one, in that one pair of side chains is not sufficiently nonpolar to form a hydrophobic bond. Hydrophobic bonds are indicated by dotted lines. For pockets of types a and b, at least five and four residues, respectively, are assumed to be required, so that the pockets are large enough for the bends to be stereochemically feasible. Pockets of both types are included in the search algorithm.

To summarize, we shall neglect most electrostatic effects, all hydrogen bonding, and the influence of short-range interactions on the nucleation process. The first two of these may reasonably be ignored, and we shall try to compensate for the neglect of the third.

According to our model, the nucleation step is considered to be the reaction in which a section of an unfolded polypeptide chain bends back on itself and creates a pocket which permits hydrophobic bonding.[34] We assume that any section of a polypeptide chain can participate in such a reaction, independent of the conformations of remote residues. While this assumption seems quite reasonable for the initial step of a folding scheme for a non-crosslinked protein, it has to be altered when we treat the activation energy of the nucleation reaction in crosslinked proteins (see section IIIC). This reaction is shown schematically in Figure 1, where hydrophobic bonds are indicated by dotted lines; a hydrophobic bond between two adjoining residues is indicated in both *unfolded* forms of Figure 1. The probability of formation of such neighbor–neighbor interactions in the unfolded forms can be quite high,[34] and such bonds make a significant contribution to the free energy of the unfolded state. Also, it may be noted that no hydrophobic bond has been shown between the third and sixth residues of the pocket in Figure 1b. While such residues may contribute no hydrophobic bonds to stabilize the pocket, their internal rotation must be restricted nonetheless.

For the purpose of assessing the role of nucleation in the *kinetics* of protein folding, it is further assumed that fluctuations between competing pockets involve a transition through the unfolded conformation. For example, we may regard a pocket which involves residues i through $(i + 9)$ to be in competition with one from residues $(i + 3)$ through $(i + 15)$ since they consist of different pairings of some of the same residues. We assume that the $i - (i + 9)$ pocket must unfold, and then the chain must refold, to form the $(i + 3) - (i + 15)$ pocket from the "completely" unfolded chain. This is consistent with our basic concept of the nucleation site as an independently folding local structure.

To summarize, our model for the nucleation of protein folding rests on the following five basic assumptions: (1) the nucleation step consists of the formation of a pocket in the polypeptide chain; (2) interactions among nonpolar groups are the dominant stabilizing forces in this pocket, and the pocket which gains the most stability from these nonpolar interactions (subject to favorable electrostatic interactions implied in assumption 3) can be identified as the nucleation site; (3) electrostatic interactions are the only ones other than hydrophobic bonds that need to be considered, and they contribute *only when hydrophobic bonds constrain charged groups to be near one another;* (4) the stability and rate of formation of a nucleating pocket are independent of residues that are not part of the pocket; (5) all pockets can be considered as being formed from the "completely" unfolded polypeptide chain. Specific calculations based on this model will be discussed in the next section, and applications to individual proteins will be presented in section IV.

III. Identification of the Nucleating Contacts

A. Proteins with Disulfide Bonds Reduced. The free energy change for reactions similar to those illustrated in Figure 1 is given by

$$\Delta G_{\rm nuc} = \sum_k (1/2)(\Delta G_{\rm k}^{{\rm H}\phi,{\rm min}} + \Delta G_{\rm k}^{{\rm H}\phi,{\rm max}}) - \sum_l f_l \Delta G_l^{{\rm H}\phi,{\rm min}} + \left(N - \sum_l f_l\right) RT \ln \Omega \quad (1)$$

where $\Delta G^{{\rm H}\phi,{\rm min}}$ and $\Delta G^{{\rm H}\phi,{\rm max}}$ are the free energies of formation of minimum- and maximum-strength hydrophobic bonds,[31] respectively, between pairs of nonpolar side chains, N is the number of peptide bonds involved in the pocket, R is the gas constant, T is the absolute temperature, and Ω is the effective number of isoenergetic positions for internal rotation (expressed "per peptide bond"). The index k runs over the hydrophobic bonds in the pocket, the index l runs over the neighbor–neighbor bonds broken (left side of equilibria in Figure 1) in order to form the pocket, and f_l is the fraction of unfolded species containing the lth neighbor–neighbor hydrophobic bond. The quantity f_l is given[34] by

$$f_l = [1 + \Omega \exp(\Delta G^{{\rm H}\phi,{\rm min}}/RT)]^{-1} \quad (2)$$

These are standard state free energies. However, to avoid complicating the notation, we have omitted the superscript zero on all ΔG's.

Equation 1 implies that only *pairwise* hydrophobic bonding exists in the pocket. However, there is an additional stability in such a pocket because the interior side chains are shielded from the solvent to a greater extent than is implied by pairwise hydrophobic bonding. We thus add an increment of free energy to take this effect into account. For a consideration of the free energy of formation of a *triple* contact,[31] a reasonable amount of free energy to add is ¾ of the free energy of transfer,[38,39] $\Delta G^{\rm Tr}$, of a side chain from water to a nonpolar solvent. However, in small pockets, this degree of shielding varies among the residues in the pocket. From an examination of molecular models, it appears reasonable to select an *average* shielding per residue by multiplying $0.75\Delta G^{\rm Tr}$ by $(N - 3)/8$ and to apply this factor to *each* residue in the pocket. Thus, this increment is zero for $N = 3$ (four residues) and increases linearly with N up to $0.75\Delta G^{\rm Tr}$ for $N \geq 11$; it is our judgment, from an examination of molecular models, that 11 peptide bonds constitute the minimum size pocket in which the interior residues would be completely shielded from the solvent. For pockets with $N \geq 11$, the value of $0.75\Delta G^{\rm Tr}$ is used for all N. We thus modify eq 1 to

$$\Delta G_{\rm nuc} = \sum_k \{(1/2)(\Delta G_k^{{\rm H}\phi,{\rm min}} + \Delta G_k^{{\rm H}\phi,{\rm max}}) + 0.75[(N - 3)/8]\Delta G_k^{\rm Tr}\} - \sum_l f_l \Delta G_l^{{\rm H}\phi,{\rm min}} + \left(N - \sum_l f_l\right) RT \ln \Omega \quad (3)$$

where $\Delta G_k^{\rm Tr}$ corresponds to the two side chains involved in the kth contact. The use of the factor $0.75[(N - 3)/8]\Delta G_k^{\rm Tr}$ for *all* residues in a pocket is an overestimate for all but the most perfectly packed conformations; this overestimate compensates, to some extent, for the overestimate of the entropy (due to neglect of short-range interactions), mentioned in section II.

With the aid of a computer, the complete amino acid sequence of a protein is searched for stable pockets, using eq 3 to calculate the value of $\Delta G_{\rm nuc}$ for each possible pocket of the two types shown in Figure 1. As a potential pocket increases in size (i.e., as k increases), the value of $\Delta G_{\rm nuc}$ generally increases. This behavior arises because long sequences of nonpolar residues are uncommon in proteins, and only a substantial negative contribution from the $\Delta G_k^{\rm Tr}$ term (see eq 3) permits the positive contribution from the entropy loss (last term of eq 3) to be fully compensated. With this in mind, we terminated the search for larger pockets centered on any particular residue when $\Delta G_{\rm nuc}$ exceeded +15 kcal/mol. The chances of $\Delta G_{\rm nuc}$ decreasing (for further increases of k) to a value near zero are remote. The values of $\Delta G^{{\rm H}\phi,{\rm min}}$ and $\Delta G^{{\rm H}\phi,{\rm max}}$ are those of Némethy and Scheraga,[31] with one exception, viz., when one or both of the partners is an aromatic residue. As discussed in section 3c of ref 31, the aromatic residues were *assumed* to form hydrophobic bonds with their large side chains kept as close as possible to their own backbones. If this is not the case, then the total number of water molecules removed from contact with the nonpolar surface is decreased, and the stability of the hydrophobic bonds involving aromatic side chains will be smaller than the values published in ref 31. The assumption of section 3c of ref 31 is reasonable for a maximum-strength hydrophobic bond, but it is not reasonable for a neighbor–neighbor minimum-strength hydrophobic bond. Thus, the free energy of formation of the hydrophobic bonds in a significant fraction of neighbor–neighbor contacts involving aromatic residues will be overestimated (to an extent that depends on the size of the other partner) unless some account is taken of the motion of the aromatic side chain away from its own backbone. We have chosen to account for this effect approximately by assigning to the minimum-strength hydrophobic bond when an aromatic side chain is involved in a neighbor–neighbor contact one-half of the free energy listed in ref 31.

In order to decide which pairs of residues contribute to the various sums in eq 3, the amino acids are assigned to three

Table I
Predicted Nucleation Sites for 14 Proteins

Protein	Nucleation site	Pocket size (no. of residues)	ΔG_{nuc},[a] kcal/mol	Δ,[b] kcal/mol
Ribonuclease A (cow)	106–118	13	−1.2[c] (−0.8)	0.9[c] (1.1)
Chymotrypsinogen A (cow)	206–216	11	−2.1[c] (−2.1)	1.9[c] (1.9)
Trypsinogen (cow)	91–94	4	1.4[c] (1.4)	0.3[c] (0.3)
	212–222	11	1.3[c] (1.5)	
Lysozyme (hen egg white)	54–64	11	1.2[c] (1.4)	0.2[c] (0.2)
	55–65	11	1.1[c] (2.0)	
α-Lactalbumin (cow)	89–92	4	1.5[c] (1.8)	0.2 (0.1)
Pancreatic trypsin inhibitor (cow)	2–12	11	1.7[c] (2.1)	0.4[c] (0.2)
Cytochrome *c* (horse)	79–82	4	1.6	0.4
Parvalbumin (carp)	28–31	4	1.5	0.3
Staphylococcal nuclease (Aureus)	29–39	11	1.9	0.3
	30–40	11	1.9	
	35–38	4	1.9	
	36–39	4	1.9	
Myoglobin	103–115	13	0.5	0.7
Hemoglobin (human β chain)	31–42	12	−0.9	0.4
Proinsulin (pig)	15–28	14	−0.4[c] (−0.2)	1.2[c] (0.8)
Elastase (pig)	105–115	11	0.3	0.2
Carboxypeptidase A (cow)	279–289	11	−1.8[c] (−1.8)	0.4[c] (0.4)

[a] Calculated at 298 K according to eq 3. No consideration is given to electrostatic interactions at this stage. When included at a later stage, electrostatic interactions were assigned such a magnitude that they did not change the selection of the pocket with the most favorable ΔG_{nuc}, but they could modify the magnitude of ΔG_{nuc} and hence that of α_2 (defined in eq 4). [b] Δ is defined as the difference (in kcal/mol) between the most favorable value of ΔG_{nuc} and that for the next most favorable pocket. If Δ is less than or equal to 0.1 kcal/mol, then the two pockets are considered to have equal values of ΔG_{nuc} and, hence, are equally favorable nucleation sites. In such a case, more than one nucleation site is listed for the protein, and the value of Δ pertains to the difference between the values of ΔG_{nuc} for the most favorable one of these and for the next best (unlisted) pocket. [c] With disulfide bonds intact. The values in parentheses correspond to the protein with its disulfide bonds reduced. The nucleation sites are found to be the same in the presence and absence of disulfide bonds.

groups: (I) Ala, Cys/2, Cys, Phe, Ile, Leu, Met, Pro, Val, and Trp; (II) Glu, Lys, Gln, Arg, and Tyr; and (III) Asp, Gly, His, Asn, Ser, and Thr. The interactions included in the sum over k are those of type I with type I or type I (except Ala) with type II. The sum over l includes only interactions between residues of type I. Interactions involving residues of type III are not included in the sums except for their contribution to ΔG^{Tr}. The data of Nozaki and Tanford[38] and Levitt[39] are used for ΔG^{Tr}. The pocket with the lowest value of ΔG_{nuc}, from eq 3, is predicted to be the nucleation site. Predictions for 14 proteins are presented in Table I.

As an aid in assessing the uniqueness of our choice of the most favorable nucleation pocket, we have tabulated the difference between the best value of ΔG_{nuc} and the next best value for each protein in the last column of Table I. For most proteins, this difference is less than RT (0.6 kcal/mol), so that the best pocket is not favored overwhelmingly according to the ΔG_{nuc} criterion. On the other hand, of the enormous number of possible nucleation sites, only a few pockets (10 for reduced trypsin inhibitor which is a bad case) have values of ΔG_{nuc} within RT of the best pocket, so that the number of reasonable alternatives is small. It is not known whether protein folding starts from one or several nucleation sites, so that this lack of an overwhelmingly favored nucleation pocket may accurately reflect the physical situation. However, given the number and severity of the approximations made in computing ΔG_{nuc}, this cannot be assessed at this time. For simplicity, we consider only the pocket(s) with the best values of ΔG_{nuc}, with the caveat that there *may* be a *few* more pockets of importance as nucleation sites.

B. Proteins with Disulfide Bonds Intact. If a protein contains one or more disulfide bonds, which are never broken in the folding process, as is the case in some types of experiments, we must consider the possible effect of the disulfide bonds on nucleation. One effect of cross-linking is to reduce the conformational entropy (i.e., change the value of Ω) of the unfolded molecule relative to that of the reduced unfolded molecule.[40–42] This effect is small and usually negligible in comparison to other uncertainties. Thus, it will not be necessary to treat it further in the present work. A second effect of cross-linking is the introduction of an additional covalent bond so that the chain segment in the pocket may follow along the disulfide bond instead of the peptide backbone. For example, the disulfide which bridges half-cystines 26 and 84 in ribonuclease A would permit new pockets of the type -85-84-26-25- or -83-84-26-25-, etc.

This second effect can be treated within the framework of our basic model. Recognizing the flexible and nondirectional nature of hydrophobic bonds, the differences in geometry between a peptide link and the -CSSC- linkage in cystine can be assumed to be negligible. Thus, the problem is reduced to that of obtaining values for Ω and for the "side-chain" parameters for this "new" residue. The crude nature of our model does not justify an effort to obtain highly accurate estimates for these values. Thus, we have made the simple reasonable choices: that $\Omega = 7$, that the value of $\Delta G^{\mathrm{H}\phi}$ is that for the asparagine side chain, and that the free energy of transfer of asparagine applies to this new type of residue. For pockets that follow the peptide chain instead of the disulfide bond (e.g., the sequence -24-25-26-27-28- in ribonculease, where residue 26 is cross-linked to residue 84), the side chain of residue 26 is treated as a methionine side chain.[31]

C. Tests for Reliability of Identification of Nucleation Site. The success or failure of the method proposed here for identifying nucleation sites in proteins cannot be judged on the basis of a single, generally accepted test. This is because neither the existence of nucleation sites nor their common properties has been established. Therefore, we summarize here those observations for which our model can account, in order to judge the reliability of the method. Applications to individual proteins are treated in the next section, and Table II presents a summary of those results.

Table II
Some Properties of the Nucleation Sites

Protein	Nucleation site	Bend	α_2 (predicted)	pH dependence of α_2
Ribonuclease A	106–118	113–114[a]	0.61[b]	Strong
Chymotrypsinogen A	206–216	203-204-205[c]	0.92	No
Trypsinogen	91–94	92–93[d]	0.09	Weak
	212–222	222–223		No
Lysozyme	54–64	55–56	0.09	Weak
	55–65			
α-Lactalbumin	89–92	*e*, *f*	0.06	Weak
Pancreatic trypsin inhibitor	2–12	5–6	0.04	No
Cytochrome *c*	79–82	*g*	0.05	No
Parvalbumin	28–31	*e*	0.06	No
Staphylococcal nuclease	29–39	37–38	0.03	No
	30–40		0.03	Weak
	35–38		0.03	No
	36–39		0.03	
Myoglobin	103–115	*e*	0.29	Strong
Hemoglobin	31–42	*e*	0.59	Strong
Proinsulin	15–28	20–23[h]	0.27	Strong
Elastase	105–115	107–108–109	0.26	No
Carboxypeptidase	279–289	284–285	0.49	No

[a] The central two residues in a β bend, found by using the X-ray coordinates from the Brookhaven Data Bank. [b] The number given is the fraction of molecules that have the predicted nucleation pocket intact in the unfolded state of the protein at a low pH (ca. 2). [c] The central three residues of a multiple β bend. For this protein the location of the bend was calculated from the X-ray data of the active protease instead of the zymogen so that the multiple bend at 203-204-205 of chymotrypsinogen is the same as the one that is centered on residues 55-56-57 of the C chain of α-chymotrypsin. [d] The bends cited here are for the crystal structure of β-trypsin (diisopropylphosphoryl inhibited) from cow pancreas since the coordinates of trypsinogen were not available. [e] All or most of the nucleation site is found in an α-helical conformation in the native structure. See discussion in section IIIC of the text. [f] The helix referred to in footnote *e* is assigned by homology with lysozyme; see ref 43. [g] No bend is observed. Residue 80 (Met) is coordinated to the iron atom of the heme. [h] Residues in a hairpin bend in the native structure of the B chain of *insulin* as described in ref 44.

Our model requires the formation of a pocket or bend in a polypeptide chain as the initial step in folding and contains no provision that the pocket should disappear in subsequent stages of folding. Thus, we should expect to find such bends in the final native structures, and they should be located near the centers of our predicted nucleation sites. Because of the weak and flexible character of hydrophobic bonds, some distortion of the initial pocket could occur during folding, but the existence of a bend somewhere in each nucleation site is a reasonable prediction of our model. Where X-ray coordinates are available, this is usually borne out by experiment (see Table II and the following section). For the case of cytochrome *c*, the nucleation site contains a residue (Met 80) which is coordinated to the iron atom of the prosthetic heme group. For the other cases in which a bend is not found, the native conformation of all or most of the predicted nucleation site is an α helix. Moreover, with the exception of the 103–115 pocket of myoglobin, the pockets are found to lie very close to the ends of helical sequences. The difference in residue backbone dihedral angles between a bend and a helical segment can be quite small. In fact, a helical sequence of two residues is indistinguishable from a bend.[37] Thus, it is easy to imagine that a pocket, that initially has randomly ordered residues at its borders, can be transformed into a portion of α helix, since most of the conformational changes involved in the transformation would involve the initially randomly ordered residues; the bend in the pocket would require little if any distortion. Since the short helical sequences typically found in proteins are frequently rather irregular, it is quite plausible that predicted nucleating pockets (particularly short ones) may be classified as parts of helical sequences (especially near the ends) in the final native structures.

Since nucleation is postulated to be a necessary step for the generation of the native structure of a protein, we expect the residues in our predicted nucleation site generally to be conserved in the sequences of a protein isolated from various species. Absolute conservation is not required, but the nonpolar character must be preserved and generally few changes are to be expected (see section IV). For several proteins, enough sequences from different species are available so that the degree of sequence conservation can be measured quantitatively. For this purpose, we have defined a correlation coefficient as the number of identical amino acids in the sequence of a protein from a particular species (compared to some arbitrarily selected reference sequence) divided by the chain length of the protein. Analogously, a correlation coefficient is defined for the nucleation site by counting only the common residues within the homologous sites of the reference and test proteins and dividing by the length of the predicted nucleation site. Deletions or insertions are not counted as changes, and the reference and test sequences are aligned prior to calculation of the correlation coefficients. Then the mean values of this correlation coefficient for the entire protein and for the predicted nucleation site can be compared. This is done in Table III, where we have computed the probability that the difference in these two means for a given protein is due to chance, *assuming* that the distributions are independent. This assumption is obviously not fulfilled rigorously since the sequences are aligned prior to examination, and changes in the region of the nucleation site are also counted in the entire sequence. Nevertheless, the calculation is useful as an indication of how very well the sequences of the nucleation site are in fact conserved.

Several protein systems exhibit biphasic kinetics of refolding when studied with fast-reaction techniques.[7–9,46–52] Biphasic behavior itself is a natural consequence of our model, where one phase (usually the slowest) of refolding corresponds to the formation of the nucleating pocket, and the second phase corresponds to further folding. Moreover, the activation parameters, relative amplitude of absorbance of species (compared to the fast kinetic phase), and relaxation times are interpretable in terms of this model (see section IV).

The formation of the nucleating pocket in a denatured protein is considered to result from random fluctuations in

Table III
Conservation of the Amino Acid Sequence at the Nucleation Site

Protein	Ref sequence	No. of sequences compared	Mean corr coeff[a] of whole sequence	Mean corr coeff[a] of nucleation site sequence	Probability[b] that diff in means is not significant
Ribonuclease A[c]	Cow pancreas	23	0.808	0.946	<0.01
Cytochrome *c*[d]	Horse	27	0.800	0.889	<0.01
Myoglobin	Sperm whale	10	0.842	0.908	<0.1
Hemoglobin[e]	Human β chain	14	0.816	0.952	<0.01
Proinsulin	Pig	8	0.487[f]	0.946	<0.01

[a] Defined in section IIIC of the text. [b] This probability is computed (using a standard t test) as if the mean correlation coefficients of the whole sequence and of the nucleation site are independent. [c] Sequences and alignments taken from ref 45. [d] Alignments taken as in ref 35. [e] Alignment of sequences from sheep and cow β chain with that from human was accomplished by treating Val 1 in human chain as an insertion. [f] This value is so small because the peptide that is excised in the conversion of proinsulin to insulin is quite variable in sequence from species to species. However, the nucleation site is *not* in this peptide.

conformation among the residues of the pocket, starting from any of the ensemble of conformations which permit creation of the pocket. In a protein *without* disulfide bonds included in the nucleation site, the activation parameters are assumed to be determined largely by the process of hydrophobic-bond formation (solvent exclusion). The energy of activation, E_A, associated with the exclusion of solvent will be taken as the proper fraction of the enthalpy of transfer associated with the side chains of the pocket, i.e., $0.75[(N - 3)/8]\Delta H_k^{\text{Tr}}$. Thus, the energy of activation should depend on temperature as the enthalpy of formation of hydrophobic bonds does; i.e., ΔH^{Tr}, and thus E_A, should be positive but decrease with increasing temperature.

For proteins *with* intact disulfide bonds which create a loop spanning the nucleation site, it is not possible at present to make even a semiquantitative estimate of the activation parameters. The presence of such a cross-link would require a correlation of more than one backbone bond rotation in the process of pocket formation. If we assume that the influence of such a cross-link will dominate over the other factors that determine the properties of the nucleation step, then at least we can say that all proteins with such a cross-link would be expected to have approximately the same activation energy. This is borne out by experiment.[29] While experiments designed to detect "crankshaft" (or correlated) motions have been unsuccessful for noncross-linked polymers in solution,[53,54] their existence in cross-linked or in solid-phase systems is assured.[54] The extent to which such correlations will raise the activation energy for a conformational transition may be quite large and may depend on the size of the loops created by the cross-links and by their interdependency.[42] At least as an initial approximation, their effect can reasonably be assumed to be large and insensitive to specific details of the protein; i.e., as stated above, all proteins with cross-links will be assumed to have approximately the same activation energy.

The relative amplitudes of absorbance of species in the fast and slow kinetic phases can be predicted semiquantitatively by our model. By using the values of ΔG_{nuc} for the nucleating pocket and for all other pockets that compete with it (i.e., contain one or more of the same residues), the fraction, α_2, of the fast-folding intermediate (see eq 6) can be calculated as

$$\alpha_2 = \frac{\exp(-\Delta G_{\text{nuc,min}}/RT)}{1 + \sum_i \exp(-\Delta G_{\text{nuc},i}/RT)} \qquad (4)$$

where the "min" subscript denotes the lowest value of ΔG_{nuc}, which is identified with the nucleating pocket, and the index i runs over all competing pockets. This procedure is independent of the presence or absence of disulfide bonds, except for the specific procedures (mentioned in section IIIB) for calculating ΔG_{nuc} when disulfides are involved in the nucleation site.

The "slow" phase of protein folding can exhibit relaxation times, τ, which vary over at least four orders of magnitude,[52] from tens of seconds to milliseconds. For proteins without cross-links, we may reasonably except the relation

$$\tau_{\text{slow}}^{-1} = Ae^{-E_A/RT} \qquad (5)$$

to hold, with E_A obtained as discussed above, and A to be a temperature-independent constant. By analogy with the behavior of synthetic polymers, A can be regarded as a function of the molecular weight of the polymer.[55] However, A would also be expected to depend on the details of solvent–polymer interactions[55] and, thus, would be a complicated entity. For purposes of qualitative discussion of available data (see section IV), we have taken A as proportional to the square root of the molecular weight of a protein.[55] For proteins with disulfide cross-links included in the nucleation site, the situation is even more difficult. Again we may suppose that cross-linking effects dominate and, thus, the relaxation times for the slow step in the folding of proteins containing disulfide bond(s) will all be of the same order of magnitude (as is the case[29]). Thus, eq 5 will be used only to order the values of τ_{slow}^{-1} for uncross-linked proteins, and we cannot say more at this time about proteins with cross-links than that they should all have approximately the same value of τ_{slow}^{-1}, provided that a disulfide bond spans all or part of the predicted nucleation site (which *is* the case for the proteins considered here).

Finally, we may ask why nucleation must be the "slow" step in protein folding since it involves conformational changes of only a few residues (~10) while complete folding affects many residues (~10^2). We are not aware of any reason why this should be so. Empirically, the fast step follows the slow step in refolding of the systems studied to date (e.g., see ref. 9), but our model does not require that nucleation be a relatively slow step. Conceptually, nucleation is a *first* step in refolding, but this step can be slower or faster than later steps. These subsequent steps of folding presumably make use of the nucleation site as a template, or as a gross restriction on chain flexibility or in some other manner, in those cases where later folding is much faster than nucleation. Our model suggests that the nonpolar character of a nucleation site could be important for these interactions in the later steps as well as for the intrinsic stability of the nucleation site.

IV. Results for Specific Proteins

A. Ribonuclease A (Bovine). The predicted nucleation site for ribonuclease A is the 13-residue pocket from Ile 106 to Val 118. This prediction correlates well with the existence of a β turn centered on Asn 113-Pro 114 in the X-ray structure of ribonuclease S. In the sequences 106–118 of ribonuclease from 23 mammalian species[45] there are only 15 substitutions, five of which involve conversions of Ile to Val at positions 106 and 107 or of Val to Ile at position 108, seven involve exchanges of polar residues at positions 111 or 113, and the remaining three are substitutions at position 115 (Tyr in cow, Phe in muskrat, Pro in pig, and Ser in bull semen). This minor amount of generally conservative substitution is also consistent with the identification of this 13-residue sequence as the nucleation site. It should be noted that, in the crystal structure of ribonuclease S,[56] there is a pronounced loop which permits nonpolar interaction between Val 108 and Phe 120, which reasonably can be supposed to be the "distorted" final structure of the initial 106–118 pocket.

Immunological experiments[57] suggest that the nucleation site is near the C terminus. Additionally, many experimental studies[14] indicate that, in the reversible thermally induced unfolding, the C-terminal portion of the chain from Lys 104 to Val 124 loses its native structure within the thermal transition region for the entire molecule.

Extensive fast-reaction studies[9,46,51] on ribonuclease A (with intact disulfide bonds) have revealed biphasic kinetic behavior consistent with the scheme[9]

$$\mathrm{U}_1 \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathrm{U}_2 \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} \mathrm{N} \qquad (6)$$

where U_1 and U_2 are the slow and fast refolding forms, respectively, of the unfolded enzyme, and N is the native form. We would identify U_2 as the species possessing the 106–118 pocket and Tyr 115 as the residue responsible for the *slight* difference in the p*K*'s of the nitrotyrosine derivative which is observed between U_1 and U_2.[58] Since the cystine link between residues 58 and 110 was intact in all of the kinetic studies, the energy of activation and relaxation time for the conversion of U_1 to U_2 would be those characteristic of other proteins with cross-links spanning all or part of the nucleation site. Indeed, these parameters are similar in magnitude[29] for ribonuclease A, chymotrypsinogen, α-chymotrypsin, trypsinogen, trypsin, and lysozyme.

In the refolding of cross-linked ribonuclease A, the fraction of molecules, α_2, in the state U_2 has been observed[59] to be temperature independent but pH dependent. This behavior can be explained by a competition between our predicted pocket ($\Delta G_{\mathrm{nuc}} = -1.3$ kcal/mol) and the less stable ones involving residues Lys 104–Glu 111 ($\Delta G_{\mathrm{nuc}} = 6.2$ kcal/mol), His 105–Glu 111 ($\Delta G_{\mathrm{nuc}} = 5.1$ kcal/mol), and Lys 61–Glu 111 ($\Delta G_{\mathrm{nuc}} = 5.8$ kcal/mol); the latter loop includes a disulfide bond as part of the "backbone". At pH values well below the p*K* of Glu 111, there are no electrostatic interactions in these four pockets, and the value of α_2 calculated with eq 4 is 0.61. If we assign an energy of -5 kcal/mol to an electrostatic interaction between positively and negatively charged side chains (see footnote *a* of Table I) in the 104–111, 105–111, and 61–111 pockets, then these pockets will become more favorable as the pH increases and Glu 111 becomes charged. In other words, the predicted *nucleating* pocket (106–118) becomes relatively less favored, and α_2 decreases to a value of 0.24. Experimentally,[59] α_2 is ~0.2 at pH 2 and decreases to zero at pH 5. Our model accounts for the decrease in α_2 qualitatively but not quantitatively. (Of course, the model could be made to fit exactly at pH 5 by taking a value more negative than -5 kcal/mol for the electrostatic energy.) However, by application of this method of accounting for electrostatic interactions in all proteins, and retaining the value of -5 kcal/mol, the model can predict whether or not the value of α_2 will depend on pH or not. These predictions are given in the last column of Table II, but there are no experimental data available at the present time (for proteins other than ribonuclease A) with which to compare these predictions. An additional prediction of our model for the folding of ribonuclease A is that α_2 should increase again as the pH approaches the p*K* of histidine since the favorable His···Glu electrostatic interaction in the competing pocket would then be lost.

Using the temperature dependence[31] of $\Delta G^{\mathrm{H}\phi}$ (and taking ΔG^{Tr} independent of temperature because of lack of available data), the value of ΔG_{nuc} for ribonuclease A was calculated at different temperatures from 15 to 67 °C, with eq 3. The nucleating pocket was 106–118 at all temperatures, and α_2 changed only slightly. This accounts for the observed[59] temperature independence of α_2.

The unfolding/folding of ribonuclease A has also been followed by fast kinetic measurements in which unfolding was produced by guanidine hydrochloride or urea[51] When either of these denaturants was present in the initial system, α_2 became a pH independent constant of about 0.2, a number identical to the observed[51] low pH value in the absence of denaturants. This can also be explained by our model. Urea and guanidine hydrochloride influence the thermodynamic activity of residues in proteins by both "hydrophobic" and "nonhydrophobic" effects.[60] The latter involve interactions with the amide backbone of the protein and should affect all denatured forms of the same protein to about the same extent. The "hydrophobic" effect clearly must depend on the types of side chains affected by the denaturants. However, in our model, the factor which is responsible for determining α_2 is the competition among various unfolded species which differ only in the structure (into pockets) of the *same* region of the chain in which the nucleation site appears. Thus, the hydrophobic effect will also be *nearly* the same in all conformations other than the "completely" random coil whose mutual competition determines α_2. Since electrostatic effects are significant only when they occur in pockets already stabilized by nonpolar interactions (see assumption 3 in section II), any destabilization of these nonpolar interactions that leads to "loss" of a pocket will lead to suppression of electrostatic effects according to our model. For the case of ribonculease A, the pH dependence of α_2 arises from the interaction of charged residues in pockets with $\Delta G_{\mathrm{nuc}} \sim 4$–6 kcal/mol relative to the "completely" random coil. The experimental data can be accounted for if we assume that all pockets (including the nucleating one of residues 106–118) are destabilized by urea or guanidine hydrochloride by the same amount, viz., *RT*. Thus, the competing pockets are too weak for an electrostatic effect to be included, in our model; hence, the model predicts this pH independent value of α_2 to be 0.55. A more rigorous quantitative analysis does not seem justifiable, since at this point it would require the introduction of adjustable parameters characterizing the positions and affinities of denaturant binding sites and these topics are not well understood quantitatively.[61]

In summary, the prediction of residues 106–118 as the nucleation site for folding of ribonuclease A (with disulfides intact) is consistent with a broad variety of experimental evidence. Our model provides qualitative explanations for the kinetic parameters observed in fast kinetic studies.

For reduced ribonuclease A, refolding studies have examined the order of reformation of the disulfide bonds.[62–64] All of these studies concur that there is little initial preference for one particular disulfide pairing to dominate. Creighton[64] also demonstrated that the rate of formation of disulfide bonds after the first is not faster than the initial cross-linking. Our model is not inconsistent with these results. These studies do point out the importance of S–S bond formation in determining the rate of refolding for proteins with reduced cystines, although the pathway may not be altered from that for folding of the same proteins with intact disulfides. Species with incorrect pairings must wait for a free SH to aid in breaking wrong disulfide bonds and permit S–S reshuffling and refolding. Under the conditions used for the renaturation of ribonuclease A (pH >7, ca. 25 °C), the value of α_2 is very small (calculated to be less than 0.1). Also, since the enthalpy of transfer for the side chains of this pocket is so large (16 kcal/mol), the activation energy for nucleation will be large and nucleation correspondingly slow. Thus, it is not at all unreasonable that incorrect cysteine pairings can occur and, once mispaired, their reshuffling becomes rate limiting. In short, incorrect disulfide pairing would not occur if α_2 were large, or if α_2 were small and nucleation were very rapid, since, in either case, the passage from nucleated to native structure would be rapid. From the observation[62-64] that incorrect disulfide pairing does occur, we conclude that α_2 is small and nucleation is slow (due to the large computed activation energy for nucleation); i.e., our nucleation model is not inconsistent with the observation of incorrect disulfide pairing. Because of the high activation energy for nucleation, and random fluctuations in the ensemble of unfolded molecules, incorrect disulfide bonds can form prior to nucleation. Once wrong disulfide bonds are formed, the "normal" folding pathway is blocked. Instead, the rate of breaking and reshuffling of wrong disulfides (with an attendant formation of proper nuclei) dominates in determining the rate of attainment of the native structure. Although there is an indication[63] that the slightly preferred initial cysteine pairing is Cys 65–Cys 72, this is not necessarily evidence for a nucleation site in this region of the chain, since disulfide pairings between residues close in the primary sequence are expected to be favored statistically.[65]

B. Chymotrypsinogen and Trypsinogen. The nucleation site for chymotrypsinogen A or α-chymotrypsin is predicted to be the 11 residue segment from 206–216 (using the sequence numbering scheme appropriate to chymotrypsinogen). The predicted value of α_2, 0.92, given in Table II, appears to be in significant disagreement with the observed value[7] of ca. 0.15 for the fraction of fast folding species. The discrepancy may arise from the fact that only about two of the eight tryptophan residues participate in the folding at pH 2.[66] Since two Trp residues occur in the calculated nucleation site, and the absorbance of Trp was used to follow unfolding,[7,67,68] the spectroscopic properties of different species may differ, and thus introduce ambiguity in the comparison of our predicted value of α_2 with the experimental value.

For trypsinogen (or trypsin), the most favorable values of ΔG_{nuc} are found for the sequence Ile-Met-Leu-Ile, residues 91–94, and for the 11-residue sequence from residues 212–222.

For both the zymogens and proteases (with disulfides intact), τ_{slow} (25 °C) and E_{A} (low temperature) are of similar magnitudes[67,68] to those found for ribonuclease A, as would be expected for proteins with disulfide cross-links that span the nucleation site. No pH dependence of α_2 is to be expected for either chymotrypsinogen or α-chymotrypsin and only a weak pH dependence is predicted for trypsinogen (or trypsin).

C. Lysozyme and α-Lactalbumin. These two proteins exhibit exceptional sequence homology,[69] and this has formed the basis for calculation[43] of the structure of α-lactalbumin by assuming a structural homology with lysozyme. However, while lysozyme is predicted to have two equally favorable nucleation sites (54–64 and 55–65), α-lactalbumin is predicted to have only one (89–92) which is not homologous to either of those in lysozyme. The 89–92 site in α-lactalbumin is located in an α-helical sequence, and the other two predicted nucleation sites contain β bends. Not many sequences of lysozyme or α-lactalbumin are available, but of those[35] which are we observe only conservative substitutions from species to species.

Studies of the folding of lysozyme often reveal only a single kinetic phase[70] but two stages can be observed.[8,11,71] Our model leads to the expectation of biphasic kinetics typical of other proteins with disulfide cross-links in the nucleation site.· α_2 is predicted to be 0.09, but experimental values with which to compare it are not available.

α-Lactalbumin has been studied extensively by the Hokkaido University group.[72,73] They propose a three-stage pathway with a partially helical state (the A state) as an intermediate. Our model is not inconsistent with their proposals since they also find evidence for "local hydrophobic interactions"[73] in the A state, and long-range interactions which may originate in our nucleation site are known to stabilize helices.[74]

D. Non-Cross-Linked Proteins (Reduced Pancreatic Trypsin Inhibitor, Cytochrome *c*, Parvalbumin, and Staphylococcal Nuclease). These proteins without cross-links (the heme of cytochrome *c* bridges only two residues) may also nucleate by formation of hydrophobic pockets. Pancreatic trypsin inhibitor and staphylococcal nuclease possess β bends within the predicted nucleation sites (see Table II) but the nucleation site for parvalbumin is in an α helix. Sequence conservation is illustrated strikingly by cytochrome *c*. The cytochrome *c*'s of 28 species show total conservation of 3 of the 4 positions predicted to constitute the nucleation site and, at the other position (residue 81), nonpolar character is always preserved.

The proteins without disulfide bonds show the wide variation in the values of τ_{slow} discussed in section IIIC. While we cannot account for the observed values of τ_{slow} quantitatively, we can use eq 5 to give a relative ordering. In addition, if two proteins differ only in the factor A (see eq 5), we would expect their respective values of τ_{slow} to be of similar magnitude, whereas variation in E_{A} would produce large differences. Using these assumptions, calculations with eq 5 lead to the expectation that the values of τ_{slow} for these proteins should be in the order: parvalbumin $\lesssim$ staphylococcal nuclease $\lesssim$ cytochrome *c* $\ll$ pancreatic trypsin inhibitor. This is in accord with experiment[47,50,52] except for pancreatic trypsin inhibitor which might be expected to differ because of the uncertainties involving species with incorrect disulfides.[75]

The predictions of α_2 (see Table II) are in reasonable agreement with the experimental value of 0.25 for parvalbumin[52] but not that of 0.5 for cytochrome *c*.[50] Kinetic experiments[76] on the unfolding of cytochrome *c* have shown that deliganding of the heme group can be associated with both a very fast (ca. 10–20 ms) and a slightly slower (ca. 100–200 ms) phase. Thus, interactions of the polypeptide chain with the prosthetic group in cytochrome *c* are known to be important for cytochrome *c* kinetics and provide a plausible explanation for this discrepancy. Pancreatic trypsin inhibitor presents a somewhat confused picture since it is not clear how to correlate α_2 with the order or relative amount of disulfide bond formation.

The structures of fragments of staphylococcal nuclease have been studied with immunological techniques by Sachs et al.[77] They have observed that addition of the fragment 6–43 (which contains the predicted nucleation sites) to solutions of other fragments causes a sharp increase in the amount of native structure at the antigenic determinants of those other fragments. Fragment 6–48 was itself examined for native structure[78] and little was found. Unfortunately, the probe in this case was circular dichroism[78] which should be quite insensitive to pocket formation, especially in view of the fact that no helices occur in the native structure of 6–48.[79] Also in connection with staphylococcal nuclease, it is interesting to note that the predicted nucleation sites are nearly coincident with the β turns which permit the formation of an extensive region of β sheets in the native structure.[79]

E. Other Proteins (Myoglobin, Hemoglobin, Proinsulin, Elastase, and Carboxypeptidase A). The nucleation sites predicted for these proteins have been included in order to make predictions on proteins whose folding behavior has been studied little if at all. The first three show exceptional sequence conservation for the limited number of species for which sequences are available.[35]

V. Discussion

Since we have not demonstrated conclusively that our model of the nucleation process is completely accurate, it should be asked to what extent it accords with other ideas about nucleation, what are its likely deficiencies, and what new experiments does it suggest. Many of the successes of our model in rationalizing the results of fast kinetic experiments are shared by the proline-isomerization model proposed by Brandts et al.[29] In this section, we discuss the general aspects of our model and, in section VI, we compare it to that of Brandts et al.

The model presented in this paper is based on the view that unfolded proteins are ensembles of molecules with different amounts of local structure and that hydrophobic bonding is the dominant force stabilizing these transient local structures. It is generally recognized that, even though the total disruption of small regions of local structure is rarely if ever achieved in protein denaturation,[61] the origin of the local stability is unknown. In section II, we have tried to justify our specification of hydrophobic bonding as this source of stability. The calculations performed on specific proteins have shown that the order of magnitude of this stability is quite reasonable, ΔG_{nuc} being typically of the order of ±1 or 2 kcal/mol (relative to the random chain with only neighbor-neighbor hydrophobic bonds) for the most stable pockets. Moreover, the flexible and nondirectional nature of hydrophobic bonding make it a plausible source of stability for structures that can readily interchange with one another, such as the competing pockets which determine α_2. The size of the nucleating pockets predicted by our model (ca. 10 residues) is near that which other authors[80] have believed to be reasonable.

Tanaka and Scheraga have proposed a three-step model of protein folding.[18,25] In it, contacts are formed first among residues close together in the primary sequence (step A), and then these regions coalesce (step B). Interactions in step A are considered to be dominant, but may proceed simultaneously with those of step B. The model presented here differs in emphasis, but not in essentials, from this view. Neighbor-neighbor hydrophobic bonds (interactions of the type that occur in step A) are allowed for in the "completely" unfolded chain and their influence is to decrease the amount of conformational entropy which is lost due to formation of a pocket. Thus, as in the Tanaka–Scheraga model, medium-range interactions influence the attainment of structures involving contacts of more distant residues. However, in our model, these medium-range interactions of type A do not dominate and need not even persist into step B. The sizes of pockets which are nucleation sites in our model are comparable to the sizes of contact regions arising in step B. Thus, rather than viewing step B as a more or less logical extension of step A (as in the Tanaka–Scheraga picture), the present model views nucleation (i.e., a particular conformational change of step B) as a qualitatively distinct step, affected only slightly by interactions of step A. Additionally, the present model is based on the assumption that not all types of contacts which can be classified as type A or B need to be considered in order to understand nucleation; only nonpolar contacts are important. The Tanaka–Scheraga model does not make this restriction specifically, although hydrophobic bonds could lead to the same result in the Tanaka–Scheraga model.

It should be emphasized that the segment of a polypeptide chain which has the lowest value of ΔG_{nuc} does not necessarily have the native conformation nor does it necessarily have the structure of lowest *total* free energy for those particular residues. As already stated, the nucleating pocket can change its conformation during the subsequent stages of folding to form the native protein. Further, since the nucleating pocket (selected on the basis of the most favorable hydrophobic interaction) can rearrange in a subsequent stage of folding, then additional interactions (not considered in the calculation of ΔG_{nuc}) could lower the free energy of the given (nucleating) segment below ΔG_{nuc}. The method proposed here assumes that, in the early stages of folding, hydrophobic bonding is the only type of interaction that can compete effectively with the entropy that stabilizes the unfolded state.

Gō[26] has suggested that significant intermediates should be observable in studies of protein folding when large blocks of predominantly nonpolar residues exist in the protein sequence. Calculations of ΔG_{nuc} with our model (see Table I) indicate that the most likely proteins in which to observe such intermediates are carboxypeptidase and chymotrypsinogen A. To our knowledge, no search for intermediates in these proteins has been carried out. Studies of the pressure denaturation of lysozyme[11] also indicate that an intermediate may exist on the unfolding pathway. If so, then our model predicts that an intermediate should also be observable with all proteins that have values of ΔG_{nuc} that are more negative than that for lysozyme. While other proteins (notably staphylococcal nuclease) must show at least biphasic kinetics, they are predicted to be poor choices in which to detect intermediates. This is because of the low stability of their most nonpolar sections.

Our model does have some obvious deficiencies at this initial stage. First is the use of three-fourths of the total free energy of transfer to approximate the stability gained by forming large pockets of nonpolar residues. The defect in this procedure is evident from the fast kinetic studies of ribonuclease A; i.e., Tyr 115 is predicted to be buried in the nucleating pocket but exposed in many of the slow folding species. Since the pK of a nitrotyrosine derivative shifts only slightly, and no spectroscopic differences are observed between the fast and slow refolding forms, it is not likely that such an appreciable burial of Tyr 115 occurs. In general, all of the parameters used in calculations with eq 3 could probably be improved by further investigations. A second deficiency is the inability of our model, in its present form, to predict the values of τ_{slow} quantitatively. Success of such predictions would constitute persuasive evidence that the model is accurate and general. A third shortcoming lies in the tacit assumption that each protein has but a single nucleation site, although four overlapping sites have the same values of ΔG_{nuc} in staphylococcal nuclease and two in two other proteins. There probably is more than one nucleation site per polypeptide chain, but it is difficult to predict the additional ones at the present stage of development of our model.

In spite of these and other shortcomings, our model is useful in that it indicates a number of directions for future experimental studies. One such direction, indicated previously, would be a search for stable intermediates in the folding of chymotrypsinogen A or carboxypeptidase. A second would be studies of folding of derivatized proteins where changes are produced in the predicted nucleation sites. A third would be experiments directed toward the acquisition of a more complete understanding of the location of denaturant binding sites to see whether these are also nucleation sites. A fourth direction would involve experiments designed to gain an understanding of the molecular basis of the rapid rate of folding after the nucleation step. Experiments along these lines are suggested by the qualitative features of the model, which have been rationalized and partially confirmed by this present work. Quantitative refinement of the details of the model will be necessary before more subtle experiments to test the predictions of the model need be undertaken. However, it should be emphasized that it is only the *relative* (rather than the *absolute*) values of $\Delta G_{\rm nuc}$ that have any reliability.

VI. Comparison with the Hypothesis of Cis-Trans Proline Isomerization

Brandts et al.[29] have proposed that the random cis–trans isomerization of proline residues is the basis of the slow step of protein folding. Conceptually, this is the same as proposing that proline isomerization constitutes the nucleation step (see definition, section I). This proposal can account[29] adequately for the observed fractions of fast-folding molecules in a variety of systems (except LiBr-denatured lysozyme[29] and guanidine-denatured ribonuclease A[29]), their temperature insensitivity, the order of magnitude of $\tau_{\rm slow}$, and the activation energy of the slow step in protein folding (nearly always ~20 kcal/mol for the cases studied to date, all of which involved cross-linked proteins). However, it cannot explain the pH dependence of the fraction of fast-folding ribonuclease A.

The model proposed in this paper approximately accounts for the fractions of fast-folding molecules in a number of systems including guanidine-denatured ribonuclease A, but not cytochrome *c* (where α_2 might be influenced by the heme group). It explains the temperature insensitivity of this quantity, its pH dependence for the case of ribonuclease A, and provides a plausible, although far from impressive, rationalization for the constancy of the observed magnitudes of $\tau_{\rm slow}$ for proteins with intact disulfide bonds. Additionally, the predictions of this model are consistent with, and so help to understand, many experimental facts about protein folding other than fast kinetic results.

Thus, except for the activation energy of its slow step, protein folding seems better treated by our model than by the proline-isomerization approach. The slow-step activation energy (for cross-linked proteins) can be understood qualitatively within the framework of our model but is not yet predicted quantitatively (see section IIIC). We expect that all appropriately cross-linked proteins should have roughly equivalent energies of activation for nucleation in accord with observation.[29]

Finally, a recent paper by Nall et al.[81] has cast doubt on the validity of the proline-isomerization hypothesis for the folding of ribonuclease A.

VII. Conclusions

The method of predicting possible nucleation sites based on considerations of long-range hydrophobic bonding leads to good agreement with experiment. It explains many aspects of protein folding in molecular terms and is of very general applicability. It can be refined further by improved parameterization. The important assumptions are that the hydrophobic bonding of nonpolar residues dominates over other noncovalent interactions in nucleation and that hydrophobic bonding must compensate for the entropy loss in forming a pocket. This procedure for predicting possible nucleation sites may serve as an additional component of a protein folding algorithm.

Acknowledgment. We are indebted to Drs. J. S. Anderson, L. G. Dunfield, E. L. Elson, F. R. Maxfield, G. Nemethy, and R. K. Scheule for helpful discussions of this work and Dr. H. Morawetz for a helpful discussion of correlated motions in conformational changes.

The computer program, and associated parameters, for calculating $\Delta G_{\rm nuc}$ from eq. 3 can be found in the Ph.D. thesis of R. R. Matheson, Jr. (Cornell University, August 1978).

References and Notes

(1) This work was supported by research grants from the National Science Foundation (PCM75-08691) and from the National Institute of General Medical Sciences of the National Institutes of Health, U.S. Public Health Service (GM-14312).

(2) NIH Predoctoral Trainee, 1974–1978.

(3) C. B. Anfinsen and H. A. Scheraga, *Adv. Protein Chem.*, **29**, 205 (1975).

(4) R. L. Baldwin, *Annu. Rev. Biochem.*, **44**, 453 (1975).

(5) G. Némethy and H. A. Scheraga, *Q. Rev. Biophys.*, **10**, 239 (1977).

(6) D. C. Poland and H. A. Scheraga, *Biopolymers*, **3**, 401 (1965).

(7) T. Y. Tsong and R. L. Baldwin, *J. Mol. Biol.*, **69**, 145 (1972).

(8) A. Ikai, W. W. Fish, and C. Tanford, *J. Mol. Biol.*, **73**, 165 (1973).

(9) P. J. Hagerman and R. L. Baldwin, *Biochemistry*, **15**, 1462 (1976).

(10) T. Y. Tsong, *Biochemistry*, **15**, 5467 (1976).

(11) T. M. Li, J. W. Hook, III, H. G. Drickamer, and G. Weber, *Biochemistry*, **15**, 5571 (1976).

(12) J. F. Brandts and L. Hunt, *J. Am. Chem. Soc.*, **89**, 4826 (1967).

(13) P. L. Privalov and N. N. Khechinashvili, *J. Mol. Biol.*, **86**, 665 (1974).

(14) A. W. Burgess and H. A. Scheraga, *J. Theor. Biol.*, **53**, 403 (1975).

(15) A. W. Burgess and H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.A.*, **72**, 1221 (1975).

(16) M. Levitt and A. Warshel, *Nature (London)*, **253**, 694 (1975).

(17) O. B. Ptitsyn and A. A. Rashin, *Biophys. Chem.*, **3**, 1 (1975).

(18) S. Tanaka and H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.A.*, **72**, 3802 (1975); **74**, 1320 (1977).

(19) M. Levitt, *J. Mol. Biol.*, **104**, 59 (1976).

(20) I. D. Kuntz, G. M. Crippen, P. A. Kollman, and D. Kimelman, *J. Mol. Biol.*, **106**, 983 (1976).

(21) A. Warshel and M. Levitt, *J. Mol. Biol.*, **106**, 421 (1976).

(22) G. M. Crippen, *Macromolecules*, **10**, 21, 25 (1977).

(23) A. Ikegami, *Biophys. Chem.*, **6**, 117 (1977).

(24) M. I. Kanehisa and A. Ikegami, *Biophys. Chem.*, **6**, 131 (1977).

(25) S. Tanaka and H. A. Scheraga, *Macromolecules*, **10**, 291 (1977).

(26) N. Gō, *Int. J. Pept. Protein Res.*, **7**, 313 (1975).

(27) S. Tanaka and H. A. Scheraga, *Macromolecules*, **9**, 142, 159, 168, 812 (1976); **10**, 9, 305 (1977).

(28) T. Y. Tsong, R. L. Baldwin, and P. McPhie, *J. Mol. Biol.*, **63**, 453 (1972).

(29) J. F. Brandts, H. R. Halvorson, and M. Brennan, *Biochemistry*, **14**, 4953 (1975).

(30) W. Kauzmann, *Adv. Protein Chem.*, **14**, 1 (1959).

(31) G. Némethy and H. A. Scheraga, *J. Phys. Chem.*, **66**, 1773 (1962).

(32) J. F. Brandts in "Structure and Stability of Biological Macromolecules", Vol. 2, S. N. Timasheff and G. D. Fasman, Ed., Marcel Dekker, New York, N.Y., 1969, p. 213.

(33) C. Tanford, *Adv. Protein Chem.*, **24**, 1 (1970).

(34) D. C. Poland and H. A. Scheraga, *Biopolymers*, **3**, 283, 315, 335 (1965).

(35) M. O. Dayhoff, "Atlas of Protein Sequence and Structure", Vol. 5, National Biomedical Research Foundation, Silver Spring, Md., 1972.

(36) S. J. Leach, G. Némethy, and H. A. Scheraga, *Biopolymers*, **4**, 369 (1966).

(37) S. S. Zimmerman and H. A. Scheraga, *Biopolymers*, **16**, 811 (1977).

(38) Y. Nozaki and C. Tanford, *J. Biol. Chem.*, **246**, 2211 (1971).

(39) M. Levitt, *J. Mol. Biol.*, **104**, 59 (1976).

(40) J. A. Schellman, *C. R. Trav. Lab. Carlsberg, Ser. Chim.*, **29**, 230 (1955).

(41) P. J. Flory, *J. Am. Chem. Soc.*, **78**, 5222 (1956).

(42) D. C. Poland and H. A. Scheraga, *Biopolymers*, **3**, 379 (1965).

(43) P. K. Warme, F. A. Momany, S. V. Rumball, R. W. Tuttle, and H. A. Scheraga, *Biochemistry*, **13**, 768 (1974).

(44) T. Blundell, G. Dodson, D. Hodgkin, and D. Mercola, *Adv. Protein Chem.*, **26**, 279 (1972).

(45) J. A. Lenstra, J. Hofsteenge, and J. J. Beintema, *J. Mol. Biol.*, **109**, 185 (1977).

(46) T. Y. Tsong, R. L. Baldwin, and E. L. Elson, *Proc. Natl. Acad. Sci. U.S.A.*, **68**, 2712 (1971).

(47) H. F. Epstein, A. N. Schechter, R. F. Chen, and C. B. Anfinsen, *J. Mol. Biol.*, **60**, 499 (1971).

(48) T. Y. Tsong, R. L. Baldwin, and E. L. Elson, *Proc. Natl. Acad. Sci. U.S.A.*, **69**, 1809 (1972).

(49) L. L. Shen and J. Hermans, Jr., *Biochemistry*, **11**, 1836 (1972).

(50) T. Y. Tsong, *Biochemistry*, **12**, 2209 (1973).

(51) J. R. Garel, B. T. Nall, and R. L. Baldwin, *Proc. Natl. Acad. Sci. U.S.A.*, **73**, 1853 (1976).

(52) J. F. Brandts, M. Brennan, and L. N. Lin, *Proc. Natl. Acad. Sci. U.S.A.*, **74**, 4178 (1977).

(53) H. Morawetz, *Adv. Protein Chem.*, **26**, 243 (1972).

(54) D. T. L. Chen and H. Morawetz, *Macromolecules*, **9**, 463 (1976).

(55) W. H. Stockmayer, *Pure Appl. Chem.*, **15**, 539 (1967).

(56) H. W. Wyckoff, D. Tsernoglou, A. W. Hanson, J. R. Knox, B. Lee, and F. M. Richards, *J. Biol. Chem.*, **245**, 305 (1970).

(57) L. G. Chavez, Jr., and H. A. Scheraga, *Biochemistry*, **16**, 1849 (1977).

(58) J.-R. Garel and R. L. Baldwin, *J. Mol. Biol.* **94**, 621 (1975).

(59) J.-R. Garel and R. L. Baldwin, *J. Mol. Biol.*, **94**, 611 (1975).

(60) D. R. Robinson and W. P. Jencks, *J. Am. Chem. Soc.*, **87**, 2462 (1965).

(61) C. Tanford, *Adv. Protein Chem.*, **23**, 121 (1968).

(62) R. R. Hantgan, G. G. Hammes, and H. A. Scheraga, *Biochemistry*, **13**, 3421 (1974).

(63) S. Takahashi and T. Ooi, *Bull. Inst. Chem. Res., Kyoto Univ.*, **54**, 141 (1976).

(64) T. E. Creighton, *J. Mol. Biol.*, **113**, 329 (1977).

(65) W. Kauzmann, "Symposium of Sulfur in Proteins", Academic Press, New York, 1959, pp 93–104.
(66) B. Havsteen, B. Labouesse, and G. P. Hess, *J. Am. Chem. Soc.*, **85**, 796 (1963).
(67) F. M. Pohl, *FEBS Lett.*, **3**, 60 (1969).
(68) F. Pohl, Ph.D. Thesis, University of Konstanz, Konstanz, Germany, 1970, cited in Brandts et al.[29]
(69) K. Brew, T. C. Vanaman, and R. L. Hill, *J. Biol. Chem.*, **242**, 3747 (1967).
(70) S. Segawa, Y. Husimi, and A. Wada, *Biopolymers*, **12**, 2521 (1973).
(71) C. Tanford, K. C. Aune, and A. Ikai, *J. Mol. Biol.*, **73**, 185 (1973).
(72) K. Kuwajima, K. Nitta, M. Yoneyama, and S. Sugai, *J. Mol. Biol.*, **106**, 359 (1976).
(73) K. Kuwajima, *J. Mol. Biol.*, **114**, 241 (1977).
(74) R. M. Epand and H. A. Scheraga, *Biochemistry*, **7**, 2864 (1968).
(75) T. E. Creighton, *J. Mol. Biol.*, **113**, 295 (1977).
(76) T. G. Bruckman, Ph.D. Thesis, Cornell University, 1977.
(77) D. H. Sachs, A. N. Schechter, A. Eastlake, and C. B. Anfinsen, *Proc. Natl. Acad. Sci. U.S.A.*, **69**, 3790 (1972).
(78) H. Taniuchi, C. B. Anfinsen, and A. Sodjo, *Proc. Natl. Acad. Sci. U.S.A.*, **58**, 1235 (1967).
(79) E. E. Hazen, Jr., D. C. Richardson, J. S. Richardson, and in part A. Yonath, *J. Biol. Chem.*, **246**, 2302 (1971).
(80) D. B. Wetlaufer and S. Ristow, *Annu. Rev. Biochem.*, **42**, 135 (1973).
(81) B. T. Nall, J. R. Garel, and R. L. Baldwin, *J. Mol. Biol.*, **118**, 317 (1978).

Polymer Effects under Pressure. 2. Enzymelike Catalysis in the Hydrolysis of Phenyl Ester by Copolymer Containing Imidazole[1]

Yoshihiro Taniguchi,* Kiyokazu Shimokawa, Hideo Hisatome, Shyoichi Tanamachi, and Keizo Suzuki

Department of Chemistry, Faculty of Science and Engineering, Ritsumeikan University, Kita-ku, Kyoto, 603, Japan. *Received July 29, 1977*

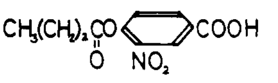
ABSTRACT: The rates of hydrolysis of 3-nitro-4-butyryloxybenzoic acid (NBBA) and 3-nitro-4-pentanyloxybenzoic acid (NPeBA) catalyzed by the copolymer of 1-vinyl-2-methylimidazole with 1-vinylpyrrolidone (MI-VP) were measured at pressures up to 1471 bar at 30 °C pH 8.0 in 0.068 M in Veronal buffer solution. The reaction was found to follow Michaelis–Menten kinetics. The Michaelis constant K_m was estimated to vary from 6.0 to 10 mM and $k_{cat.}$ from 0.024 to 0.059 min^{-1} between 1 and 1373 bar for NBBA, and K_m from 5.4 to 12 mM and $k_{cat.}$ from 0.022 to 0.077 min^{-1} between 1 and 1471 bar for NPeBA. From the pressure dependence of K_m and $k_{cat.}$, the volume changes accompanying the dissociation of the polymer–substrate complex and the activation volume of the process of the product formation were calculated to be -8.4 ± 2 and -20 ± 2 cm^3/mol for NBBA and -12 ± 2 and -20 ± 2 cm^3/mol for NPeBA, respectively. These negative values for the dissociation of the polymer–substrate complex show that hydrophobic interactions stabilize the complex. The negative activation volumes for the process of product formation may be attributed either to polarity increases in the transition state of the acylation or to bimolecular attack of water on the alkylimidazolium in the deacylation step.

In studying the pressure effect on enzyme reactions, it is desirable for the experiments to be carried out over a range of substrate concentration in order to separate the effects of pressure on the dissociation constants of the Michaelis complex, K_m, and the rate constants of the product formation, $k_{cat.}$. In 1955 the effects of pressure on the kinetics of the myosin-catalyzed hydrolysis of adenosine triphosphate were studied by Laidler and Beardell.[2] Later, similar studies were carried out by Andersen and Broe[3] (interconversion of fumarate to L-malate by fumarase), Mohankumar and Berger[4] (hydrolysis of *p*-nitrophenyl phosphate catalyzed by bovine alkali phosphotase), Williams and Shen[5] (hydrolysis of cytidine 2′,3′-phosphate catalyzed by ribonuclease), and Neville and Eyring[6] (dried *Micrococcus luteus* cell catalyzed by lysozyme).

Kunitake et al.[7] have studied the hydrolysis of phenyl esters catalyzed by a copolymer in which 1-vinyl-2-methylimidazole (MI) residues play the role of the catalytic function and 1-vinylpyrrolidone (VP) residues the role of a hydrophobic binding function. According to them, the rate of the catalytic hydrolysis of the polymer containing low imidazole residues below 20 mol % is simply described by Michaelis–Menten kinetics. The use of such a copolymer for the study of the reaction rates under high pressure is advantageous since the enzymelike catalyst is free of the pressure inactivation found in enzymatically active proteins.[8]

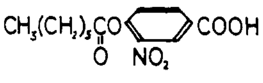
In the present paper, the hydrolysis of NBBA **1** and NPeBA **2** catalyzed by the MI-VP **3** copolymer has been measured at 30 °C and pressures up to 1471 bar in order to obtain the pressure dependences of K_m and $k_{cat.}$. The reaction mechanism is discussed in terms of the volume change accompanying Michaelis–Menten kinetics.

1



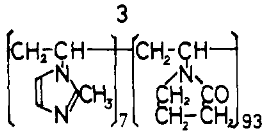
NBBA 1

2



NPeBA 2

3



MI-VP 3

Experimental Section

Copolymer. 1-Vinylpyrrolidone, commercial product, was purified by distillation before use, bp 75.5–76.5 °C (6 mm). 1-Vinyl-2-methylimidazole was provided by Shikoku Kasei Co., Japan, and distilled twice before use, bp 69.5–72.5 °C (9 mm). The copolymers of 1-vinyl-2-methylimidazole and 1-vinylpyrrolidone were obtained by radical polymerization of bulk monomer mixtures at 70 °C for 30 min using azobis(isobutyronitrile) as an initiator. The reaction mixture was diluted with methanol and the copolymer was precipitated by excess ethyl ether. The copolymer was reprecipitated twice then dried in vacuo. The molecular weight was measured with a vapor pressure osmometer (Hitachi Perkin-Elmer type 115) to be 10500. The copolymer composition, determined by the titration of the imidazole (Im) groups, contained 7.2 mol % Im.

Substrates. 3-Nitro-4-hydroxybenzoic acid (NHBA), mp 182–184 °C, which was obtained by nitration of *p*-hydroxybenzoic acid, was

0024-9297/78/2211-0829$01.00/0