Chemical Frustration in the Protein Folding Landscape: Grand Canonical Ensemble Simulations of Cytochrome *c*[†]

Patrick Weinkam,[‡] Floyd E. Romesberg,[§] and Peter G. Wolynes*[,‡,||]

Department of Chemistry, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, Center for Theoretical Biological Physics and Department of Chemistry and Biochemistry, University of California at San Diego, 9500 Gilman Drive, La Jolla, California 92093, and Department of Physics, University of California at San Diego, 9500 Gilman Drive, La Jolla, California 92093

Received December 18, 2008; Revised Manuscript Received February 5, 2009

ABSTRACT: A grand canonical formalism is developed to combine discrete simulations for chemically distinct species in equilibrium. Each simulation is based on a perturbed funneled landscape. The formalism is illustrated using the alkaline-induced transitions of cytochrome *c* as observed by FTIR spectroscopy and with various other experimental approaches. The grand canonical simulation method accounts for the acid/base chemistry of deprotonation, the inorganic chemistry of heme ligation and misligation, and the minimally frustrated folding energy landscape, thus elucidating the physics of protein folding involved with an acid/base titration of a protein. The formalism combines simulations for each of the relevant chemical species, varying by protonation and ligation states. In contrast to models based on perfectly funneled energy landscapes that contain only contacts found in the native structure, this study introduces "chemical frustration" from deprotonation and misligation that gives rise to many intermediates at alkaline pH. While the nature of these intermediates cannot be easily inferred from available experimental data, this study provides specific structural details of these intermediates, thus extending our understanding of how cytochrome *c* changes with an increase in pH. The results demonstrate the importance of chemical frustration for understanding biomolecular energy landscapes.

Specific chemistry often plays a dominant role in folding, especially in metalloproteins. Covalent chemistry is clearly discrete, while folding is so complex that it often appears to be nearly continuous. A general way to combine detailed coordination chemistry with energy landscape ideas describing the inter-residue forces guiding folding is to employ a grand canonical formalism. The formalism allows the use of discrete simulations using perturbed structure-based (native topology-based) models to examine how ensemble properties change with pH and metal coordination. Structure-based models, utilizing a perfectly funneled energy landscape, have already been used to predict the dominant folding route of cytochrome *c* which is strongly influenced by its heme cofactor (*1*) (Figure 1). These models have also been successfully applied to many other naturally occurring proteins without cofactors (*2*−*8*). However, changes in solvent conditions can perturb the energy landscape, giving rise to "chemical frustration" that allows chemically distinct intermediates to appear (Figure 2). The observation of non-native protein ligand interactions with the iron of cytochrome *c* provided one of the first experimentally documented examples of frustration on a biomolecular energy landscape (*9*). Especially beautiful examples of how the folding landscape may be modulated by ligation and misligation events are provided by the alkaline-induced transitions of cytochrome *c*. With an increase in pH, the native Met80 ligand is displaced from the heme center, resulting in misligation of the heme and eventual unfolding of the protein (*10*−*13*). Experiments that probe cytochrome *c* mutants with NMR (*13*−*15*), hydrogen exchange (*16*), EPR

FIGURE 1: Crystal structure of cytochrome *c* colored by the folding units: N- and C-terminal helices (blue), 60s helix (green), two-stranded β sheet (yellow), Met80 loop (red), and the omega loop (gray). H-Bonds included in the model are shown with blue lines.

[†] This work was supported by National Institutes of Health Grant 5R01 GM44557 and by the Center for Theoretical Biological Physics (PHY-0822283).

* To whom correspondence should be addressed. Phone: (858) 822-4825. Fax: (858) 822-4560. E-mail: pwolynes@ucsd.edu.

[‡] Center for Theoretical Biological Physics and Department of Chemistry and Biochemistry, University of California at San Diego.

[§] The Scripps Research Institute.

[||] Department of Physics, University of California at San Diego.

[1] Abbreviation: AMH, associative memory Hamiltonian; rmsd, root-mean-square deviation.

10.1021/bi802293m CCC: $40.75
Published on Web 02/06/2009

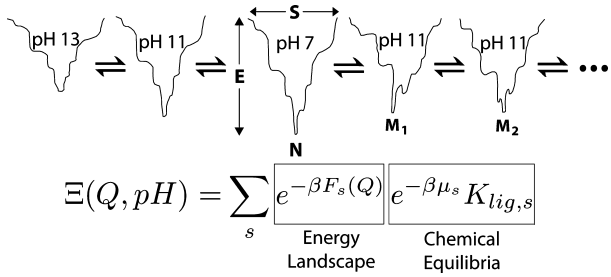

FIGURE 2: Scheme for combining energy landscapes for folding individual chemical species with acid/base and coordination chemistry. In the scheme, each chemical species is represented by its own folding funnel of varying depths. The pH 7 folding funnel has an energy minimum at the native state (N), but as the pH or ligation conditions change, the energy landscape is perturbed, giving rise to slightly different funnels with distinct minima such as at a misligated state (M_1 or M_2). The grand canonical partition function contains a sum over each chemical species (s). The folding landscape contributions are clearly separated from the terms containing chemical potentials that give rise to the chemical equilibria between species. For more information on the grand canonical partition function, see Methods.

(*15*), and resonance Raman (*17*) have identified Lys73 and Lys79 as ligands for the heme. However, at least under certain conditions, kinetic pH-jump experiments detect an intermediate in which the Met80 ligand is displaced and not replaced with a lysine (*18*). A more recent FTIR study (*19*) based on the incorporation of carbon−deuterium bonds at specific residues has identified a potentially related intermediate under equilibrium conditions (*19*). On the basis of these experiments, the equilibrium alkaline-induced transitions may be written as III ⇌ 3.5 ⇌ (IV_a ⇌ IV_b) ⇌ V ⇌ U. State III is the native state at pH 7; states IV_a and IV_b are misligated by Lys73 and Lys79, respectively, and U is the unfolded state. States 3.5 and V are less well characterized experimentally than the others. The identities of the heme ligands in states 3.5 and V have been suggested (*17, 19*), but this assignment remains tentative. Providing a detailed structural description of these states is the main goal of this paper.

To describe the different partially unfolded states induced by an increase in pH, chemical frustration must be introduced into what may be taken otherwise to be a perfectly funneled energy landscape to a good approximation. This frustration can be described by introducing a misligation potential (with an explicit heme) and by accounting for energetic changes due to deprotonation and accompanying electrostatics. Separate simulations are carried out for the relevant chemical species, which are selected by enumerating the probable ligation and protonation states. The simulations are used to generate free energy profiles that determine the stabilities of the different states in various degrees of folding. By knowing the chemical potential balances for the acid/base chemistry and for the heme coordination chemistry, we can combine the resulting set of free energy profiles using the statistical mechanics of the grand canonical ensemble to generate a rather complete description of the alkaline-induced transitions.

The grand canonical formalism allows us to model in a unified way a system which has multiple conformational transitions involving chemically distinct species. All atom simulations have previously been used to determine pH-dependent properties and pK_a values for relatively static conformations (*20*−*24*), but such expensive calculations are not currently feasible to apply to large conformational transitions like those involved in folding. Using a coarse-grained model, however, free energy profiles can efficiently be generated. The method allows us to separately analyze each individual species among a complex set of equilibria. The free energy profiles from different chemical species are combined by relating the free energies of completely unfolded ensembles, which in these models consist of structures that are near-random coils, to chemical equilibrium constants involving free amino acids. The free energies of these unfolded ensembles, before accounting for the chemical equilibria, should be the same since entropy dominates these ensembles and is nearly identical for unstructured chains of the same length. The difference in stability between the unfolded ensembles is determined by the relevant chemical potentials, such as for the hydronium ion in this study (Figure 2). One advantage of employing the grand canonical ensemble method is the flexibility to include or exclude specific chemical species. In this way, we can compare the effects of including different species on the properties of the entire ensemble to address experimental results that report on such intermediates, but without structural detail. This method also allows for approximations that can reduce the necessary simulation time by several orders of magnitude. For example, there are 2^{25} potential protonation states in the alkaline-induced transitions of cytochrome *c* if one enumerates all 25 basic residues. If we consider only the possible lysine-misligated states, then nearly 10^9 simulations would be needed. An approximation is appropriate given that the majority of those states are not individually observable experimentally. Therefore, we include the most probable protonation states (native pH 7, all lysine and tyrosine residues deprotonated, or all lysine and tyrosine and arginine residues deprotonated) and ligation states (native methionine, each lysine, or hydroxide bound). One of the main advantages of the current grand canonical approach is that it utilizes effectively only two free parameters. These correspond to the heme ligation constants which are furthermore restricted by experimental data. Apart from the heme coordination chemistry, the current method provides a simple energy landscape with essentially no adjustable parameters. In effect, the grand canonical ensemble method leads to results that are mostly dependent on features of the funneled part of the energy landscape which in turn is determined by the native state topology.

We therefore employ separate models to see which aspects of the energy landscape are most important for reproducing different experimental results. First, the most simple model, or pure-funnel model, consists of, with the exception of a misligation potential, interactions that are based entirely on the native structure of cytochrome *c*. In a separate model, we try to isolate the effects of long-range electrostatic and hydrophobic forces by including Coulomb and hydrophobic collapse potentials in addition to native topology-based terms. We explore the role of side chain hydrogen bonding (H-bonding) in yet another model. The use of these three models provides an interesting comparative study of the forces leading to the intermediate structures populated in the alkaline-induced transitions of cytochrome *c*.

Table 1: Simulations Performed and Parameters Used

simulated chemical species		
ligated chemical group	deprotonated residues	parameters
Met80	none	
Lys5	Lys, Tyr	pure-funnel and electrostatics/
Lys7	Lys, Tyr	collapse models
Lys8	Lys, Tyr	$pK_{Tyr} = 10.0$
Lys13	Lys, Tyr	$pK_{Lys} = 10.5$
Lys22	Lys, Tyr	$pK_{Arg} = 12.5$
Lys25	Lys, Tyr	$pK_{H_2O} = 15.6$
Lys27	Lys, Tyr	$K_{Lys} = 1 \times 10^{-4}$
Lys39	Lys, Tyr	$K_{OH} = 1 \times 10^{-7}$
Lys53	Lys, Tyr	$K_{Met} = K_{H_2O} = 1$
Lys55	Lys, Tyr	
Lys60	Lys, Tyr	
Lys72	Lys, Tyr	hydrogen bonding model
Lys73	Lys, Tyr	$pK_{Tyr} = 10.0$
Lys79	Lys, Tyr	$pK_{Lys} = 10.5$
Lys86	Lys, Tyr	$pK_{Arg} = 12.5$
Lys87	Lys, Tyr	$pK_{H_2O} = 15.6$
Lys88	Lys, Tyr	$K_{Lys} = 1 \times 10^{-4}$
Lys99	Lys, Tyr	$K_{OH} = 1 \times 10^{-9}$
Lys100	Lys, Tyr	$K_{Met} = K_{H_2O} = 1$
$^{-}$OH	Lys, Tyr	
$^{-}$OH	Lys, Tyr, Arg	

METHODS

The grand canonical ensemble is a general statistical mechanical tool that can be used to combine known chemical properties with results from simulation. Especially when dealing with large biomolecules, such a strategy can save tremendous amounts of computing time and in some cases allows calculations that would otherwise be impossible. Grand canonical ensemble simulations have been previously used to simplify the treatment of bulk water in calculations of solvation in proteins and DNA (*25, 26*). Similar methods have been used to improve the accuracy of calculations involving ligand−protein binding interactions (*27*).

Grand canonical ensemble simulations are used here to describe the alkaline-induced transitions of cytochrome *c*. The grand canoncial partition function is composed of three terms that describe the folding, the acid/base chemistry, and the heme coordination chemistry: $\Xi(Q,\mathrm{pH}) = \sum_s e^{-\beta F_s(Q)} e^{-\beta \mu_s} K_{lig,s}$. The subscript s runs over each chemical species. These have varying protonation and ligation states that occur at alkaline pH. $\beta = 1/k_B T$. In this study, we include 22 distinct species as shown in Table 1. The free energy profiles, $F_s(Q)$, that describe the folding and stability of each state are generated using the weighted histogram analysis method WHAM (*28*). Q is a folding reaction coordinate that describes the similarity to the native state topology and is defined as $Q = \{2/[(N-1)(N-2)]\} \sum_{i<j-1} \exp[-(r_{ij} - r_{ij}^{nat})^2/(2|i-j|^{0.3})]$. The summation runs over all C^α residues; r_{ij}^{nat} is the distance between C^α atoms in the crystal structure, and N refers to the number of residues. Deprotonation is described by the following chemical potential: $\mu_s = \sum_j k_B T(\mathrm{pH} - pK_{a,j}) N_j$. In this equation, j indicates three residues that are deprotonated (lysine, tyrosine, and arginine) and water, N_j equals 1 if the base is deprotonated within state s and 0 if it is not, and the pK_a is set to the value for the free group in solution (Table 1). Ligation is summarized in the dissociation constants, $K_{lig,s}$, for each of the potential heme ligands. The values of $K_{lig,s}$ are not individually known by experiment to a high degree of accuracy. Fortunately, only the ratios of the dissociation constants are important in the grand canonical method. We chose parameters, consistent with the principles of inorganic chemistry, such that heme is more strongly bound by hydroxide than by deprotonated lysine, which are, in turn, both bound more strongly than methionine or water. The dissociation constants are furthermore restricted by resonance Raman (*29*) and FTIR experiments (*19*) which suggest that, for example, two hydroxide-bound states dominate under highly alkaline conditions with a midpoint between pH 10.5 and 11. Varying $K_{OH,s}$ merely shifts the transition midpoint and does not change the structural details or the identity of the intermediates involved. Varying any dissociation constant by <1 order of magnitude does not significantly change the results.

These simulations are based on the associative memory Hamiltonian (AMH)[1] (*30, 31*) which utilizes structural information in multiple "memory" proteins which is used for ab initio structure prediction. By using only a single memory protein conforming to the native X-ray structure, we ensure a perfect funnel. The structure of horse heart cytochrome *c* is used for the simulations in this study. The AMH is written as the sum of backbone and contact terms. The backbone potential, V_{back}, contains only local terms that maintain backbone geometry (bond angles and distances) and includes excluded volume. The contact potential, $V_{contact}$, includes Gaussian well potentials summed over all C^α atoms and C^β atoms for protein−protein contacts or C^α atoms and heme pseudoatoms for heme−protein contacts. The contact potential includes 40% nonadditivity, which results in barrier heights and cooperativity consistent with those observed in small globular proteins (*32*). The details of the backbone and contact potentials were laid out in our earlier studies (*33*) and are also discussed in the Supporting Information. The heme is introduced explicitly with a potential V_{heme} that includes terms for maintaining the heme geometry as well as ligation terms for the iron to protein residues. The current heme model contains five heme pseudoatoms. Four atoms are fixed in a square-planar geometry, while the fifth atom is placed in the center. Well-chosen interactions between the pseudoatoms of the heme ensure that the heme has a realistic size and shape. Harmonic potentials, $k(r - r_o)^2$, account for ligation between the heme center pseudoatom and the C^α atom of a residue in which r_o for lysine is set to 7 Å. Simulations are run at a constant temperature near the temperature used in the experiment.

We use three models to study the role of electrostatics in the alkaline-induced transitions. A protein's native state at pH 7 is usually optimized electrostatically, but upon deprotonation, favorable electrostatic effects are generally lost. Therefore, deprotonation of a residue is modeled by energetically destabilizing some of the native state contacts. The pure-funnel model, used in this study, is structure-based and includes a mean field treatment of electrostatics in which the pairwise energy of contacts involving a deprotonated residue is decreased homogeneously. The pure-funnel Hamiltonian can be written as $H_{pure-funnel} = V_{back} + V_{contact} + V_{heme} + V_{deprot}$, in which $V_{deprot} = -\Sigma_{i \text{ or } j \text{ deprot}} 0.1\varepsilon_{ij}(r_{ij})$. The factor 0.1 was estimated using a calculation of the electrostatic energy change upon deprotonation obtained by placing charges on the C^β atoms (using a Coulomb potential) in the crystal structure. The total destabilization energy distributed among all contacts involving deprotonated residues is

~5−15% of γ_{ij}. Neither the nature of the intermediate ensembles nor their relative stabilities change significantly when the γ_{ij} values are varied within this range.

A second model was also employed additionally to account for long-range and non-native electrostatic effects by using a simple Coulomb potential. In a separate study (*34*), we compared pair distribution functions calculated from simulation to those derived from experimental FRET data for cytochrome *c* (*35, 36*). The results of the previous study suggest that long-range electrostatic effects and a general hydrophobic collapse term improve the accuracy of the unfolded and partially unfolded conformations of cytochrome *c*. The Hamiltonian for the electrostatic/collapse model is given by $H_{\text{elec/col}} = V_{\text{back}} + V_{\text{contact}} + V_{\text{heme}} + V_{\text{elec}} + V_{R_g} + V_{\text{deprot}}$, where $V_{\text{elec}} = \sum_{li-jl>2} \gamma V_{R_g} \times (q_i q_j)/(4\pi\varepsilon_0\varepsilon r_{ij}) \times e^{-r_{ij}/\lambda_D}$. We account for solvent and ion screening effects using a Debye screening term (e^{-r_{ij}/λ_D}). The charge on residue *i* is ±1 depending on the identity of the residue. r_{ij} is the distance between C^{β} atoms of residues *i* and *j*. γ is the strength of a native contact. $\lambda_D = 2.5$ Å. Because the funneled energy terms implicitly contain native electrostatic effects, we chose a dielectric constant so that the long-range and non-native electrostatics have a more dominant effect in a manner similar to those of other studies (*34, 37*). The dielectric constant ($\varepsilon = 30$) is less than that for water but is still significantly more than the value of 3.5 calculated for ferricytochrome *c* (*38*). Furthermore, the electrostatic terms are scaled by the amount of collapse to avoid overcompensating for electrostatics implicitly included in the contact terms. Such a model mimics the expected interplay of electrostatic effects and hydrophobic collapse on the frustration of the energy landscape in the unfolded ensemble and allows for a nearly unfrustrated energy landscape near the native ensemble.

The last model includes the effects of specific side chain H-bonding changes in addition to the local electrostatics to capture the large energetic costs of deprotonating a residue and disrupting a H-bond. With an increase in pairwise contact energies between residues that participate in side chain H-bonding in the native state, the H-bonding potential captures specific interactions that are affected by deprotonation. The Hamiltonian for the H-bonding model is $H_{\text{H-bond}} = V_{\text{back}} + V_{\text{contact}} + V_{\text{heme}} + V_{\text{deprot}} + V_{\text{H-bond}}$. The possible H-bonds in the crystal structure were determined using WHATIF (*39*) and can be seen in Figure 1. H-Bonds are included if either the donor or acceptor is a side chain and the residues are more than three residues apart. The pairwise contact energies between residues determined to have H-bonds involving a side chain are increased by a factor of 3: $V_{\text{H-bond}} = \sum_{\text{H-bonded,not deprot}} 3\varepsilon_{ij}(r_{ij})$. If the side chain of the residue involved in the H-bond is deprotonated, the energetic bonus for the H-bond is removed.

RESULTS

Alkaline-Induced Intermediates. As the pH is varied over alkaline conditions, we calculate the most probable structural ensembles as a function of the folding reaction coordinate *Q*. The appropriate free energy profile is calculated using the grand canonical partition function $F(Q,\text{pH}) = -k_B T \ln[\Xi(Q,\text{pH})]$ (Figure 3). The pure-funnel model clearly reveals a transition at pH 9 where the depth of the folded state free energy minimum ($Q \approx 0.9$) becomes similar to

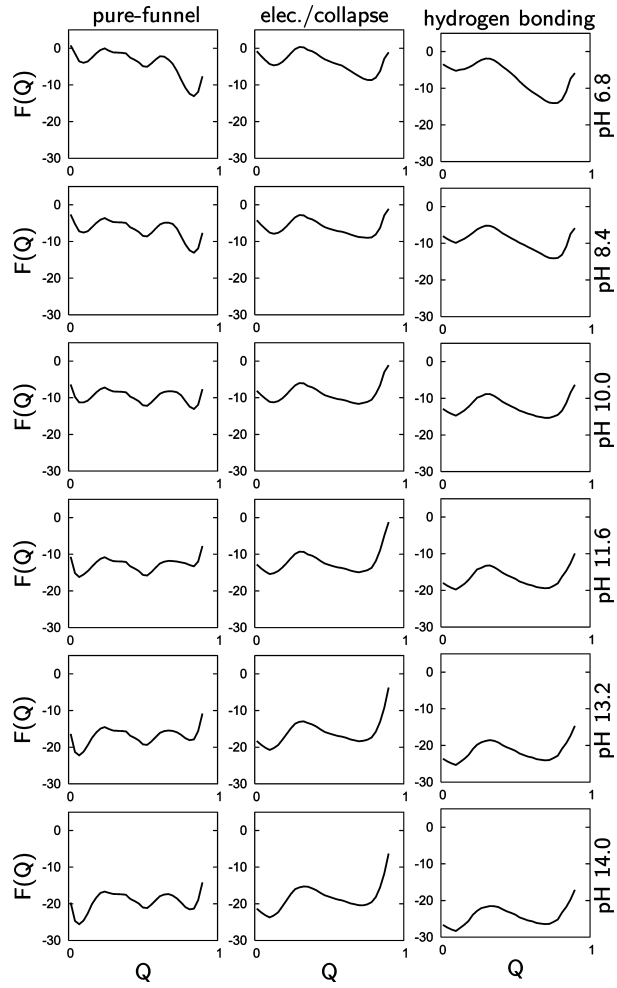

FIGURE 3: Free energy curves calculated using the grand canonical partition function as a function of reaction coordinate *Q* at specific pH values of 6.8, 8.4, 10.0, 11.6, 13.2, and 14.0. The curves were obtained from simulations using the pure-funnel (left column), electrostatic/collapse (middle column), and H-bonding models (right column).

the depth of the partially and globally unfolded state minima ($0.5 < Q < 0.6$ and $Q \approx 0.1$, respectively). The electrostatic/collapse and H-bonding models exhibit a more modest change in *Q* with pH. A partially unfolded minimum begins to appear around pH 9 ($Q \approx 0.8$). This minimum is broader than that associated with the fully folded state at pH 7. A second transition involves the global unfolding of the protein at pH >12. The prediction of several distinct transitions agrees with numerous experimental studies (*10−13, 17, 19*) that have identified multiple intermediates at alkaline pH. To analyze the ensemble calculations in more detail, we determined the population of each species as a function of pH (Figure 4). The population of species s in a grand canonical ensemble is

$$P(s) = \frac{\sum_{Q} e^{-\beta F_s(Q)} e^{-\beta \mu_s} K_{\text{lig,s}}}{\sum_{Q} \Xi(Q, \text{pH})} \quad (1)$$

Over the entire pH range, four intermediates are found to dominate using either the pure-funnel or electrostatic/collapse model: the Lys73- and Lys79-misligated states, an unfolded hydroxide-bound state, and an unfolded lysine- or water-

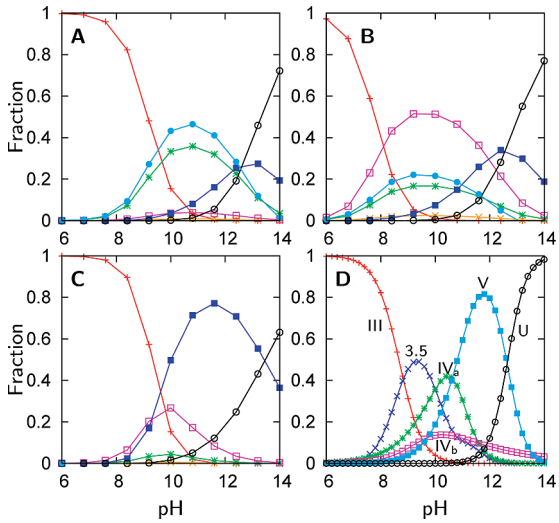

FIGURE 4: Fractional concentrations of the most stable chemical species as a function of pH for (A) the pure-funnel model, (B) the electrostatic/collapse model, and (C) the H-bonding model. The native (+), Lys73-misligated (*), Lys79-misligated (□), Lys72-misligated (×), unfolded and water- or lysine-misligated (●), OH-misligated and lysine- or tyrosine-deprotonated (■), and OH-misligated and lysine-, tyrosine-, or arginine-deprotonated (○) states are shown. (D) Probability distribution of states inferred from FTIR studies of semisynthetic protein with carbon- and deuterium-labeled residues (*19*).

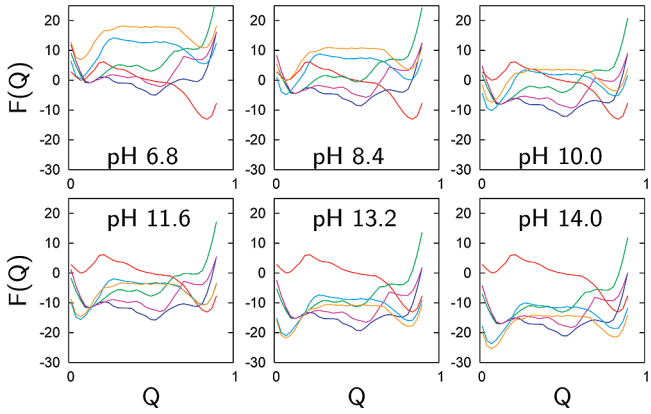

FIGURE 5: Free energy curves of the most stable chemical species as a function of reaction coordinate Q at specific pH values of 6.8, 8.4, 10.0, 11.6, 13.2, and 14.0. The chemical species depicted are the native (red), Lys73-misligated (blue), Lys79-misligated (magenta), Lys72-misligated (green), hydroxide-misligated and lysine- or tyrosine-deprotonated (cyan), and hydroxide-misligated and lysine-, tyrosine-, or arginine-deprotonated (yellow) states. The free energy curves were calculated using the grand canonical partition function based on pure-funnel model simulations.

bound state. However, only three intermediates are needed to describe the results of the model with explicit H-bonding: the Lys73- and Lys79-misligated states and a partially folded hydroxide-bound state. The free energy profiles of the most probable species after evaluation by the grand canonical partition function are illustrated in Figure 5. The figure displays the effect of misligation and deprotonation on the shape of the free energy profiles as well as the effect of the grand canonical partition function on the their stability. The predicted presence of several intermediates agrees with many studies (*13*−*17, 19*) which demonstrate that the transition at pH 9 is an equilibrium between the native state (III), the two lysine-misligated states (IV_a and IV_b), and another state with limited structural information (3.5). The calculations

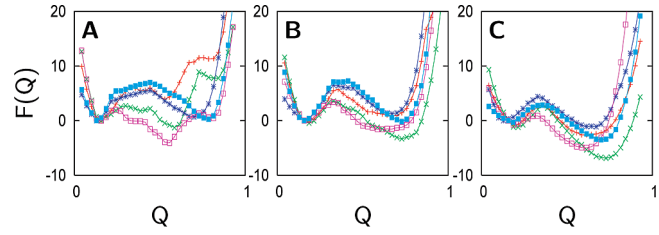

FIGURE 6: Free energy curves as a function of reaction coordinate Q for the five most stable lysine-misligated intermediates from simulations using the (A) pure-funnel, (B) electrostatic/collapse, and (C) H-bonding models. The most stable lysine intermediates are Lys53 (■), Lys55 (*), Lys72 (+), Lys73 (□), and Lys79 (×).

Table 2: Structural Accuracy of Lys73-Misligated Structures

model	rmsd[a]	Q[a]	Q_{cut}[a]	CE score[a]
pure-funnel	6.5 ± 0.6	0.21 ± 0.01	0.36 ± 0.01	2.1 ± 0.1
electrostatic/ collapse	3.2 ± 0.4	0.25 ± 0.01	0.37 ± 0.02	4.4 ± 0.3
H-bonding	3.5± 0.4	0.25 ± 0.1	0.36 ± 0.01	4.2 ± 0.4

[a] The NMR structure of the Lys79Ala Lys73-misligated structure is compared to 30 structures randomly selected from the free energy minima calculated from each simulation.

also show that, in agreement with experiment, there is a transition at pH 11−12 that involves a global unfolding event with multiple partially unfolded states (*12, 17, 19*).

Lysine-Misligated Intermediates. Our simulations provide information about the structures and relative stability of the lysine intermediates. There is much experimental evidence that Lys73 and Lys79 form stable misligates with maximal probability near pH 10.5 (*13*−*17, 19, 40*). Simulations were carried out in which each of the 19 lysine residues was forced to misligate the heme center and with protonation states corresponding to pH 10.5 (Table 1). Free energy curves were calculated to visualize the relative stability and energy barriers between the unfolded state ($Q \approx 0.1$) and the folded state ($0.6 < Q < 0.8$). The pure-funnel model, as discussed previously (*19*), predicts that Lys73- and Lys79-misligated ensembles are more stable than all other lysine misligates (Figure 6). Similar results are obtained with the electrostatic/collapse and H-bonding models, but the relative stability of the Lys79 intermediate is increased. Both the electrostatic/collapse and H-bonding models predict that the lysine misligates' free energy minima are at a Q higher than that found with the pure-funnel model which shows a larger number of native contacts. All three models predict that Lys53- and Lys55-misligated species could normally be present in small concentrations and perhaps would dominate if Lys79 and Lys73 were not available [i.e., due to trimethylation (*41*) or protein binding (*42*)].

To assess the accuracy of the simulations, we compare the predicted structures from the Lys73-misligated minima to the published NMR structure of this misligated state (*43*). As shown in Table 2, the rmsd, the Q score, and the combinatorial extension (CE) score show that the electrostatic/collapse and H-bonding models correspond more closely to the NMR results than does the pure-funnel model. Nevertheless, structures from all three models possess nearly the same average fraction of native contacts, Q_{cut}, as a measure of the distance pair similarity between the predicted and NMR structure (with an 8 Å cutoff). The comparison of the structures from the simulated ensembles, and the NMR model

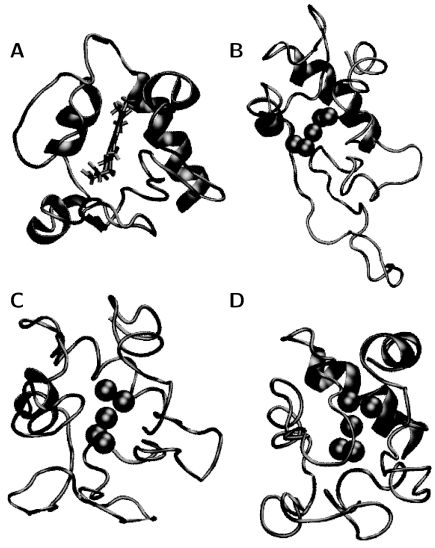

FIGURE 7: (A) NMR structure of a Lys79Ala Lys73-misligated cytochrome *c* (PDB entry 1LMS). Representative structures of the Lys73-misligated cytochrome *c* taken from the free energy minima of the (B) pure-funnel, (C) electrostatic/collapse, and (D) H-bonding simulations.

deposited in the Protein Data Bank (PDB) is shown in Figure 7. The primary difference is that the omega loop (residues 40−56) seems to more closely dock to the rest of the protein in the electrostatic/collapse and H-bonding models than it does in the pure-funnel model.

Solvent Exposure Calculations. Hydrogen exchange experiments (*44*) have been used to infer the dominant folding route of cytochrome *c* which, in agreement with results of theoretical modeling (*1*), involves the sequential folding of five cooperative folding units (Figure 1). However, the hydrogen exchange experiments suggest that the alkaline transition at pH 9 involves the unfolding or restructuring of only the two least stable folding units: the Met80 loop (residues 70−88) and the omega loop (residues 40−56) (*16, 45*). Calculations involving the number of contacts per residue have been shown to reproduce protection factors determined from hydrogen exchange data (*46*−*52*) and will be used in this work to discuss the degree of solvent exposure. By computing the ensemble average of the number of contacts per residue over all conformations in the ensemble, we confirm that the two least stable units are as exposed to solvent at pH 10.8 as they are at pH 14.0 (Figure 8A). Averaging the number of contacts per residue over each folding unit reveals that the least stable folding units are less protected than the more stable folding units and show unfolding transitions at pH 9 (Figure 8B). The remaining units unfold near pH 12.

We can examine the simulated alkaline-induced ensembles in a manner closely analogous to what is done for the site specific FTIR experiments. FTIR studies of semisynthetic protein with carbon- and deuterium-labeled residues provide information about the local environment of the labeled residues because the observed signals are sensitive to solvent exposure, ligation to the heme iron, and protonation changes (*19, 53*−*55*). To relate the simulations to FTIR signals that are mainly sensitive to changes in solvent exposure, we determine the probability distributions of the number of contacts at five residues as a function of pH (Figure 9). The probability of c contacts for a residue at a given pH is

$$P(c) = \sum_s \frac{\sum_Q e^{-\beta F_s(Q)} e^{-\beta \mu_s} K_{\mathrm{lig},s} \theta[c_s(Q)]}{\sum_Q \Xi(Q, \mathrm{pH})} \quad (2)$$

The value of $\theta[c_s(Q)]$ is usually 0 but is given a value of 1 if c equals $c_s(Q)$, the number of contacts for residue i in which the distance, $\langle r_{ij}(Q) \rangle$, is less than 8 Å and $|i - j| > 1$. The results can be qualitatively described via a two-state fit at Leu68 and by three-state fits at Lys73, Lys79, and Met80, consistent with the observed changes in the FTIR experiment (Figure S1 of the Supporting Information). The electrostatic/collapse and H-bonding models more correctly predict the Leu68 transition midpoint which occurs at pH 12.7 experimentally but is predicted to occur at pH <10.0 in the pure-funnel model (Figure 9A). In contrast to the Leu68 FTIR signals which are sensitive primarily to changes in solvent exposure and the Met80 signals which are sensitive to both solvent exposure and ligation, the Lys72, Lys73, and Lys79 signals are affected by changes in solvent exposure, ligation, and deprotonation. While comparison of the lysine signals to simulation is difficult, it is interesting to note that the calculated transitions at Lys73 and Lys79 appear to involve more distinct structural ensembles than the Lys72 transition does (Figure 9D−F). This agrees with the interpretations of the experimental IR data that suggest there is a three-state transition for Lys73 and Lys79 and a two-state transition for Lys72. The simulations also qualitatively capture the character of the transition at Met80 (Figure 9B,C). Experimental evidence suggests that between pH 9 and 11, Met80 experiences two different environments, a buried state that is more heterogeneous than the native state at pH 7 and a solvent-exposed state. Another transition near pH 11 results in complete solvent exposure at Met80. The probability distributions of the Met80 contacts show similar behavior (Figure 9C and Figures S1 and S2 of the Supporting Information).

DISCUSSION

FTIR studies of site selectively deuterated cytochrome *c* have resolved the existence of two intermediates, which are maximally populated at pH 9.3 (3.5) and 11.8 (V), in addition to the lysine-misligated intermediates (*19*). However, the experiments provide only limited structural information about these intermediates. The data suggest that both intermediates are partially folded since the FTIR signal at Leu68 indicates protection from solvent but the Met80 signal indicates significant solvent exposure. Both the pure-funnel and electrostatic/collapse models fail fully to reproduce these structural details because they predict Leu68 is solvent-exposed in both 3.5 and V (Table 3). In these models, the protein unfolds upon deprotonation of lysine and tyrosine and is only modestly affected by the deprotonation of arginine. However, due to the stabilizing H-bonds between the two most stable folding subunits (the N- and C-terminal helices and the 60s helix) (*44*), the model with modified H-bonding captures the global unfolding event upon arginine deprotonation. The modified H-bonding model also predicts that the partially unfolded intermediate structures for 3.5 and

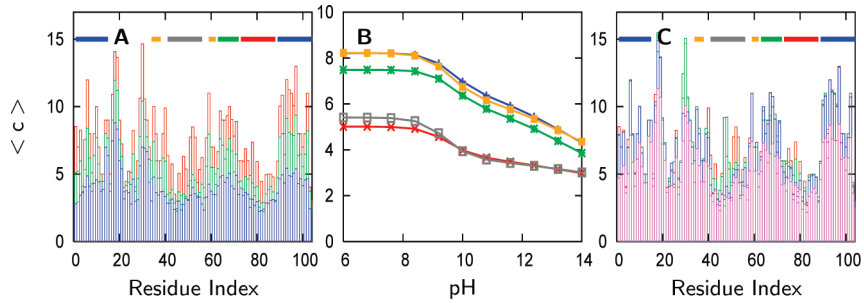

FIGURE 8: (A) Ensemble average of the number of contacts [$\langle c \rangle = \sum_c cP(c)$] per residue at pH 6.0 (red), pH 10.8 (green), and pH 14.0 (blue) calculated using the entire structural ensemble at those pH values. The lines across the top indicate the folding units as defined by hydrogen exchange experiments (*44*) and simulation (*1*). (B) Values of $\langle c \rangle$ averaged over each folding unit. (C) Average of the number of contacts per residue calculated for individual species: native (red), Tyr48-misligated (green), Tyr67-misligated (blue), and hydroxide-ligated and lysine- or tyrosine-deprotonated (magenta). Each plot was generated using the H-bonding model simulations. Similar results were calculated for the electrostatic/collapse and pure-funnel models.

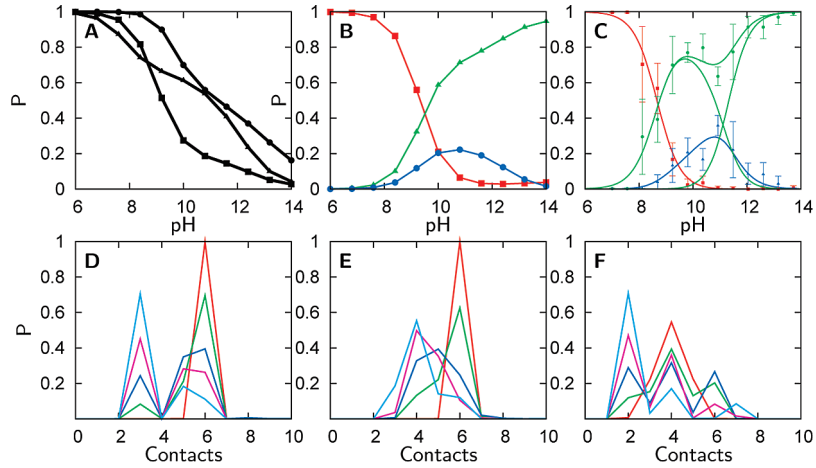

FIGURE 9: (A) Probability of finding 10 contacts, as found in the native state, with Leu68 calculated using the pure-funnel (squares), electrostatic/collapse (triangles), and H-bonding (circles) models. (B) Probability of finding six (red), four (blue), or two (green) contacts with Met80 calculated using the pure-funnel model. (C) Fractional concentrations of the folded (red), high-pH and solvent-exposed (green), and intermediate (blue) signals observed for Met80 by FTIR (*19*). (D−F) Probability of the number of contacts (c) with Lys72, Lys73, and Lys79, respectively, calculated using the H-bonding model. The curves plotted are for pH 6.0 (red), pH 9.2 (green), pH 10.0 (blue), pH 11.6 (magenta), and pH 14.0 (teal). The complete set of probability distributions of the number of contacts for all models is shown in the Supporting Information.

Table 3: Structural Description and Ligation State of the Dominant Species in the Simulation As Compared to the Experiments

experimentally observed state[a]	pure-funnel	electrostatic/collapse	H-bonding	experiment
III	F/Met80	F/Met80	F/Met80	F/Met80
3.5	U/H_2O	U/H_2O	PF/$^-$OH	PF/?
IV_a	PF/Lys73	PF/Lys73	PF/Lys73	PF/Lys73
IV_b	PF/Lys79	PF/Lys79	PF/Lys79	PF/Lys79
V	U/$^-$OH	U/$^-$OH	PF/$^-$OH	PF/?
U	U/$^-$OH	U/$^-$OH	U/$^-$OH	U/$^-$OH

[a] States are either folded (F with $Q > 0.8$), partially folded (PF with $0.3 < Q < 0.8$), or unfolded (U with $Q < 0.3$) and are ligated with either an amino acid or solvent molecule.

V are similar since the distribution of the partially unfolded hydroxide-bound state (Figure 4C), which dominates at intermediate pH, may represent the sum of 3.5 and V as inferred from experiment (Figure 4D). Previous experiments, in fact, suggest that the ligand in state V is hydroxide (*17*). An alternative possibility is that the 3.5 intermediate results from a chemical species not included in the current simulations. Another chemical species having a maximum population at pH 9.3 could shift the population of the predicted intermediates so that they form at a higher pH. Because of the lower pK_a of tyrosine compared to that of lysine, its deprotonation could favor a tyrosine-misligated intermediate or trigger an unfolding event that is localized to the Met80 loop, both of which could be consistent with experimental observations. Analogous to the calculations of the relative stability between the lysine misligates, we predict that the Tyr48 and Tyr67 misligates would be most stable as indicated by all three models (Figure S3 of the Supporting Information). However, the structures of both misligates do not agree with interpretations of hydrogen exchange and residue specific FTIR data based upon examination of the ensemble average of the number of contacts per residue (Figure 8C). Consistent with experimental interpretations, the predicted structures of the Tyr67 misligate and the hydroxide-bound species (V) demonstrate solvent-exposed Met80 and omega loops (residues 70−88 and 40−56, respectively). In contrast, the Tyr48 misligate has solvent exposure similar to that for the native state. Upon closer

analysis, the Tyr67-misligated species and the hydroxide-bound species (V) are predicted to have a significantly decreased number of contacts at Met80 but a similar number of contacts at Leu68 when compared to those of the native state which agrees with interpretations of FTIR experiments. Therefore, this work strongly suggests that the 3.5 intermediate is either a Tyr67 misligate or a hydroxide-bound species. All three models predict that state V is a hydroxide-bound species, in agreement with experimental work (*17*).

Using these three related but distinct models provides insight into the dominant forces that govern these multiple pH transitions. Protein folding theory indicates that local interactions, interactions close in distance but not necessarily close in sequence, are the most dominant contributors to the funneled nature of the energy landscape. Deprotonation decreases the strength of local interactions that stabilize the native state, thereby decreasing the funneled nature of the energy landscape. In the pure-funnel model, long-range electrostatic effects are ignored and side chain interactions are treated with homogeneous nonadditive terms. Remarkably, even this most simplistic model is able to predict intermediate stabilities well because the destabilizing effects of deprotonation are captured in the changes of the local interactions. Thus, the homogeneous nonadditive pure-funnel model describes well the relative stability of the lysine-misligated intermediates in addition to the relative folding sequence of the folding units (*1*). However, some detailed structural information appears to be poorly captured in this simplest model's prediction. For instance, the agreement of structural features of the Lys73-misligated state with experiment is relatively poor because the omega loop (residues 40−56) is predicted to be randomly coiled but appears to be partially folded on the basis of the NMR studies. The electrostatic/collapse model shows better agreement with the Lys73-misligated NMR structure because the collapse potential forces the omega loop to dock more completely with the rest of the structure. While structural accuracy is apparently improved by the addition of the electrostatic and collapse potentials, the more complex model fails to capture the stability change upon arginine deprotonation. This discrepancy is resolved using the model having modified H-bonding because the Arg91 side chain forms a H-bond at the interface of the two most stable folding units of cytochrome *c*. This H-bond is vital to the stability of the protein since preserving the interface allows a significant part of the folding nucleus to remain intact. In general, the H-bond potential rescales the stabilities of the folding units and the interfaces between folding units so that the dominant folding route correctly predicted with the pure-funnel model is more structurally faithful. Treating H-bonding explicitly results in improved agreement of the Lys73-misligated NMR structure with the simulation as several H-bonds help stabilize the omega loop.

This study demonstrates the importance of including the full range of chemical species in the analysis of complex multispecies processes such as the alkaline-induced transitions of cytochrome *c*. While only a few states dominate at any given pH, many different intermediates become populated, albeit at low concentrations. Consistent with the funneled landscape picture, mutating the protein to selectively destabilize one intermediate ordinarily gives rise to a competing species that is slightly less stable. Also, cytochrome *c* forms from different species are funneled into slightly different near native state ensembles which may explain experimental discrepancies between measurements of the relative concentrations of the Lys79-, Lys73-, and Lys72-misligated states on different variants of cytochrome *c* (*14*−*16, 56, 57*). The current grand canonical ensemble methodology can serve as an explanatory tool for analysis of multistate folding transitions. Since the method requires few free parameters, the same methodology can easily be applied to many metalloprotein systems. The introduction of specific non-native contacts, correlated with known inorganic chemistry, into an otherwise perfectly funneled energy landscape demonstrates the importance of energetic frustration in chemically induced unfolding as demonstrated by the alkaline-induced transitions of cytochrome *c*.

ACKNOWLEDGMENT

We thank Joerg Zimmerman for thoughtful discussions.

SUPPORTING INFORMATION AVAILABLE

Further calculations as referenced in the text. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES

1. Weinkam, P., Zong, C. H., and Wolynes, P. G. (2005) A funneled energy landscape for cytochrome c directly predicts the sequential folding route inferred from hydrogen exchange experiments. *Proc. Natl. Acad. Sci. U.S.A. 102*, 12401–12406.
2. Shoemaker, B. A., Wang, J., and Wolynes, P. G. (1997) Structural correlations in protein folding funnels. *Proc. Natl. Acad. Sci. U.S.A. 94*, 777–782.
3. Alm, E., and Baker, D. (1999) Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. Natl. Acad. Sci. U.S.A. 96*, 11305–11310.
4. Munoz, V., and Eaton, W. A. (1999) A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl. Acad. Sci. U.S.A. 96*, 11311–11316.
5. Clementi, C., Nymeyer, H., and Onuchic, J. N. (2000) Topological and energetic factors: What determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol. 298*, 937–953.
6. Clementi, C., Jennings, P. A., and Onuchic, J. N. (2000) How native-state topology affects the folding of dihydrofolate reductase and interleukin-1 β. *Proc. Natl. Acad. Sci. U.S.A. 97*, 5871–5876.
7. Galzitskaya, O. V., and Finkelstein, A. V. (1999) A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proc. Natl. Acad. Sci. U.S.A. 96*, 11299–11304.
8. Koga, N., and Takeda, S. (2001) Roles of native topology and chain-length scaling in protein folding: A simulation study with a Go-like model. *J. Mol. Biol. 313*, 171–180.
9. Jones, C., et al. (1993) Fast events in protein-folding initiated by nanosecond laser photolysis. *Proc. Natl. Acad. Sci. U.S.A. 90*, 11860–11864.
10. Theorell, H., and Akesson, A. J. (1941) Studies on cytochrome c. II. The optical properties of pure cytochrome c and some of its derivatives. *J. Am. Chem. Soc. 63*, 1812–1818.
11. Moore, G. R., and Pettigrew, G. W. (1990) *Cytochromes c. Evolutionary, Structural and Physicochemical Aspects*, Springer-Verlag, Berlin.
12. Scott, R. A., and Mauk, A. G. (1996) *Cytochrome c: A Multidisciplinary Approach*, University Science Books, Sausalito, CA.
13. Banci, L., Bertini, I., Bren, K. L., Gray, H. B., and Turano, P. (1995) pH-dependent equilibria of yeast Met80Ala-iso-1-cytochrome c probed by NMR spectroscopy: A comparison with the wild-type protein. *Chem. Biol. 2*, 377–383.
14. Hong, X. L., and Dixon, D. W. (1989) NMR study of the alkaline isomerization of ferricytochrome c. *FEBS Lett. 246*, 105–108.
15. Rosell, F. I., Ferrer, J. C., and Mauk, A. G. (1998) Protein-linked protein conformational switching: Definition of the alkaline con-

formational transition of yeast iso-1-ferricytochrome c. *J. Am. Chem. Soc. 120*, 11234–11245.
16. Maity, H., Rumbley, J. N., and Englander, S. W. (2006) Functional role of a protein foldon: An omega-loop foldon controls the alkaline transition in ferricytochrome c. *Proteins 63*, 349–355.
17. Dopner, S., Hildebrandt, P., Rosell, F. I., and Mauk, A. G. (1998) Alkaline conformational transitions of ferricytochrome c studied by resonance Raman spectroscopy. *J. Am. Chem. Soc. 120*, 11246–11255.
18. Kihara, H., et al. (1976) Kinetic study of isomerization of ferricytochrome c at alkaline pH. *Biochim. Biophys. Acta 430*, 225–243.
19. Weinkam, P., et al. (2008) Characterization of Alkaline Transitions in Ferricytochrome *c* Using Carbon-Deuterium IR Probes. *Biochemistry 47*, 13470–13480.
20. Yang, A., and Honig, B. (1993) On the pH dependence of protein stability. *J. Mol. Biol. 231*, 459–474.
21. Alexov, E. G., and Gunner, M. R. (1997) Incorporating protein conformational flexibility into the calculation of pH-dependent protein properties. *Biophys. J. 72*, 2075–2093.
22. Khandogin, J., and Brooks, C. L. (2005) Constant pH molecular dynamics with proton tautomerism. *Biophys. J. 89*, 141–157.
23. Zhou, H. X., and Vijayakumar, M. (1997) Modeling of protein conformational fluctuations in pK_a predictions. *J. Mol. Biol. 267*, 1002–1011.
24. McDonald, S. M., Willson, R. C., and McCammon, J. A. (1995) Determination of the pK_a values of titratable groups of an antigen-antibody complex, HyHEL-5-hen egg lysozyme. *Protein Eng. 8*, 915–924.
25. Resat, H., and Mezei, M. (1996) Grand canonical ensemble Monte Carlo simulation of the dCpG/proflavine crystal hydrate. *Biophys. J. 71*, 1179–1190.
26. Woo, H. J., Dinner, A. R., and Roux, B. (2004) Grand canonical Monte Carlo simulations of water in protein environments. *J. Chem. Phys. 121*, 6392–6400.
27. Clark, M., Guarnieri, F., Shkurko, I., and Wiseman, J. (2006) Grand canonical Monte Carlo simulation of ligand-protein binding. *J. Chem. Inf. Model. 46*, 231–242.
28. Kumar, S., Bouzida, D., Swendsen, R., Kollman, P., and Rosenberg, J. (1992) The weighted histogram analysis method for free-energy calculations on biomolecules. 1. The method. *J. Comput. Chem. 13*, 1011–1021.
29. Dopner, S. (1999) The structural and functional role of lysine residues in the binding domain of cytochrome c in the electron transfer to cytochrome c oxidase. *Eur. J. Biochem. 261*, 379–391.
30. Friedrichs, M. S., and Wolynes, P. G. (1989) Toward protein tertiary structure recognition by means of associative memory Hamiltonians. *Science 246*, 371–373.
31. Sasai, M., and Wolynes, P. G. (1992) Unified theory of collapse, folding and glass transitions in associative-memory Hamiltonian models of proteins. *Phys. Rev. A 46*, 7979–7997.
32. Ejtehadi, M. R., Avall, S. P., and Plotkin, S. S. (2004) Three-body interactions improve the prediction of rate and mechanism in protein folding models. *Proc. Natl. Acad. Sci. U.S.A. 101*, 15088–15093.
33. Eastwood, M., and Wolynes, P. G. (2001) Role of explicitly cooperative interactions in protein folding funnels: A simulation study. *J. Chem. Phys. 114*, 4702–4716.
34. Weinkam, P., Pletneva, E. V., Gray, H. B., Winkler, J. R., and Wolynes, P. G. (2009) Electrostatic effects on funneled landscapes and structural diversity in denatured protein ensembles. *Proc. Natl. Acad. Sci. U.S.A. 106* (6), 1796–1801.
35. Lyubovitsky, J. G., Gray, H. B., and Winkler, J. R. (2002) Structural features of the cytochrome c molten globule revealed by fluorescence energy transfer kinetics. *J. Am. Chem. Soc. 124*, 14840–14841.
36. Pletneva, E. V., Gray, H. B., and Winkler, J. R. (2005) Many faces of the unfolded state: Conformational heterogeneity in denatured yeast cytochrome c. *J. Mol. Biol. 345*, 855–867.
37. Cho, S. S., Weinkam, P., and Wolynes, P. G. (2008) Origins of barriers and barrierless folding in BBL. *Proc. Natl. Acad. Sci. U.S.A. 105*, 118–123.
38. Simonson, T., Perahia, D., and Brunger, A. T. (1991) Microscopic theory of the dielectric-properties of proteins. *Biophys. J. 59*, 670–690.
39. Vriend, G. (1990) What if: A molecular modeling and drug design program. *J. Mol. Graphics 8*, 52–56.
40. Rosell, F. I., et al. (2000) Characterization of an alkaline transition intermediate stabilized in the Phe82Trp variant of yeast iso-1-cytochrome c. *Biochemistry 39*, 9047–9054.
41. Paik, W. K., Cho, Y. B., Frost, B., and cKIM, S. D. (1989) Cytochrome c Methylation. *Biochem. Cell Biol. 67*, 602–611.
42. Rieder, R., and Bosshard, H. R. (1978) Cytochrome c-oxidase Binding Site on Cytochrome c: Differential Chemical Modification of Lysine Residues in Free and Oxidase-Bound Cytochrome c. *J. Biol. Chem. 253*, 6045–6053.
43. Assfalg, M., et al. (2003) Structural model for an alkaline form of ferricytochrome c. *J. Am. Chem. Soc. 125*, 2913–2922.
44. Maity, H., Maity, M., and Englander, S. W. (2004) How cytochrome c folds, and why: Submolecular foldon units and their stepwise sequential stabilization. *J. Mol. Biol. 343*, 223–233.
45. Hoang, L., Maity, H., Krishna, M. M. G., Lin, Y., and Englander, S. W. (2003) Folding units govern the cytochrome c alkaline transition. *J. Mol. Biol. 331*, 37–43.
46. Miller, D. W., and Dill, K. A. (1997) Ligand binding to proteins: The binding landscape model. *Protein Sci. 6*, 2166–2179.
47. Gsponer, J., et al. (2006) Determination of an ensemble of structures representing the intermediate state of the bacterial immunity protein Im7. *Proc. Natl. Acad. Sci. U.S.A. 103*, 99–104.
48. Hilser, V. J., and Freire, E. (1996) Structure-based calculation of the equilibrium folding pathway of proteins. Correlation with hydrogen exchange protection factors. *J. Mol. Biol. 262*, 756–772.
49. Sheinerman, F. B., and Brooks, C. L. (1998) Molecular picture of folding of a small α/β protein. *Proc. Natl. Acad. Sci. U.S.A. 95*, 1562–1567.
50. Bahar, I., Wallqvist, A., Covell, D. G., and Jernigan, R. L. (1998) Correlation between native-state hydrogen exchange and cooperative residue fluctuations from a simple model. *Biochemistry 37*, 1067–1075.
51. Garcia, A. E., and Hummer, G. (1999) Conformational dynamics of cytochrome c: Correlation to hydrogen exchange. *Proteins: Struct., Funct., Genet. 36*, 175–191.
52. Viguera, A. R., and Serrano, L. (2003) Hydrogen-exchange stability analysis of Bergerac-Src homology 3 variants allows the characterization of a folding intermediate in equilibrium. *Proc. Natl. Acad. Sci. U.S.A. 100*, 5730–5735.
53. Chin, J. K., Jimenez, R., and Romesberg, F. E. (2001) Direct observation of protein vibrations by selective incorporation of spectroscopically observable carbon-deuterium bonds in cytochrome c. *J. Am. Chem. Soc. 123*, 2426–2427.
54. Chin, J. K., Jimenez, R., and Romesberg, F. E. (2002) Protein dynamics and cytochrome c: Correlations between ligand vibrations and redox activity. *J. Am. Chem. Soc. 124*, 1846–1847.
55. Sagle, L. B., Zimmerman, J., Baldridge, K., Dawson, P., and Romesberg, F. (2004) Carbon-deuterium bonds as an IR probe in cytochrome c. *Protein Sci. 13*, 216–216.
56. Ferrer, J. C. (1993) Identification of Lys79 as an iron ligand in one form of alkaline yeast iso-1-ferricytochrome-c. *J. Am. Chem. Soc. 115*, 7507–7508.
57. Pollock, W. B., Rosell, F. I., Twitchett, M. B., Dumont, M. E., and Mauk, A. G. (1998) Bacterial expression of a mitochondrial cytochrome c. Trimethylation of Lys72 in yeast iso-1-cytochrome c and the alkaline conformational transition. *Biochemistry 37*, 6124–6131.

BI802293M