PROTEINS: Structure, Function, and Genetics 39:309–316 (2000)

# Use of MM-PB/SA in Estimating the Free Energies of Proteins: Application to Native, Intermediates, and Unfolded Villin Headpiece

**Matthew R. Lee,* Yong Duan, and Peter A. Kollman**
*Department of Pharmaceutical Chemistry, University of California, San Francisco, California*

***ABSTRACT*** **We investigated the stability of three different ensembles of the 36-mer villin headpiece subdomain, the native, a compact folding intermediate, and the random coil. Structures were taken from a 1-μs molecular dynamics folding simulation and a 100-ns control simulation on the native structure. Our approach for each conformation is to first determine the solute internal energy from the molecular mechanics potential and then to add the change resulting from solvation ($\Delta G_{solv}$). Explicit water was used to run the simulation, and a continuum model was used to estimate $\Delta G_{solv}$ with the finite difference Poisson-Boltzmann model accounting for the polarization part and a linearly surface area-dependent term for the non-polar part. We leave out the solute vibrational entropy from these values but demonstrate that there is no statistical difference among the native, folding intermediate, and random coil ensembles. We find the native ensemble to be ≈26 kcal/mol more stable than the folding intermediate and ≈39 kcal/mol more stable than the random coil ensemble. With an experimental estimate for the free energy of denaturation equal to 3 kcal/mol, we approximate the non-native degeneracy to lie between $10^{16}$ and $10.^{25}$ We also present a possible scheme for the mechanism of folding, first-order exponential decay of a putative transition state, with an estimate for the $t_{1/2}$ of folding of ≈1 μs. Proteins 2000;39:309–316.**


**Key words: free energy; proteins; villin; unfolded degeneracy; folding time**

## INTRODUCTION

Among the goals of computer simulations of protein systems are to understand the mechanism and kinetics of folding and to predict the correct native structure from the very large number of possibilities.

Because the timescale of protein folding, ranging from tens of microseconds to seconds, makes it currently prohibitive to study the entire mechanisms of folding by using all-atom models with explicit solvent, simplified models have been used and have given exciting insights.[1–4] All-atom models have been used to give insights into protein unfolding by raising the temperature to high values.[5–9] Also, advances in computer power have enabled studies on the early stages of the mechanism of protein folding,[10,11] using all-atom, explicit solvent models. Progress has also been made in predicting protein structure from sequence,[12,13] but there is still much work to be performed. A crucial element in reaching the goal of predicting protein structures is the development of a method that can discriminate between the correct native structure and other alternatives. Because native protein structures at physiological temperatures are determined by their free energies, which consist of competing enthalpic and entropic parts, gas-phase energies alone are unlikely to be effective for such a purpose, even at the atomic level.[14] As a result, two general types of approaches have emerged for adding the entropy: knowledge-based and physical. The knowledge-based methods rely on comparison with properties of known proteins[15] taken from the protein structure databases. The physically based methods use functions from molecular mechanical force fields. Recently, we[16–18] and Hermans' group[19] proposed two similar physical methods and showed they were effective in comparing different structures of free energies of nucleic acids[16] and proteins.[19] The challenge remains to try such functions on even more challenging decoy structures that come increasingly closer to the correct native structure. An interesting test case has been afforded us in this regard, because our simulation of the early phase of villin folding found a variety of structures including a metastable intermediate. We also have a control simulation on native villin (minimized average NMR structure[20]) which lasted 100 ns, approximately 10 times longer than any comparable simulation on a native protein.

We have applied the molecular mechanics-Poisson Boltzmann/surface area (MM-PB/SA) method developed by Srinivasan et al.[16] to the folding and native simulations of villin, a net 1.1 μs worth of structures. We find, encouragingly, that the native structure is calculated to have a noticeably more favorable free energy, 15–35 kcal/mol lower than all other structures, with the intermediate characterized by the lowest free energy found during the folding trajectory.

Grant sponsor: National Institutes of Health; Grant numbers: GM-0717 and GM-29072.

*Correspondence to: Matthew R. Lee, Department of Pharmaceutical Chemistry, University of California, San Francisco, CA 94143. E-mail: mrlee@cgl.ucsf.edu

Received 31 August 1999; Accepted 6 January 2000

## MATERIALS AND METHODS

As previously reported,[11] we ran molecular dynamics with the Cornell et al.[21] all-atom force field, the TIP3P model for water, periodic boundary conditions, and an 8-Å cutoff for all solute/solvent non-bonded interactions (with no cutoff for intrasolute interactions) to sample conformational space in the isothermal-isobaric ensemble. Energy calculations reported in this study were made every 100 ps, totaling 10,000 evaluations for the folding trajectory and 1,000 for the control simulation. We first approximated the free energy of each snapshot as the sum of two terms: the internal energy of the protein ($E_{MM}$) and a solvation free energy ($\Delta G_{solv}$).

$$G_1 = E_{MM} + \Delta G_{solv} \quad (1)$$

$E_{MM}$ is the sum of an internal strain energy ($E_{int}$), a van der Waals energy ($vdW_{tot}$), and an electrostatic energy ($EEL_{tot}$). $E_{int}$ is the energy associated with vibration of covalent bonds and rotation of valence bond angles and torsional angles. $vdW_{tot}$ and $EEL_{tot}$ are further broken down into short-range values, those that are within three covalent bonds ($vdW_{1-4}$ and $EEL_{1-4}$), and long-range values that are four or more covalent bonds apart ($vdW_{NB}$ and $EEL_{NB}$).

The entropy of a given snapshot ($S_{solute}$), excluding conformational entropy, can be estimated by calculating the translational, rotational, and vibrational partition functions with normal mode analysis on a Newton-Raphson minimization. Because configurational differences stem primarily from the latter, we refer to this as the "vibrational" entropy. This is the most time-intensive part of the MM-PB/SA method on a per snapshot basis, and we performed the vibrational entropy calculation on five conformers each of the native state, the metastable folding intermediate, and the denatured state.

Obtaining the solvation free energy from an implicit description of solvent as a continuum is advantageous because it affords a solvation potential that is a function only of the solute's geometry, as discussed and implemented by Srinivasan et. al.[18]:

$$\Delta G_{solv} = \langle \Delta G_{NP} \rangle + \langle \Delta G_{pol} \rangle \approx (\gamma \cdot \mathrm{SASA} + b) + \langle \Delta G_{pol} \rangle. \quad (2)$$

The non-polar solvation free energy ($\Delta G_{NP}$) includes the (largely entropic) cost of creating a solute-sized cavity in solvent and the free energy of inserting the discharged solute into that cavity. Also referred to as the first solvation shell effects, this term has been found experimentally in hydrocarbons to be linearly related to the solvent accessible surface area (SASA), which is obtained from Sanner's MSMS algorithm[22] (probe radius = 1.400 Å). The $\gamma$ coefficient is set to 5.42 cal/mol × Å,$^2$ and b is set to 920 cal/mol. The electrostatic solvation free energy ($\Delta G_{pol}$) is the cost of charging the discharged solute in the cavity. We adhered to the same Poisson-Boltzmann protocol as described by Srinivasan et. al.,[18] which uses DelPhi[23]and most of its standard default parameters, together with PARSE atomic radii[24] and Cornell et al. charges,[21] to calculate the electrostatic solvation free energy difference for the system between exterior dielectrics of 80 (solvent) and unity (gas phase) according to the position-dependent electrostatic potential. One small difference in this usage of DelPhi is to use larger grid spacing of 0.5 Å, extending 20% beyond the edge of the solute. In addition, we used fewer finite difference iterations (1,000) for each ($\Delta G_{pol}$) calculation, which were still amply sufficient because we found the values in this system to reach 90% convergence at around 50 iterations.

## RESULTS AND DISCUSSION

### The Native Structure Has the Lowest MM-PB/SA Free Energy Estimate

Figures 1 and 3a show the actual MM-PB/SA free energy data as a function of time from the folding and control simulations. As shown in Table I, we predict the native villin headpiece conformation to be on average ≈25 kcal/mol more stable than the lowest energy state encountered during the 1-μs folding simulation (15 kcal/mol at the smallest gap). This non-native low-energy state is, as previously reported, highly compact with a residence time of 160 ns.[11] In comparison, we predict the native conformation to be on average ≈35 kcal/mol more stable than the unfolded state.

### The Folding Trajectory Roughly Obeys Boltzmann Statistics, According to MM-PB/SA

In our previous work, we further characterized the folding trajectory by a clustering method, using a limit of 3 Å main-chain RMSD from the cluster's average and found 30 marginally stable states that were populated with ≈500 or more of the 50,000 total coordinate sets.[11] The relationship between the natural log of the cluster population and the MM-PB/SA free energy appears to be a reasonably well-behaved Boltzmann distribution, with a correlation coefficient of −0.54. We do not expect a perfect inverse relationship because kinetic barriers distort the Boltzmann relationship in a non-ergodic trajectory and because the MM-PB/SA free energy is not completely accurate.

### Electrostatics Are the Major Source of Fluctuation, but Not a Good Predictor of $G_1$

As can be seen in Figures 2 and 3b, $\Delta G_{solv}$ and $E_{MM}$ each exhibit much more fluctuation than their sum, $G_1$. Over a typical 10-ns period, $\Delta G_{solv}$ and $E_{MM}$ will each oscillate over a 300 kcal/mol range and $G_1$ over a 50 kcal/mol range. The standard deviations over the entire 1-μs folding trajectory are 66.5 kcal/mol for $\Delta G_{solv}$, 73.7 kcal/mol for $E_{MM}$, and only 17.6 kcal/mol for $G_1$. The reason for such a disparity in variances is that $\Delta G_{solv}$ and $E_{MM}$ are strongly inversely related with a correlation coefficient of −0.97; large changes of $\Delta G_{solv}$ are always accompanied by approximately equal and opposite changes in $E_{MM}$. This inverse relationship can be explained by looking at their dependency on their individual electrostatic components. $E_{MM}$ has a correlation coefficient of 0.95 with its electrostatic term, the intraprotein Coulombic energy ($E_{MM\text{-}eel}$), and $\Delta G_{solv}$ one of 1.00 with its electrostatic term, the cost of charging the solute ($G_{pol}$). The correlation between $E_{MM\text{-}eel}$

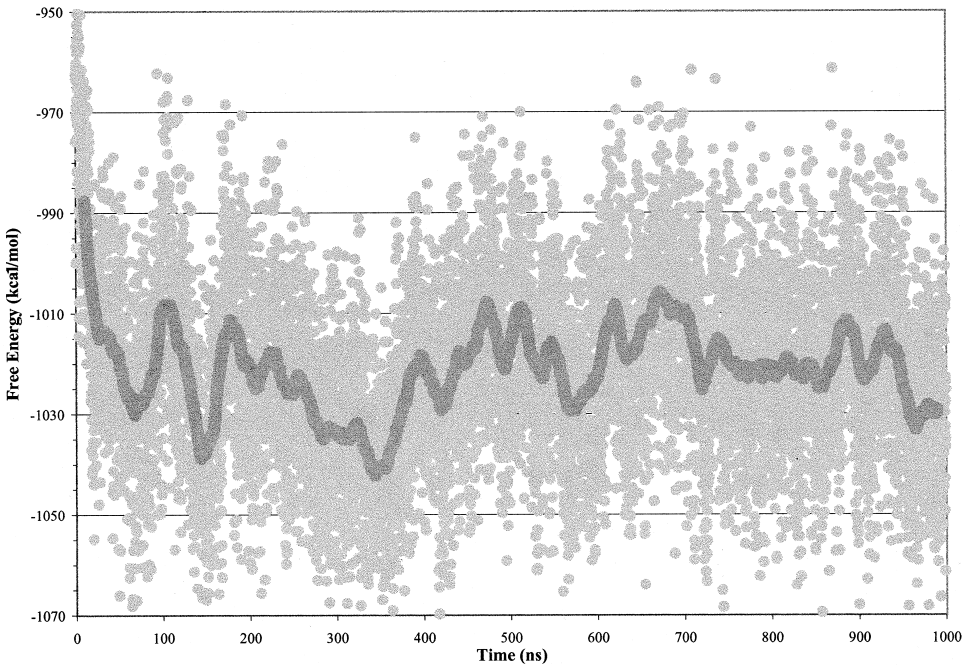


Fig. 1. MM-PB/SA free energy from the folding trajectory. The free energy was calculated once every 100 ps, a total of 10,000 times for one μs of data. For each Poisson-Boltzmann calculation, 1,000 iterations were used with grid spacing of 0.5 Å, PARSE radii, and Cornell et al. charges. The 20-ns running average of 100 ps time steps is shown as the darker solid line. The previously reported folding intermediate ensemble lies between 240 and 400 ns. We refer the structures from 500 to 1,000 ns as the "random coil ensemble" or the "unfolded state."

**TABLE I. Summary of Energies From MD Simulations on the Villin Headpiece[†]**

| Ensemble | ⟨vdW⟩ | ⟨eel⟩ | $\langle E_{int}\rangle$ | $\langle E_{MM}\rangle$ | $\langle \Delta G_{pol}\rangle$ | $\langle \Delta G_{NP}\rangle$ | $\langle \Delta G_{solv}\rangle$ | $\langle G_1\rangle^a$ | $\langle T \cdot S_{gas}\rangle$ $(n = 5)$ |
|---|---|---|---|---|---|---|---|---|---|
| Native | −104.6 | −731.9 | 582.0 | −254.5 | −821.9 | 18.7 | −803.2 | −1057.7 | 455.94 |
| (1–100 ns) | (11.1) | (31.5) | (16.1) | (35.7) | (32.2) | (0.7) | (31.9) | (15.9) | (3.8) |
| Folding intermediate | −105 | −720.1 | 600.6 | −224.6 | −824.6 | 17.7 | −806.8 | −1031.5 | 450.12 |
| (240–400 ns) | (14.9) | (30.1) | (17.6) | (38.0) | (32.1) | (0.9) | (31.6) | (15.7) | (2.0) |
| Unfolded | −65.6 | −660.9 | 598.3 | 128.2 | −912.9 | 21.7 | −891.2 | −1019.2 | 455.35 |
| (500–1,000 ns) | (17.5) | (50.9) | (16.8) | (61.4) | (55.7) | (1.7) | (54.5) | (17.2) | (5.6) |

[†]Values are given in kcal/mol with standard deviations in parentheses.
[a]$G_1 = E_{MM} + \Delta G_{solv}$.

and $G_{pol}$ is also strong with a coefficient of −0.97. As intrasolute electrostatic interactions are formed, $E_{MM\text{-}eel}$ and resultantly $E_{MM}$ decrease, whereas electrostatic interactions between solute and solvent are broken, and resultantly $G_{pol}$ and $\Delta G_{solv}$ increase. Thus, the solvation free energy and gas phase energy are inversely related because their preponderant terms are themselves inversely related. The causal factors for fluctuation of $\Delta G_{solv}$ and $E_{MM}$ are their electrostatic terms, whereas the causal factor for fluctuation in $G_1$ is the total electrostatics for the solvated system ($EEL_{tot}$), the sum of $E_{MM\text{-}eel}$ and $G_{pol}$.

Given that electrostatics provide the major source of fluctuation in the solvated protein system, a separate issue remains about whether $EEL_{tot}$ dictates the general trend of $G_1$. We find $EEL_{tot}$ over the course of the folding trajectory to have a correlation coefficient of only 0.30 with $G_1$, and the sum of the remaining non-electrostatic terms in the system (non-EEL) one of 0.77 with $G_1$. In addition, $EEL_{tot}$ in the folding trajectory has a much smaller standard deviation (16.2 kcal/mol) than non-EEL (27.3 kcal/mol). It is not the delicate balance between the sum of strongly opposing terms, $G_{pol}$ and $E_{MM\text{-}eel}$, that relates best to our estimate of the total free energy. Rather, it is the sum of all other terms not associated with electrostatics that drives the shape of the $G_1$ trajectory. This finding does not suggest that forces created by electrostatic interactions are a small contribution to the sum of all forces acting on a protein, that they do not drive the motion of the protein. What the variances in the distributions of $EEL_{tot}$ and non-EEL show are that the sum of electrostatics is much more constant, and what the correlation coefficients with $G_1$ show are that the sum of non-electrostatics is more responsible for changes in the free energy.

### Compact Structures Have Better Long-Range van der Waals Contacts

As seen in Table I, we consider the 100-ns native simulation as a single ensemble and have broken the folding trajectory into two further ensembles: folding intermediate and the unfolded (the last half μs of the trajectory). The one energy component that is similar for the native and intermediate states (≈ −105 kcal/mol) and significantly more favorable than in the ensemble of unfolded structures (≈ −66 kcal/mol) is $vdW_{tot}$. The *P*

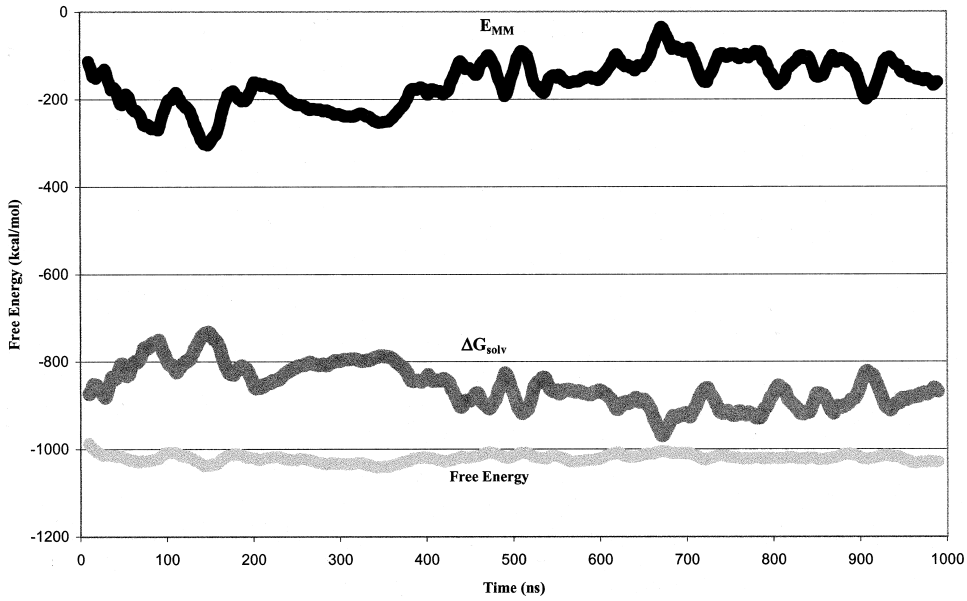


Fig. 2. Twenty-nanosecond running averages from the folding trajectory of the molecular mechanics energy ($E_{MM}$), the solvation free energy ($\Delta G_{solv}$), and the MM-PB/SA free energy. The free energy exhibits much less variation than $E_{MM}$ and $\Delta G_{solv}$, and the latter two are strongly inversely related.

value from a two-tailed Student's *t*-test between average values from native and folding intermediate is statistically insignificant (>0.05). In contrast, *P* values, even after the Bonferroni correction for multiple group comparison, between native and unfolded and between intermediate and the unfolded are highly significant (<0.0001). However, when looking at $EEL_{tot}$, the internal strain energy ($E_{int}$), and the energy of the non-polar first solvation shell effects ($\Delta G_{NP}$), none of the three pairwise comparisons of the three ensembles is significantly different for any of the three energies. In addition, we find $vdW_{1-4}$ to show virtually no fluctuation in any of our simulations, that the variance found in $vdW_{tot}$ is essentially identical to that of $vdW_{NB}$, which implies that it is the long-range van der Waals interactions (4 or more covalent bonds apart) that are more favorable in the native and intermediate states. This is reasonable because these two ensembles are more compact, and more favorable van der Waals interactions would be a rational causal factor that they might share in common.

Although the above shows that the two similarly compact native and folding intermediate states have dispersion energies ($vdW_{NB}$) that are similarly favorable over the less compact unfolded state, this does not imply that all states with native compactness will necessarily have dispersion energies as favorable as the native state. It is

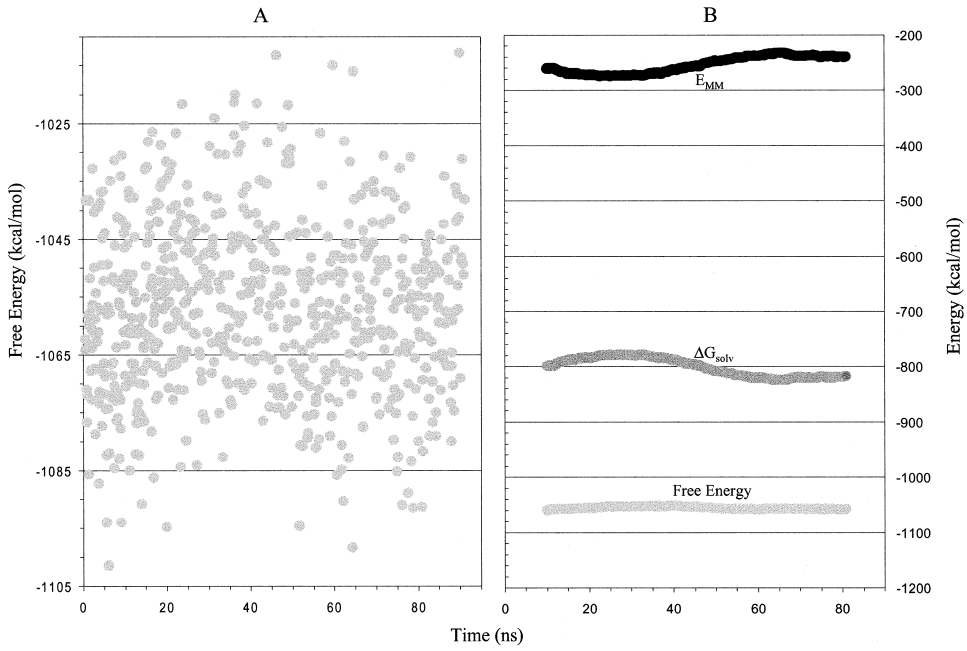


Fig. 3. Control simulation results showing (A) actual MM-PB/SA data and (B) 20 running averages. The 20-ns running average of the free energy remains relatively constant at −1,058 kcal/mol.

possible that the most highly compact structures will not have well-packed interiors and therefore higher than native dispersion energies. In this case, the native-like dispersion energies in the folding intermediate were accomplished only at the expense of internal strain energies (see below).

A more statistically meaningful way to associate van der Waals interactions with compactness is to look at the correlation coefficients between the $vdW_{NB}$ and some parameter that estimates degree of compactness, because this weights the relationship at every snapshot as opposed to comparing three group averages, which can only reveal if the relationship is direct or inverse. The most common measure for degree of compactness is the radius of gyration ($R_\gamma$), which calculates the root mean square deviation of each atom in a molecule from the center of mass. The correlation coefficient over the entire folding simulation is 0.82, and the correlation coefficient from the control simulation is 0.67, suggesting that compactness and long-range dispersive forces are indeed related, albeit more so in the less compact non-native structures. However, even when looking only at the most compact region of the folding simulation, the folding intermediate, the correlation coefficient is still much higher than the value in the native structures, 0.79. Thus, although the two similarly compact states have similarly favorable dispersion energies, the relationship between $R_\gamma$ and $vdW_{NB}$ is substantially lower in the native state. Among compact states, there can be a larger distribution of correlation coefficients than among unfolded structures. Perhaps this lesser degree of correlation in the native state can be explained if very well-packed protein side chains hinder deviations in the compactness from being accompanied by changes in $VDW_{NB}$. The hydrophobic core in the native state will have very favorable van der Waals contacts, and hence a reasonably constant $vdW_{tot}$ that will likely be less sensitive to the protein's periodic expansion and relaxation than that of the folding intermediate and unfolded ensemble whose side chains have more freedom to reorient themselves.

### The Native State Has Less Internal Strain Than Other Compact Structures

We use the same type of comparisons between native and other compact structures as we did between compact and non-compact structures in the previous section. Referring back to Table I, one can compare the energy terms between native and folding intermediate ensembles. $G_1$ is $\approx$27 kcal/mol lower and $E_{int}$ $\approx$19 kcal/mol lower in the native ensemble than in the folding intermediate. The difference seen in the averages of $E_{int}$ is highly significant with a Bonferroni-corrected $P$ value of $< 0.0001$. None of the other energy terms ($EE_{tot}$, vdW, and $\Delta G_{NP}$) are significantly different between the native and folding intermediate states.

By these group comparisons, it appears that $E_{int}$ has the greatest association with $G_1$ in the native and little association in all other ensembles. Again, correlation coefficients provide more information. In the unfolded ensemble (500 ns $-$ 1 $\mu$s), the coefficient between $E_{int}$ and $G_1$ was only 0.30, suggesting that they are relatively independent of one another. In the control simulation on the native ensemble, we observe a correlation coefficient of 0.73 and in the compact folding intermediate ensemble, a coefficient of 0.26. These data suggest that the biggest source of disparity between the native and other low $R_\gamma$ states in this study is $E_{int}$, the internal strain energy. Table II takes a more detailed look at the $E_{int}$ and finds that the major source of difference is the angle term.

**TABLE II. Summary of Internal Strain Energies From MD Simulations on the Villin Headpiece[†]**

| Ensemble | Bond | Angle | Dihedral | $E_{int}$ |
|---|---|---|---|---|
| Native | 104.5 | 274.7 | 202.8 | 582.0 |
| (1–100 ns) | (8.1) | (11.8) | (8.6) | (16.1) |
| Folding intermediate | 106.7 | 290.7 | 203.6 | 601.1 |
| (240–400 ns) | (8.6) | (12.5) | (10.4) | (15.8) |
| Unfolded | 105.9 | 285.9 | 206.7 | 598.5 |
| (500–1,000 ns) | (8.2) | (12.5) | (9.4) | (15.5) |
| (Intermediate)–(Native) | 2.2 | 16.0 | 0.8 | 19.0 |
| (Unfolded)–(Intermediate) | −0.8 | −4.8 | 3.1 | −2.6 |

[†]Values are given in kcal/mol, with standard deviations in parentheses.

### Entropic Contributions to the MM-PB/SA Method

Table I shows that vibrational entropy does not differ by much among the native, compact and unfolded states and that the $T \times S_{solute}$ term does not appear to be any more or less favorable in any of the states. The $P$ values for the three pairwise comparisons are all $>0.05$ and thus not statistically different. This is in agreement with the findings of Hermans comparing folded and misfolded protein structures by using the harmonic approximation from the covariance matrix of the positional fluctuations during the dynamics trajectory,[19] and of our group comparing different forms of nucleic acids.[16–18] Neither method is particularly accurate, but both show that the $T \times S_{solute}$ term is comparable for various similar "structures" of small proteins.

The native and random coil intrinsic "vibrational" entropies are similar, but it is the entropy associated with the hydrophobic effect that is represented in the $\Delta G_{NP}$ term. As expected, this part of the solvation free energy is least favorable in the unfolded states, which also has the highest $R_\gamma$. This $\Delta G_{NP}$ term makes the unfolded states, from 500 ns to 1 $\mu$s, about 3 kcal/mol less favorable than the native ($P < 0.0001$). However, $\Delta G_{NP}$ in the folding intermediate is about 1 kcal/mol more favorable than in the native ensemble ($P < 0.0001$). As should be expected with a simple linear relation, the fact that the folding intermediate is (statistically) significantly more favorable than the native (albeit by only 1 kcal/mol) shows that the limitations of this term arise from the uncertainty in the function, not from sample size.

### Estimating the Conformational Entropy of the Denatured State

The free energy of denaturation ($\Delta G_{denat}$) for small proteins has been estimated to fall between 5 and 10

kcal/mol.[25–28] MacKnight linearly extrapolated a series of guanidine hydrochloride (GuHCl) denaturations of villin to 0 M GuHCl at pH = 5.4 and 4°C and estimates $\Delta G_{denat}$ to be 3.3 (±0.4) kcal/mol. This may at first seem to be inconsistent with our $\Delta G_1$ between native and unfolded villin estimates. However, until now we have been considering the free energies of the individual snapshots and not the entropy associated with considering all the unfolded states as a conformational ensemble, the conformational entropy. In fact, the experimental $\Delta G_{denat}$ can be used in conjunction with $\Delta G_1$ to estimate the effective conformational degeneracy of this ensemble. Assuming a Boltzmann distribution of the two-state model, the number of individual denatured conformations, i.e., the degeneracy of the denatured state ($\Omega_{denat}$), can be estimated as follows:

$$P(\text{denat})/P(\text{native}) = e^{(-\Delta G_{denat}/RT)} = \Omega_{denat}/\Omega_{nat} \cdot e^{(-\Delta G_1/RT)} \quad (3)$$

where $\Delta G_1$ is an average effective $G_1$ difference between native and denatured states. Assuming the degeneracy of the native state ($\Omega_{nat}$) is unity,

$$\Omega_{denat} = e^{[(\Delta G_1 - \Delta G_{denat})/RT]} \quad (4)$$

$$S_{conf} = R \ln \Omega_{denat} = (\Delta G_1 - \Delta G_{denat})/T \quad (5)$$

$$\ln \Omega_{denat} = \Delta\Delta G/RT \quad (6)$$

The total conformational entropy associated with the degeneracy of the denatured state ($S_{conf}$) and hence the log of $\Omega_{denat}$ are directly proportional to $\Delta\Delta G$, the difference between protein stability, $\Delta G_{denat}$, and the effective $G_1$ difference between a typical native and denatured snapshot, $\Delta G_1$.

From looking at the free energy during the microsecond trajectory, it is not clear whether the free energy difference between the native conformation and the folding intermediate (≈26.2 kcal/mol) should be used as $\Delta G_1$ or if instead the free energy difference between the native and the average non-compact states (≈38.5 kcal/mol) is more appropriate. If the compact states are all of approximately similar free energy and together represent the dominating configuration of the denatured state, then we would use a $\Delta G_1$ = 26.2 kcal/mol, giving us the smallest estimate for $\Delta\Delta G$ of 22.9 kcal/mol. This translates to a lower bound degeneracy estimate on the order of $3.8 \times 10^{16}$. If on the other hand, the non-compact random coil configurations are Boltzmann-weighted far more than the compact ones, then we would approximate $\Delta G_1$ = 38.5 kcal/mol. This leads to upper bound estimates of $\Delta\Delta G$ = 35.2 kcal/mol and of $\Omega_{denat} = 3.0 \times 10^{25}$. The value for $\Omega_{denat}$ can then be converted into another interesting value, an average number of degrees of freedom per residue (y): $y^{36} = \Omega_{denat}$. Our range of estimates for $\Omega_{denat}$ corresponds to a range of y values from 2.9 to 5.1, which is in qualitative agreement with Dill's estimates[29] for y.

## Estimating the Free Energy of Unfolding

In the previous section, we attempted to use the experimental $\Delta G_{denat}$ together with our $\Delta G_1$ to estimate the degeneracy of the non-native state. Alternatively, we can use $\Delta G_1$ together with other degeneracy estimates to obtain a free energy of unfolding and compare that directly with the experimental value. Karplus estimates that in a 27-mer small protein, there is a mixture of $10^{10}$ "semi-compact globule" conformations and $10^{16}$ random coil conformations,[25] which correspond to y values of 2.3 and 3.9, respectively, and $\Omega_{denat}$ values for a 36-mer of $2.2 \times 10^{13}$ and $2.2 \times 10^{21}$, respectively. By splitting the denatured state probability shown in Equation (3) into a sum of two probabilities, again assuming $\Omega_{nat}$ is unity, Karplus's estimates for the degeneracy of compact ($\Omega_{compact}$) and random coil ($\Omega_{RC}$) states can be used together with the respective $\Delta G_1$ predictions in this study ($\Delta G_{1,compact}$, and $\Delta G_{1,RC}$),

$$P(\text{denat})/P(\text{native}) = P(\text{compact})/P(\text{native}) + P(\text{random coil})/P(\text{native}) = \Omega_{compact} \cdot e^{(-\Delta G_{1,\,compact}/RT)} + \Omega_{RC} \cdot e^{(-\Delta G_{1,\,RC}/RT)} \quad (7)$$

This results in a $\Delta G_{denat}$ of 7.7 kcal/mol with 89% of the denatured state being a compact structure and 11% random coil; thus, the total error is 4.4 kcal/mol.

## Estimating the Folding Kinetics

Figure 4 attempts to separate the noise of the folding trajectory by looking at the 200-ns running average (thick line). This plot captures the folding intermediate and suggests that a transition state might lie around the 700-ns mark. If this is the case and villin continues to undergo first-order exponential decay toward its native state without encountering any further kinetic traps, extrapolating out the smoothed plot leads to a half time for folding of 1.05 μs ($k_f = 6.6 \times 10^5$), leading to a total time from the denatured state to 90% "folded" of 4.2 μs.

Plaxco et al.[30] summarized the intrinsic folding rates for a set of 12 non-homologous, simple, single-domain proteins and looked at their relationship with size, stability and topology of the proteins. They found that size and stability have weak or non-existent relationships with $\ln(k_f)$, but that the relative contact order (CO), which reflects the relative amount of local and non-local contacts in a protein's native structure, shows a strong inverse correlation ($R = -0.81$). CO is the average sequence separation distance between all non-hydrogen atoms that are within 6 Å, normalized by the sequence length. CO for the average, minimized, NMR villin structure[20] comes out to 11.0%. Our estimate for villin, from the extrapolation of the folding trajectory, of $\ln(k_f) = 13.4$ is consistent with the data presented by Plaxco et. al.[30]; adding these data actually increases the magnitude of the correlation coefficient between CO and $\ln(k_f)$ to −0.84. Although the extrapolation leads to strong agreement with the estimate that would arise from villin's CO, there are a number of assumptions made, the two most important are the following.

First, the extrapolation in Figure 4 assumes that folding will proceed without going through any further metastable intermediates, such as the one found between 240 and 400

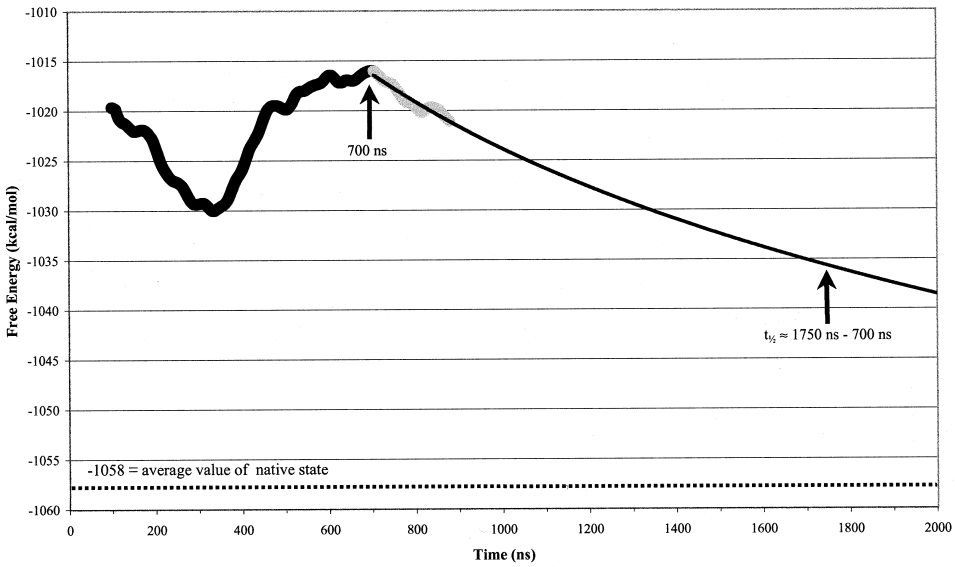


Fig. 4. A possible depiction for the folding pathway of the villin headpiece. Thick line represents a 200-ns running average of the folding simulation's MM-PB/SA free energy. The thin line is an extrapolation from a putative transition state at 700 ns, using first-order linear kinetics to describe folding as a cooperative process without additional barriers on its way to the native state. The dotted line at the bottom is the reference value from the control simulation. This leads to a half time of folding from a transition state of $\approx 1$ μs.

ns. In our previous article,[11] it was concluded, based on villin's CO of 11.0% and the Plaxco/Baker CO least squares line, that villin folds on a 10- 100-μs timescale. There, we suggested that the villin headpiece may continue to fall into metastable intermediates until it found one that was close enough to the more stable native structure to allow it to reach the native state by further subtle readjustments of the structure. Again approximating the 90% "folded" state as native, this range of folding times (10–100μs) would correspond to a range for the half time of folding between 3.3 and 33.3 μs, each of which also increases the correlation between CO and $\ln(k_f)$ to $-0.84$ and $-0.85$, respectively. At this point, it is not clear which picture is correct.

Second, although the native structure is of significantly lower free energy according to the combined molecular mechanical/continuum model (MM-PB/SA) than anything found in the folding trajectory so far, it is not known at this point how closely this free energy model can reproduce the "true" native global free energy minimum of villin.

## CONCLUSIONS

The recent completion of a 1-μs folding simulation has allowed us to demonstrate that the MM-PB/SA method can successfully identify the native conformation from other compact structures in a small, single-domain protein, the villin headpiece. As the folding trajectory formed the very compact intermediate, the biggest change in energy was a drop in the dispersion energy to levels as favorable as in the native state, and during this simulation we found a high correlation between the dispersion energy and Rγ. However, the folding intermediate was only able to accomplish such favorable van der Waals contacts at the expense of exhibiting more internal strain, particularly in the angle term, which was the key term more favorable in the native state than in the intermediate.

The differences in MM-PB/SA free energies of villin between native and the non-native structures, combined with the estimated free energy of unfolding, leads to an estimate of the conformational degeneracy in the non-native state between $10^{16}$ and $10^{25}$ or an average number of conformations per residue between 2.9 and 5.1. Smoothing the energies over a large window leads to an apparent transition state for folding at 700 ns in the trajectory. If one assumes no further kinetic traps, our estimate is that it may take an additional 3.5 μs to fold villin from this point.

## REFERENCES

1. Dill KA, Chan HS. From Levinthal to pathways to funnels. Nat Struct Biol 1997;4:10–19.
2. Dill KA, Bromberg S, Yue K, Feibig KM, Yee DP, Thomas PD, Chan HS. Principles of protein folding—a perspective from simple exact models. Protein Sci 1994;4:561–602.
3. Skolnick, J, Kolinski A, Ortiz AZ. MONSSTER: a method for folding globular proteins with a small number of distance restraints. J Mol Biol 1997;265:217–41.
4. Kolinski A, Skolnick J. Lattice Models of Protein Folding, Dynamics and Thermodynamics. Austin: R. G. Landes Company, 1996.
5. Alonso DO, Daggett V. Molecular dynamics simulations of protein unfolding and limited refolding: characterization of partially unfolded states of ubiquitin in 60% methanol and in water. J Mol Biol 1995;247:501–20.
6. Alonso DO, Daggett V. Molecular dynamics simulations of hydrophobic collapse of ubiquitin. Protein Sci 1998;7:860–74.
7. Li A, Daggett V. Molecular dynamics simulation of the unfolding of barnase: characterization of the major intermediate. J Mol Biol 1998;275:677–94.
8. Tirado-Rives J. Orozco M, Jorgensen WL. Molecular dynamics simulations of the unfolding of barnase in water and 8 M aqueous urea. Biochemistry 1997;36:7313–29.
9. Lazaridis T, Karplus M. "New view" of protein folding reconciled with the old through multiple unfolding simulations. Science 1997;278:1928–1931.
10. Duan Y, Wang L, Kollman PA. The early stage of folding of villin headpiece subdomain observed in a 200-nanosecond fully solvated molecular dynamics simulation. Proc Natl Acad Sci USA 1998;95: 9897–9902.
11. Duan Y, Kollman PA. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science 1998;282:740–744.
12. Moult J, Hubbard T, Bryant SH, Fidelis K, Pedersen JT. Critical assessment of methods of protein structure prediction (CASP): round II. Proteins 1997; Suppl 1:2–6.

13. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 1999;34:82–95.
14. Novotný J, Bruccoleri R, Karplus M. An analysis of incorectly folded protein models. Implications for structure predictions. J Mol Bio 1984;177:787–818.
15. Martin ACR, MacArthur MW, Thorton JM. Assessment of comparative modeling in CASP2. Proteins 1997;Suppl, 1:14–28.
16. Srinivasan J, Miller J, Kollman PA, Case DA. Continuum solvent studies of the stability of RNA hairpin loops and helices. J Biol Struct Dyn 1998;16:671–82.
17. Cheatham TE 3rd, Srinivasan J, Case DA & Kollman PA. Molecular dynamics and continuum solvent studies of the stability of polyG-polyC and polyA-polyT DNA duplexes in solution. J Biol Struct Dyn 1998;16:265–80.
18. Srinivasan J, Cheatham TE 3rd, Cieplak P, Kollman PA & Case DA. Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate-DNA helices. J Am Chem Soc 1998;120:9401–9409.
19. Vorobjev YN, Almagro JC, Hermans J. Discrimination between native and intentionally misfolded conformations of proteins: ES/IS, a new method for calculating conformational free energy that uses both dynamics simulations with an explicit solvent and an implicit solvent continuum model. Proteins 1998;32:399–413.
20. McKnight CJ, Doering DS, Matsudaira PT, Kim PS. A thermostable 35-residue subdomain within villin headpiece. J Mol Biol 1996;260:126–134.
21. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM Jr, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 1995;117:5179–5197.
22. Sanner MF, Olson AJ, Spehner JC. Reduced surface: an efficient way to compute molecular surfaces. Biopolymers 1996;38:305–20.
23. Gilson MK, Honig B. Calculation of the total electrostatic energy of a macromolecular system: solvation energies, binding energies, and conformational analysis. Proteins 1998;4:7–18.
24. Sitkoff D, Sharp KA, Honig B. Accurate calculation of hydration free energies using macroscopic solvent models. J Phys Chem 1994;98:1978–1988.
25. Sali A, Shakhnovich E, Karplius M. How does a protein fold? Nature 1994;369:248–251.
26. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins 1995;21:167–195.
27. Onuchic JN, Wolynes PG, Luthey-Schulten Z, Socci ND. Toward an outline of the topography of a realistic protein-folding funnel. Proc Natl Acad Sci USA 1995;92:3636–3630.
28. Pande VS, Grosberg AY, Tanaka T. On the theory of folding kinetics for short proteins. Fold Des 1997;2:109–114.
29. Dill KA. Theory for folding and stability of globular proteins. Biochemistry 1985;24:1501–1509.
30. Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 1998;277:985–994.