Structure, Vol. 9, 905–916, October, 2001, ©2001 Elsevier Science Ltd. All rights reserved. PII S0969-2126(01)00660-8

Free-Energy Calculations Highlight Differences in Accuracy between X-Ray and NMR Structures and Add Value to Protein Structure Prediction

Matthew R. Lee[1,2] **and Peter A. Kollman**[3]
Department of Pharmaceutical Chemistry
University of California, San Francisco
San Francisco, California 94131

Summary

Background: While X-ray crystallography structures of proteins are considerably more reliable than those from NMR spectroscopy, it has been difficult to assess the inherent accuracy of NMR structures, particularly the side chains.

Results: For 15 small single-domain proteins, we used a molecular mechanics-/dynamics-based free-energy approach to investigate native, decoy, and fully extended alpha conformations. Decoys were all less energetically favorable than native conformations in nine of the ten X-ray structures and in none of the five NMR structures, but short 150 ps molecular dynamics simulations on the experimental structures caused them to have the lowest predicted free energy in all 15 proteins. In addition, a strong correlation exists ($r^2 = 0.86$) between the predicted free energy of unfolding, from native to fully extended conformations, and the number of residues.

Conclusions: This work suggests that the approximate treatment of solvent used in solving NMR structures can lead NMR model conformations to be less reliable than crystal structures. This conclusion was reached because of the considerably higher calculated free energies and the extent of structural deviation during aqueous dynamics simulations of NMR models compared to those determined by X-ray crystallography. Also, the strong correlation found between protein length and predicted free energy of unfolding in this work suggests, for the first time, that a free-energy function can allow for identification of the native state based on calculations on an extended state and in the absence of an experimental structure.

Introduction

While methods for the experimental determination of protein structure have had an enormous impact on the study of molecular action, protein design, and interpretation of chemical, kinetic, or thermodynamic experiments, they are often quite challenging. Elucidation of a protein structure by X-ray crystallography demands a supersaturated concentration, which can usually only be achieved upon the addition of agents that compete with the protein for water [1]. These foreign agents and packing effects of crystallization itself can induce structural defects [1]; while this artifactual information is reported with the structure in known instances, it is not possible to realize all of the errors caused by these model-specific systematic limitations. Another pitfall of crystallography occurs on segments having very low or nonexistent electron densities, which presumably contain highly disordered atoms that are in motion and thus difficult to detect in the time scale of crystallography [1]. Additionally, oxygen, nitrogen, and carbon atoms usually cannot be distinguished from one another [2]. Other, smaller deviations almost certainly exist in all X-ray structures due to differences between the crystal environment, which is only 50% aqueous by volume [1, 2], and the natural surroundings; this fundamental difference between crystal and native structures, as well as the non-static nature of proteins, creates an average atomic uncertainty of around 0.5 Å in structures, with the best data.

In comparison, protein structures solved by nuclear magnetic resonance (NMR) are completely solvated and free of the constraints of a crystal lattice; this allows for better description of the inherent flexibility, with the protein in surroundings much closer to what it actually experiences under physiological conditions. However, despite the more realistic environment that NMR structures experience, they are inherently less reliable than X-ray data because crystallographic models contain far more experimental data per atom. Differences among the various models of an NMR ensemble are usually much greater than 1 Å, often 2 Å. Determination of a protein structure by NMR involves a refinement process, usually starting from a randomly generated conformation that satisfies some local distance constraints. The determination then proceeds with a sampling protocol that attempts to satisfy as many nuclear Overhauser effects (NOEs) as possible until a point is inevitably reached where the structure is incapable of being improved further [1]. While more NOEs generally allow for more accurate structures, we suggest that the shortcomings of the refinement stage are what preclude greater precision in the method. Probably, the most severe approximation made during the refinement stage of NMR structure determination is an inaccurate representation of the solvent. This systematic error in the energy potential can prevent finding a solution with lower positional inaccuracies, even if the refinement stage were capable of exploring every possible conformation. Generating tens of structures with low average rmsd values compared to those of the mean structure does not necessarily imply *accuracy*. This only implies that there is less *uncertainty* in each NMR model having satisfied both the NOE constraints and the flawed energy potential, which is likely flawed due to the highly simplistic treatment of solvent. In the vast majority of NMR structures, inclusion of solvent effects is accom-

[1] Correspondence: matthew.lee@lionbioscience.com
[2] Present address: Lion Bioscience, 9880 Campus Point Drive, San Diego, California 92121.
[3] Deceased.

Key words: decoys; MM-PBSA; ranking

plished by the use of a distance-dependent dielectric constant in the Coulombic term of the potential energy function, and it is thus not very accurate.

To take a step toward understanding some of the qualitative differences between NMR and X-ray structures, we investigated the Molecular Mechanics-Poisson Boltzmann/Surface Area (MM-PBSA) [3] free energies of X-ray and NMR structures, before and after short, computationally inexpensive molecular-dynamics simulations, in comparison to large sets of decoy conformations on a total of 15 small, single-domain proteins. Sets of decoys for eight proteins came from the "Rosetta All Atom Decoy Set" [4, 5], and those for seven proteins came from the Park & Levitt four-state reduced decoy set [6, 7]. While it is widely believed that the native structure lies at the global free-energy minimum [8], which would satisfy the demands of thermodynamics, alpha-lytic protease has recently emerged as an exception, with the native state exhibiting a half life of unfolding on the order of 1 year [9]. Because proteins are translated sequentially, it is not surprising that kinetic traps govern the overall structure in some cases. However, we expect that the proteins, in the majority of cases, obey macroscopic thermodynamics, with the native state lying at the global free-energy minimum, irrespective of whether the native structure has been solved by X-ray crystallography or NMR spectroscopy. At the very least, the native state should have a free energy substantially lower than unfolded and poorly folded conformations. This work suggests that NMR structures can benefit significantly from short, aqueous molecular-dynamics simulations and that free-energy calculations can be used to identify the native state in the absence of an experimental structure.

Results

Decoys Compared to Crystal Structures

The four-state reduced decoy set [6, 7] consists of approximately 650 conformations for seven proteins, with each conformer differing from the native conformation at ten specific dihedral angles that always lie in regions between or at the ends of secondary-structure elements. Each dihedral may adopt only one of four possible discrete values, leading to an exhaustive enumeration of 1,048,576 (4^{10}) possible conformations per protein, of which approximately 650 were physically reasonable after the removal of those with steric conflicts and unreasonably extended chains. Thus, the decoys for any given protein differ only in their tertiary structure but cover a wide range of native similarity in terms of tertiary structure. Three of these proteins are purely alpha, and the other four are mixed alpha/beta, with the native counterpart being an X-ray structure in all seven cases.

For each protein in the four-state reduced set, we performed single-point minimization MM-PBSA calculations on all the decoys, on the initial crystal structure, and on a 150 ps snapshot from an explicit-solvent molecular-dynamics simulation that started with the minimized crystal structure. The MM-PBSA free energy is simply the sum of an internal energy, as determined by the AMBER force field, and a solvation free energy, based predominantly on DelPhi's calculation of the Poisson equation (see Experimental Procedures). Figure 1 shows the resulting MM-PBSA free energies as a function of Cα rmsd. This free energy function does better than any of the 18 scoring functions studied by Park et al. (1997) [10] and at least as well as other recently reported physically based functions that have successfully examined this same decoy set [11–13]. The crystal structures, shown as the gray tube diagrams on the pictorial inlays and represented by the solid red circles, have lower, more favorable free energies than all of the decoys in six out of the seven proteins, with the crystal structure coming out third best among 654 decoys on 3icb, although, even for this protein, the best structure with MM-PBSA had a Cα rmsd of only approximately 1 Å from the native.

The Z score has been widely used for evaluating the goodness of a protein structure scoring function [14]. However, good Z scores, which are the number of standard deviations separating the native from the rest of the population (see Experimental Procedures), only imply that the native structure receives a much better score than the average score of all the conformers in the decoy set. Table 1 shows the X-ray rank results of two distance-dependent contact potential-energy functions from Park et al. (1997) [10] that were among the four best (in terms of average Z scores of all the proteins in the four-state reduced set) out of the 18 functions investigated. These results are shown alongside the X-ray rank results from MM-PBSA and its van der Waals component alone (VDW). While the average Z scores are comparable in each of the four, VDW(MJ) clearly does a relatively poor job in picking out the crystal structure as best. Our VDW correctly identifies all seven crystal structures, MM-PBSA identifies six out of seven, VDW(MJ)12 correctly predicts four out of seven, and VDW(MJ) does not predict any correctly. These X-ray rank results indicate that energy functions, which result in good Z scores, are not necessarily good at correctly identifying the native fold.

The Z score also does not address the strength of the relationship between native similarity and the scoring function. Instead, correlation coefficients provide a far more direct criterion for establishing the strength of association between two variables and are thus more fitting for judging the predictive value of a scoring function for structure prediction. For parametric samples, in which both variables are normally distributed on an interval scale, which implies a linear relationship, the standard (Pearson product-moment) correlation coefficient (r) is most appropriate, but for nonlinear relationships on an ordinal scale, in which one or both variables are not normally distributed, the Spearman rank correlation coefficient (r_s) is most appropriate. In a Boltzmann distribution, conformations are weighted exponentially according to their free energies, $P(i) = \exp(-\Delta E_i/RT)$, where ΔE_i is the difference in free energies between two states, *i* and some reference, such as the native state. If one subscribes to the common notion that the vast majority of proteins obey microscopic thermodynamics, one should expect that protein conformations roughly populate in a Boltzmann distribution, rather than a Gaussian distribution. Thus, for evaluating the strength of association between any variable and a free energy such as

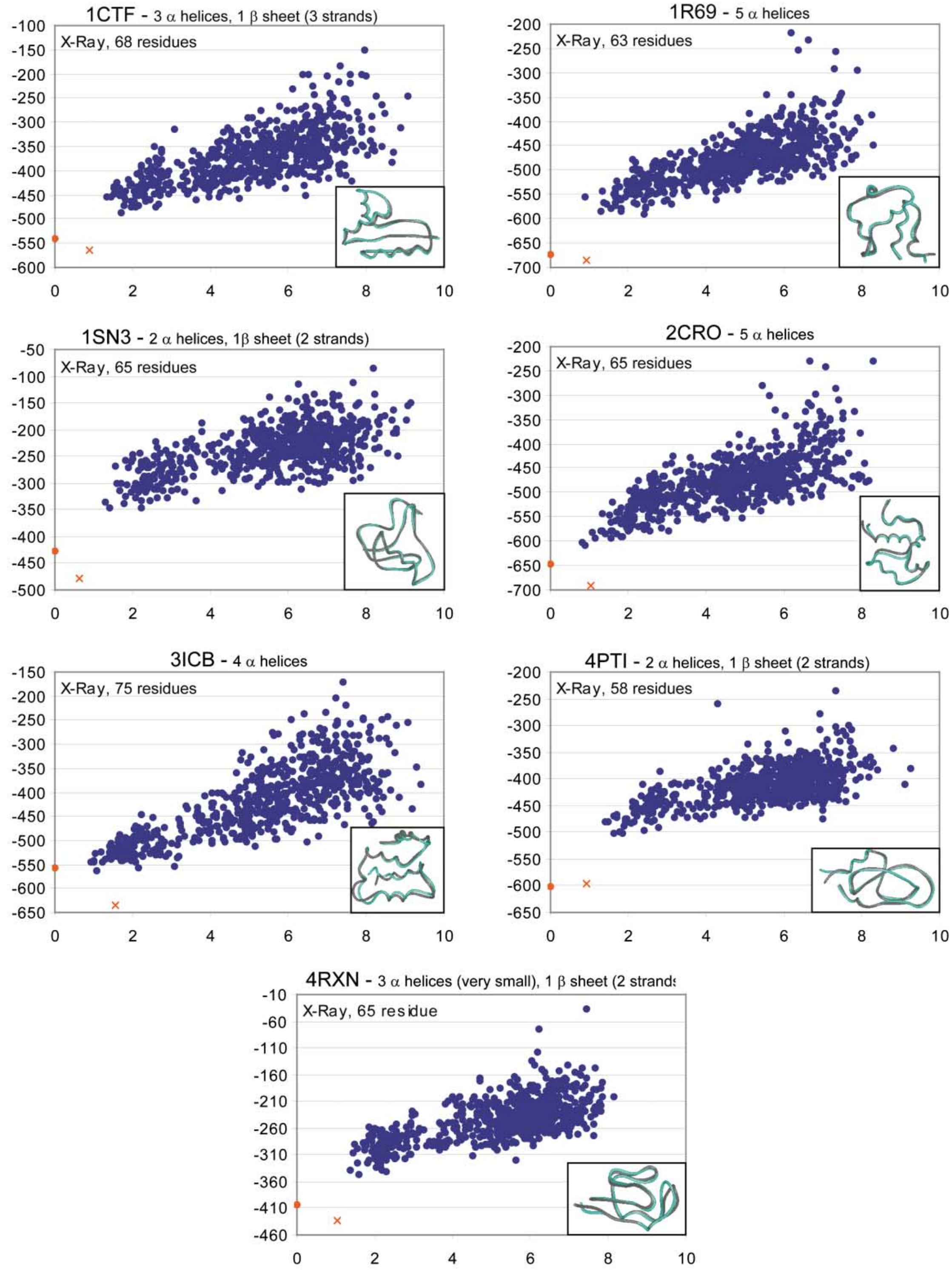

Figure 1. Single-Point MM-PBSA Energy on "Park & Levitt Four-State Reduced Decoy Set"

Single-point minimization MM-PBSA (*y* axes) versus Cα rmsd (*x* axes) on crystal structures and on decoys containing native secondary structure (Park and Levitt four-state reduced set). Each blue dot represents a single decoy. There were approximately 650 decoys for each of the seven proteins. Red circles are minimized X-ray crystal structures. Red exes are crystal structures that have been minimized after 150 ps of molecular dynamics in explicit solvent. Figure inlays contain an overlap of the crystal structure before (gray tube) and after (cyan tube) dynamics.

Table 1. X-Ray Rank among Park and Levitt Set

Protein	VDW(MJ)[1]	VDW(MJ)12[1]	VDW[2]	MM-PBSA[3]
1ctf	2	1	1	1
1r69	77	1	1	1
1sn3	2	1	1	1
2cro	160	1	1	1
3icb	1327	3	1	3
4pti	286	4	1	1
4rxn	49	3	1	1
<Z>	−3.95	−3.98	−3.92	−3.57

[1] VDW(MJ) is a distance-dependent contact potential, and VDW(MJ)12 is the same, but with a sharper repulsive term. The results of these energy functions are taken from Park et al., 1997.
[2] VDW is the attractive dispersion energy between nonbonded atoms in the MM-PBSA calculation.
[3] MM-PBSA is described in the Experimental Procedures.

MM-PBSA, r_s is more appropriate than r. For predictive value in protein structure prediction, a strong correlation with native similarity is highly desired, so we evaluated the Spearman rank correlation between MM-PBSA and Cα rmsd in Table 2, which shows a reasonably good correlation, slightly better than that reported by Gatchell et al. [13] and by Dominy and Brooks [12]. While Table 1 indicates the lack of association between good Z scores and the ability to correctly identify the native fold, Table 2 shows that good Z scores do not imply good predictive value. Although the VDW potential did slightly better than MM-PBSA in terms of Z score, it has no meaningful relationship with Cα rmsd, as illustrated in Figure 2 on two representative proteins and quantified in Table 2. Eight angstrom conformations have the same VDW energy as 2 Å structures

While some have suggested that there is no physical requirement for a relationship between free energy and native similarity [15], Dill and Chan popularized the now widely accepted view of a funnel-shaped free-energy landscape [16–18] to describe proteins. In this view, the native state has the lowest free energy, and the more distant the native similarity, the less favorable the free energy. If the free-energy landscape is indeed globally convex, the relationship between native similarity and free energy should be approximately linear for only those conformations immediately surrounding the native state, and the farther structures lie from the native state, the less linear the relationship should be, until finally on the level surface of the funnel, where native similarity would be very low, there should be no relationship at all. We investigated this by separating the four-state reduced decoy sets into three bins of native similarity: close (<2.5 Å), medium (2.5–5.0 Å), and distant (>5 Å). For each similarity bin, we evaluated the Pearson product-moment correlation coefficient, which again is the strength of the relationship between two variables that are *linearly* related. As expected, the close structures showed the greatest degree of linear association between Cα rmsd and MM-PBSA ($r = 0.64$), with the distant structures showing only a slight tendency, and medium structures falling in between (Table 3). Together with the rank correlation results in Table 2, these results suggest that the free energy and native similarity are related on an ordinal scale, with that relationship becoming increasingly linear as native similarity increases.

Table 2. Assessing Predictive Value of Energy Functions

	MM-PBSA		VDW	
Protein	r_s[1]	Z[2]	r_s[1]	Z[2]
1ctf	0.77	−2.47	−0.18	−3.36
1r69	0.55	−3.88	−0.27	−5.01
1sn3	0.52	−4.57	−0.32	−3.97
2cro	0.66	−3.03	−0.03	−4.57
3icb	0.75	−1.86	0.13	−3.58
4pti	0.44	−5.21	−0.37	−3.27
4rxn	0.65	−4.00	−0.48	−3.66
Average	0.62	−3.57	−0.22	−3.92

[1] r_s is the Spearman rank correlation coefficient between Cα rmsd and the energy; it is more meaningful than the standard Pearson product-moment correlation coefficient in nonparametric relationships that are not linearly related.
[2] Z score is the number of standard deviations separating the energy of the native conformation from the average energy of the entire set. (see Experimental Procedures)

The Rosetta “All Atom Decoy Set” consists of 1000 decoy conformations for each protein, with each conformer generated by the Baker group in the same manner as that used for the 1998 community-wide Critical Assessment of Structure Prediction experiment (CASP III) [5]. Three of the eight that we investigated from this decoy set had crystal structures. In contrast to the four-state reduced set, the Rosetta set usually does not populate the low Cα rmsd regions very well, which should lead to a limited relationship at best between functions with good predictive value and Cα rmsd among these decoys because, as discussed earlier, the linear correlation falls off beyond the 5 Å mark (Table 3). Furthermore, because the structures in this set differ from one another immensely more than they do in the four-state reduced set, where 10 dihedral angles are the only degrees of freedom, the noise of the energy should be much greater. Thus, we cannot hope to distinguish 8 Å from 15 Å structures, even with a free-energy function that was entirely precise. All that can be hoped for in this Rosetta decoy set is the ability to distinguish native structures from everything else, which MM-PBSA effectively accomplishes (Figure 3).

Decoys Compared to NMR Structures

The five other proteins that we investigated in the Rosetta decoy set had NMR structures. What clearly distinguishes these five, shown in Figure 4, from the 10 sets with crystal structures is that the minimized NMR structures (closed red circles) do not have the lowest free energies in any of the proteins. For 1gb1, there is a 6 Å decoy lower in free energy, a 16 Å one for 1ksr, a 9 Å one for 1res, an 18 Å one for 1tit, and a 17 Å one for 1wiu. We presume that this arises because of the unsophisticated refinement methods used for solving the NMR structures, as discussed above. While these reported NMR structures are undoubtedly in the correct global structural fold, these single-point minimization MM-PBSA results suggest that the NMR structures are nowhere near the bottom of the native energy basin, that minimization

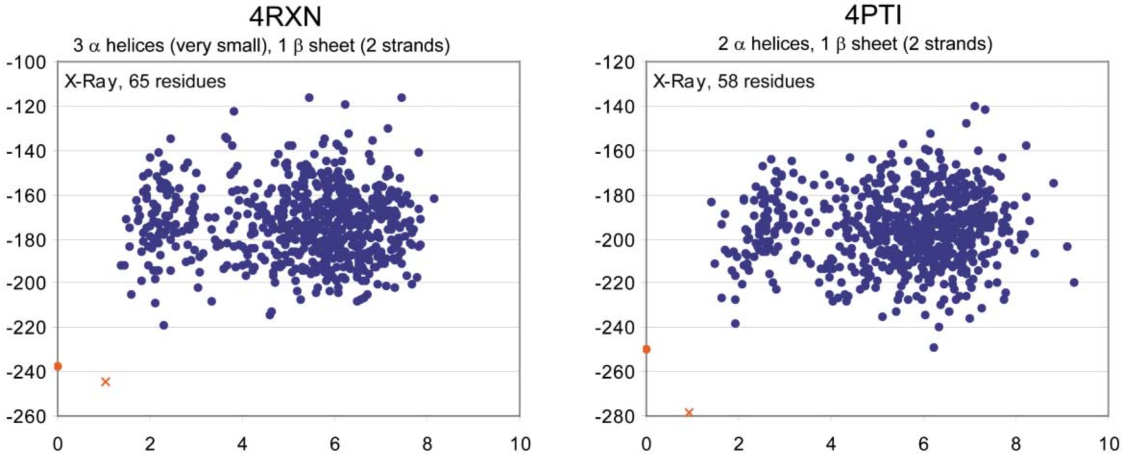

Figure 2. Single-Point Van der Waals Energy on "Park & Levitt Four-State Reduced Decoy Set"

Single-point minimization VDW (*y* axes) versus Cα rmsd (*x* axes) on crystal structures and on decoys containing native secondary structure (Park & Levitt four-state reduced set). Only two representative proteins are shown, demonstrating the lack of a relationship between native similarity and van der Waals energies, despite identification of native folds from all decoys. Red circles are minimized X-ray crystal structures. Red exes are crystal structures that have been minimized after 150 ps of molecular dynamics in explicit solvent.

alone is insufficient to overcome the numerous bad contacts, bond lengths, angles, and dihedrals, which additively can lead to many tens to hundreds of kcal/mol penalties, with only minor perturbations to the correct native topology and structure, in terms of rmsd.

Effect of Molecular Dynamics

Figures 1, 3, and 4 also show the effect of molecular dynamics on the native structure compared to single-point minimization MM-PBSA calculations. Experimental structures that have undergone 150 ps of molecular dynamics, followed by minimization, are shown as the cyan tube diagrams on the inlays and are represented by the red exes. These native 150 ps snapshots have the best free energies in all ten of the X-ray examples (seven from the four-state reduced model and three from Rosetta), including 3icb, where the minimized X-ray structure ranked third best, and the best free energies in four of the five NMR examples where none of the minimized NMR structures ranked best. In the 150 ps snapshot of 3icb, the region that deviated most from the crystal was one of two Ca^{2+} binding loops in the protein. While the 3icb deposited Protein DataBank structure contains hetero-atom records for two Ca^{2+} ions, the structures in the decoy set do not, so to be consistent with the decoy set and have a level playing field, we removed these divalent cations from the crystal structure prior to evaluating the single-point minimization MM-PBSA. We thus created a locally unfavorable hole in the system, which was filled in the 150 ps snapshot. Because hetero-atoms are not included in structure predictions, it is appropriate to remove them from the experimental structures as well when one is trying to evaluate a scoring function's ability to pick out the native conformation. This leaves crystal structures with locally unfavorable regions, where the missing hetero-atoms may have been involved in stabilizing the protein, and it creates an artifactual energy penalty for the native structure, which short 150 ps dynamics simulations can correct.

Table 3. Pearson Product-Moment Correlation Coefficient between Cα Rmsd and MM-PBSA

	Cα rmsd bin		
Protein	0–2.5	2.5–5.0	>5.0
1ctf	0.62	0.38	0.36
1r69	0.63	0.45	0.37
1sn3	0.66	0.38	0.23
2cro	0.72	0.37	0.41
3icb	0.48	0.55	0.38
4pti	0.84	0.38	0.32
4rxn	0.54	0.47	0.34
Average	0.64	0.43	0.34

The Cα rmsd bins contain every structure in the decoy set within the specified values. The linear relationship between Cα rmsd and MM-PBSA is strongest in the bin of close structures.

Table 4 numerically summarizes the single-point minimization data of the X-ray structures, before and after molecular dynamics, for MM-PBSA and each of its four components. Nine out of the ten crystal structures had a more favorable free energy after the dynamics simulations, with the 150 ps snapshots having moved 0.97 Å on average from their initial conformation and being 43 kcal/mol on average more favorable. Only the 4pti crystal structure, which was already 100 kcal/mol more favorable than the best decoy and, incidentally, whose crystal structure did not contain any hetero-atoms other than water molecules, did not experience an improvement. These substantial improvements in free energy and the approximately 1.0 Å movement away from the crystal structure results from (1) the absence of hetero-atoms included in the X-ray crystal, (2) differences between our more representative aqueous solution and the crystal surroundings, and perhaps (3) inaccuracies in the force field.

Table 5 shows the same results as Table 4, but for the 5 NMR examples. After 150 ps, the NMR structures,

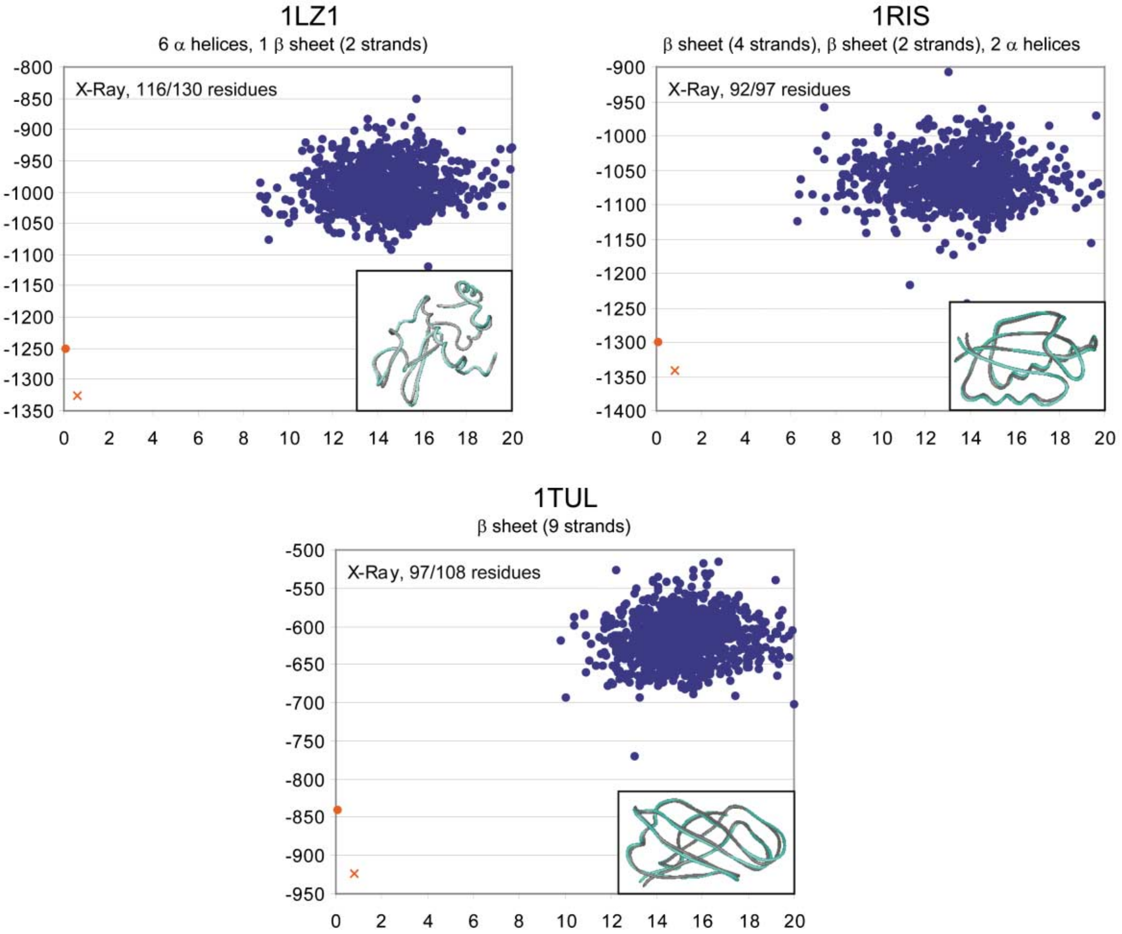

Figure 3. Single-Point MM-PBSA Energy on X-Ray Cases of "Rosetta All Atom Decoy Set"

Single-point minimization MM-PBSA (*y* axes) versus Cα rmsd (*x* axes) on crystal structures and on decoys with no secondary structural restrictions (Rosetta *ab initio* all atom decoy set). Note that these decoys were generated in a structure prediction effort without information of the native structure. Red circles are minimized X-ray crystal structures. Red exes are crystal structures that have been minimized after 150 ps of molecular dynamics in explicit solvent. Figure inlays contain an overlap of the crystal structure before (gray tube) and after (cyan tube) dynamics.

none of which had the most favorable single-point minimization MM-PBSA, moved 60% farther ($<$Cα rmsd$> =$ 1.57 Å), on average, from their starting structures than did the crystal structures. They also experienced a much greater free-energy decrease, 112 kcal/mol on average, with only the 150 ps 1ksr NMR snapshot not showing an improvement over the initial NMR structure and thus not becoming more favorable than its Rosetta decoy set.

The incorrect ranking of the 1ksr conformations stems not from flaws in MM-PBSA but rather from using it to compare single-point calculations on minimized structures. Although it is a rapid and thus desirable calculation, there are at least three reasons why this single-point minimization MM-PBSA method cannot be expected to succeed in all cases. First, minimizations effectively remove temperature and thereby alter the balance between enthalpy and entropy; they thus change the free-energy surface as well. Second, minimizations have difficulty escaping local minima, which can over-penalize conformations, such as those solved by NMR, that experience locally unfavorable energies. Third, the MM-PBSA values fluctuate considerably, with standard deviations on the order of 20–30 kcal/mol. Therefore, to obtain a more accurate MM-PBSA value, one should generate a statistically sufficient ensemble of molecular-dynamics trajectories and compare the resulting ensemble averages. An ensemble, which samples conformational space at 300 K, does not overly weight enthalpic contributions, can much more readily alleviate locally unfavorable interactions to escape local minima, and provides enough data to generate meaningful ensemble averages that one can compare by using *t* tests to evaluate the significance of differences.

Thus, for 1ksr, as well as for each of the other NMR examples, we generated six ensembles: one from the NMR structure, two from the Rosetta decoys with the lowest rmsd, and three from the Rosetta decoys with the lowest single-point minimization MM-PBSAs. The open circles in Figure 5 show the single-point minimization MM-PBSA of all the decoys and the initial NMR structure, relative to the most favorable conformation, with the red ones being the NMR structure and the dark blue ones being the five decoys selected for molecular dynamics. (Note that the energies are relative in Figure

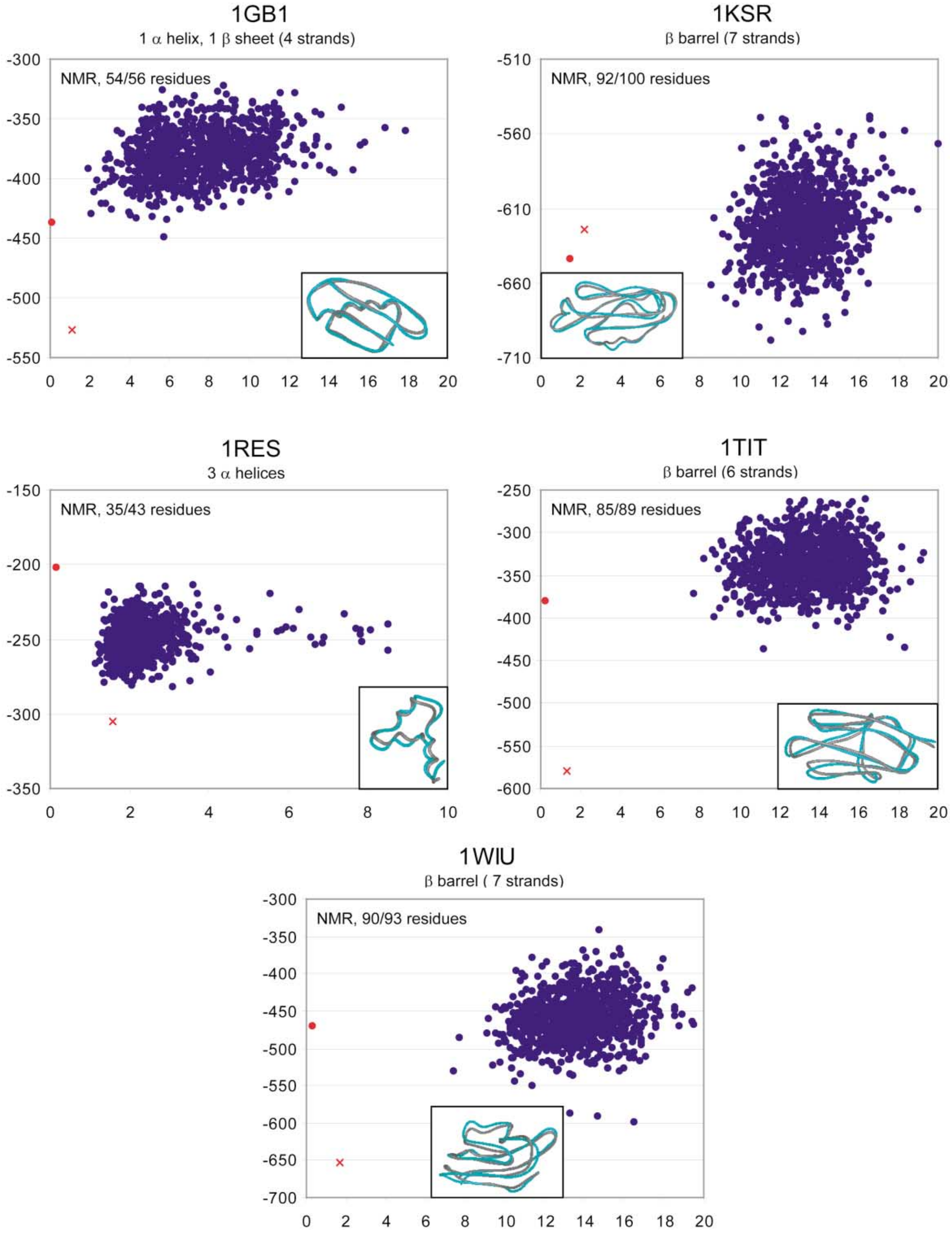

Figure 4. Single-Point MM-PBSA Energy on NMR Cases of "Rosetta All Atom Decoy Set"

Single-point minimization MM-PBSA (*y* axes) versus Cα rmsd (*x* axes) on NMR structures and on decoys with no secondary structural restrictions (Rosetta *ab initio* all atom decoy set). Red exes are NMR structures that have been minimized after 150 ps of molecular dynamics in explicit solvent. Figure inlays contain an overlap of the crystal structure before (gray tube) and after (cyan tube) dynamics.

5 and absolute in Figure 4.) The exes in Figure 5 are the resulting ensemble average MMPBSA values, relative to the best. The arrows map initial snapshot to its corresponding ensemble average. Upon comparing the ensemble averages, we found that the native state then has the most favorable MM-PBSA free energy in every protein, except 1res, where the lowest free energy has only a 1.5 Å Cα rmsd from the NMR-determined structure. It

Table 4. Free-Energy Improvement of X-Ray Structures by Molecular Dynamics

protein	ΔMM-PBSA	Δstrain[1]	ΔVDW	Δsolv_NP[2]	ΔEEL_tot[3]	Cα Rmsd	Resolution
1lz1	−73.4	−0.4	−12.1	1.1	−62.0	0.59	1.5
1ris	−40.3	6.8	−34.3	0.8	−44.6	0.79	2.0
1tul	−83.0	−10.4	−36.5	1.0	−37.2	0.84	2.2
1ctf	−25.2	−8.6	−15.2	0.7	−2.0	0.89	1.7
1r69	−9.8	0.4	−3.5	0.4	−7.1	0.93	2.0
1sn3	−50.8	−16.5	−25.8	0.5	−8.9	0.64	1.2
2cro	−43.1	−19.6	−16.9	0.5	−7.1	1.04	2.4
3icb	−79.8	−27.1	−0.9	1.3	−53.2	1.54	2.3
4pti	5.5	37.1	−28.6	0.3	−3.3	0.93	1.5
4rxn	−29.1	−17.2	−6.4	0.2	−5.7	0.93	1.2
Average	−42.9	−5.5	−18.0	0.7	−23.1	0.97	1.8

The differences are between single point calculations on the initial structure as well as on the 150 ps snapshot of the dynamics simulation. They are not as precise as ensemble average calculations, which are not possible because the minimum requirement for a statistically meaningful ensemble average is 15 snapshots over 150 ps.

[1] The internal strain energy results from deviations away from reference values in bond length, angle, and dihedral terms.

[2] The nonpolar solvation energy accounts for the cost of solvating a discharged solute.

[3] The total elctrostatics energy is the sum of intrasolute Coulombic energies and solute-solvent electrostatic energies.

is also particularly noteworthy that this approach shows the native structure to be most stable for 1ksr, whereas MD followed by minimization (red ex in Figure 4) did not lead to the NMR structure being most stable. To be sure, we only performed the MD average structure analysis on six candidates, rather than the 1000 in the entire decoy set, although we picked the ones with the lowest energy and with the lowest rmsd values from the original minimization analysis as our decoys.

Size Dependence of a Free Energy of Unfolding

Because the whole allure of protein structure prediction rests in its potential to determine structures more quickly than experimental methods, an often overlooked requirement is that the predictor have an absolute means of knowing when the native state has been found. A scoring function that has a high correlation between score and native similarity, when applied to a database of structure predictions, can only identify the lowest-scoring conformation, which it predicts to have the most native similarity, but it cannot determine if this best-scoring structure is native or not. Consequently, we investigated the possibility of using an extended state as the reference, rather than the native state, for our ensemble average MM-PBSA free energy, as Chiche et al. [19] did by using the Eisenberg and McLachlan SFE solvation energy [20]. A fully extended state is desirable for modeling the unfolded state because it normalizes all sequences by eliminating long-range interactions. In contrast, any number of more physically realistic unfolded states could be generated for any particular sequence, with widely varying long-range interactions among those conformations, but the unfolded states among a set of proteins would not be normalized because the relative number of tertiary contacts would be nonuniform. We used an all-α-helical structure as the extended reference for technical reasons (see Experimental Procedures), and we also added hydrogen atoms to sulfur atoms of cysteine residues involved in disulfide bonds of the native structure. We find, as shown in Figure 6, that among the 15 proteins studied in this work, a strong correlation exists ($r^2 = 0.86$) between the size of a protein, in terms of the number of residues, and its Δ(MM-PBSA) ensemble average free energy, the difference between its native state and its fully extended state, which is entirely alpha. Because the absolute ensemble average MM-PBSA of a fully extended helical state for any protein can always be simulated, this correlation implies that one can come up with an expected absolute ensemble average MM-PBSA value for the na-

Table 5. Free-Energy Improvement of NMR Structures by Molecular Dynamics

Protein	ΔMM-PBSA	Δstrain[1]	ΔVDW	Δsolv_NP[2]	ΔEEL_tot[3]	Cα Rmsd
1gb1	−91.2	−7.9	−38.2	0.4	−45.6	1.12
1ksr	19.7	44.5	106.4	3.5	−134.7	2.18
1res	−103.8	−26.2	−56.5	−0.1	−21.0	1.56
1tit	−200.8	−45.7	−76.8	−0.2	−78.1	1.32
1wiu	−184.7	−69.3	−68.7	2.0	−48.7	1.65
Average	−112.1	−20.9	−26.8	1.1	−65.6	1.57

The differences are between single point calculations on the initial structure as well as on the 150 ps snapshot of the dynamics simulation. They are not as precise as ensemble average calculations, which are not possible because the minimum requirement for a statistically meaningful ensemble average is 15 snapshots over 150 ps.

[1] The internal strain energy results from deviations away from reference values in bond length, angle, and dihedral terms.

[2] The nonpolar solvation energy accounts for the cost of solvating a discharged solute.

[3] The total elctrostatics energy is the sum of intrasolute Coulombic energies and solute-solvent electrostatic energies.

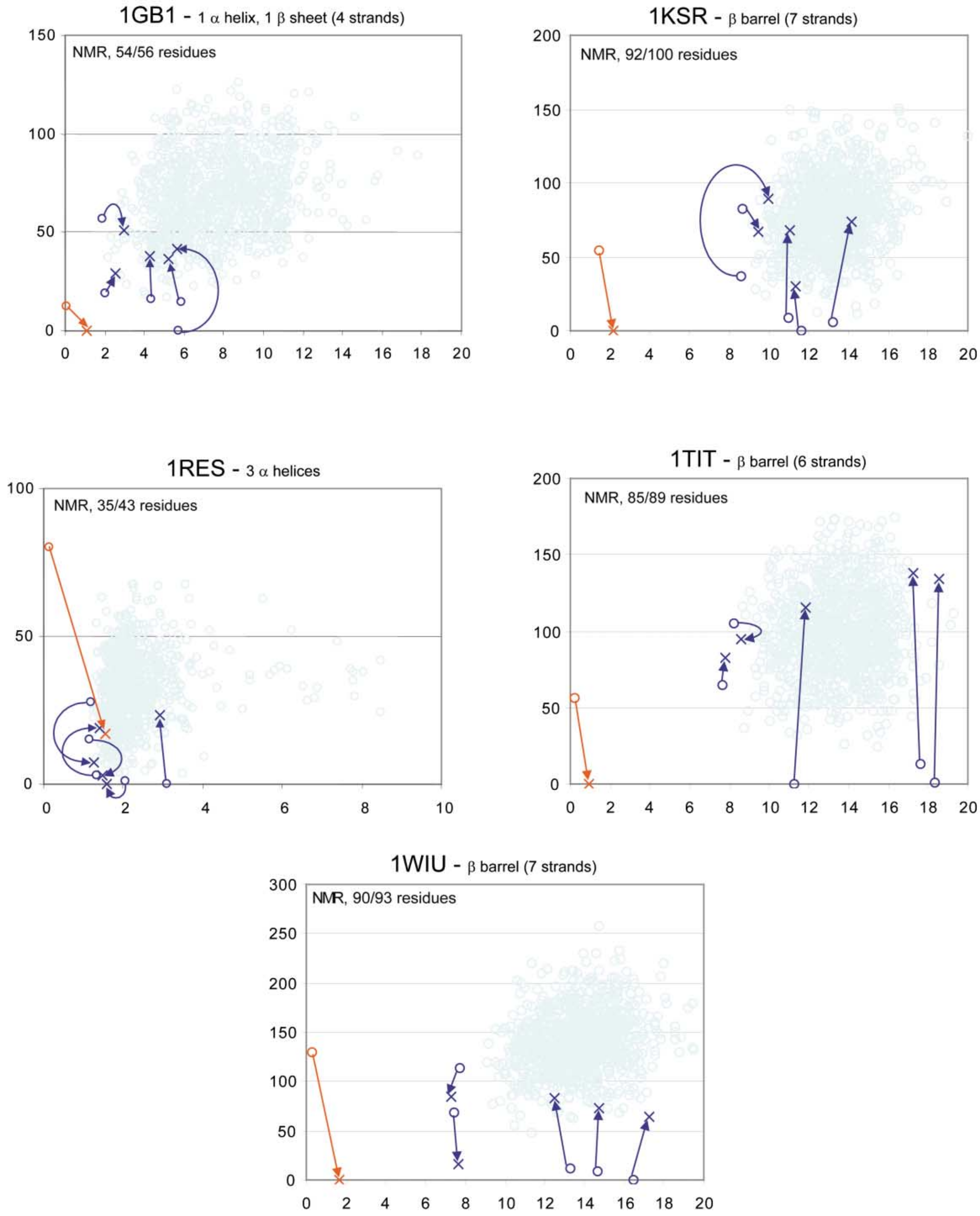

Figure 5. Ensemble Average MM-PBSA energy on NMR Cases of "Rosetta All Atom Decoy Set"

Effect of using ensemble averages on MM-PBSA (*y* axes) versus Cα rmsd (*x* axes). Circles show the same single-point minimization MM-PBSA results as Figure 4, but on a relative scale. Exes are the ensemble averages from 150 ps of molecular dynamics simulation, starting from the conformation represented by the open circle from which the arrow originates.

tive state, based only on the number of residues and thereby provides an absolute check for identifying the native state.

That $\Delta(\text{MM-PBSA})_{\alpha\text{-nat}}$ relates linearly to the size of a protein is not a coincidence and can be simply rationalized. The MM-PBSA free energy does not account for conformational entropy (S_{conf}); it predicts the intrinsic free energy of a particular snapshot without including the effects of other degenerate structures residing at the same energy level, which effectively lowers the relative free energy of this ensemble of near-degenerate structures by increasing S_{conf}. In other words, $\Delta G_u = \Delta(\text{MM-PBSA})_{u\text{-nat}} - T{\cdot}S_{conf,u}$. If we use the expression for Boltzmann's law, $S_{conf,i} = R{\cdot}\ln(\Omega_i)$, where Ω_i is the number

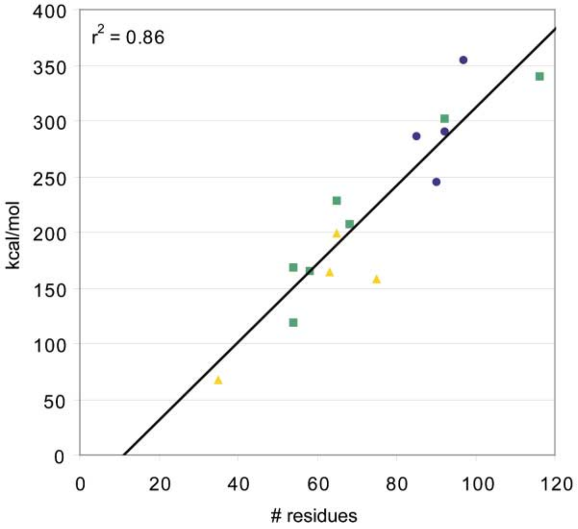

Figure 6. A Free Energy of Unfolding Correlates with Protein Length

Size dependence of Δ(MM-PBSA)$_{\alpha\text{ - nat}}$. Data on the 15 proteins in this work, four alpha proteins (yellow triangles), four beta proteins (blue circles) and seven mixed proteins (green squares), shows a strong relationship between the number of residues and a free energy of unfolding. The *x* intercept is 10 residues, suggesting that the most favorable conformation for peptides of this size may be all α-helical.

of degenerate structures at a given energy level *i*, and assume for the unfolded state that $\Omega_i = y^n$, with *y* representing the average number of conformations per residue and *n* the number of residues, $S_{\text{conf,u}}$ can be assumed to be directly proportional to *n*, $S_{\text{conf,u}} = n \cdot R \cdot \ln(y)$. In view of the empirical fact that ΔG_u remains relatively insensitive to protein size, Δ(MM-PBSA)$_{\text{u-nat}}$ should be roughly equal to $T \cdot S_{\text{conf,u}}$ and thus also proportional to the number of residues. If one finally assumes that the MM-PBSA of the α helical state is representative of the MM-PBSA of other individual members of the unfolded state, Δ(MM-PBSA)$_{\alpha\text{-nat}} \approx \Delta$(MM-PBSA)$_{\text{u-nat}}$.

A final interesting observation from Figure 6 is that the regression line has an *x* intercept of 10 residues. This suggests that peptides of 10 amino acids or fewer prefer the α-helical conformation over any other. For peptides so small, hydrophobic clusters, which are likely critical for compact conformations, would be marginally stable at best. Furthermore, a collapsed structure would probably have less favorable van der Waals interactions than the repeating (*i* to *i* + 4) attractions found in an α helix. Another possibility for interpreting the far-left end of Figure 6 is that the linear relationship adopts a much smaller slope for very small peptides.

Discussion

High-resolution X-ray crystallography structures have an average atomic uncertainty on the order of 0.5 Å. Interestingly, 150 ps snapshots from molecular-dynamics simulations on crystal structures had lower single-point MM-PBSA free energies than the initial crystal structures in nine out of the ten cases, with only 4pti not benefiting energetically from molecular dynamics. While the initial structures were already more favorable than entire ensembles of decoys in all but 3icb, the 150 ps snapshots had better single-point MMPBSA values in all 10 cases. In addition to having more-favorable predicted free energies, 150 ps snapshots moved, on average, less than 1 Å from their initial coordinates; these limited coordinate shifts may have been due to our removal of hetero-atoms, due to adverse effects caused by packing artifacts or other defects in the crystal structure, or due to the intrinsic tendency of proteins to breathe.

The average atomic uncertainty of NMR structures is difficult to quantify. While a popular idea is to evaluate the average deviation from a central average structure on an “NMR ensemble,” this does not account for systematic uncertainties caused by inaccuracies of the energy surfaces being used to refine the structure or for the inability to sample sufficiently during the refinement. Snapshots of 150 ps from simulations on NMR structures showed much greater improvements in free energy and much greater movement, over 1.5 Å, compared to their initial structures, than in the X-ray examples. All five of the NMR models were less favorable than a significant number of decoy structures, and four of the five 150 ps snapshots had a markedly improved free energy, to levels significantly below the best-scoring decoys. However, the more accurate method for evaluating free energies, ensemble average MM-PBSA, favors the native state in all five of the NMR examples. The larger structural shifts and drops in predicted free energies for NMR than for X-ray structures are consistent with the greater uncertainty in NMR structures. Moreover, this work suggests that short explicit-solvent molecular-dynamics simulations can correct, at least in part, for the errors introduced during the standard in vacuo refinement protocol of NMR structure solution.

MM-PBSA provides meaningful, physically based insight into relative free energies of proteins [21, 22], as do a few other energy functions [11–13], but an important finding of this work is that it presents the first look at using this kind of free energy to determine whether a protein structure prediction is of native quality, *sans* the actual experimental structure. We find that a strong correlation exists between the size of a protein and its MM-PBSA free energy of unfolding, from the native state to an all-α-helical state ($r^2 = 0.86$).

Biological Implications

A critical step for making use of the now abundant genomic information is having accurate three-dimensional protein structures, with X-ray crystallography and NMR spectroscopy currently being the two methods that can be used for determining these structures. However, although crystal structures are well known to be more accurate than NMR models, assessing the inaccuracies in the NMR models that are obtained through refinement of NOE constraints has been challenging. The present work suggests that short, room-temperature molecular-dynamics simulations with accurate treatment of solvent effects and long-range electrostatics, which are dramatically more computationally accessible than they were only 5 years ago, are important for escaping locally trapped, energetically unfavorable geometries that are inherent in NMR models.

While protein structure prediction methods are still not at the point where they can be used in place of experimental methods, if they are ever to reach that lofty goal, they must be capable of more than just generating the native structure, for these methods always generate a multitude of models. Structure predictors must also be able to (1) identify which among the scores of generated conformations are most native-like and (2) know if the best structures are actually in the native state or not. Molecular-mechanics free-energy functions such as MM-PBSA that include implicit-solvation free energies and are physically grounded perform better than statistical and empirical functions at ranking structure predictions. We also show in this work that MM-PBSA can be used, together with an α helix-extended state, to accurately predict when a protein conformation is in the native state without any *a priori* native-state information, such as tertiary contacts or secondary structure. This method is based only on the protein length and the difference in free energy between a given conformation and the α-extended conformation.

Experimental Procedures

The AMBER 5 suite of programs [23] was used for all molecular mechanics simulations. The Cornell et al. all-atom force field [24] (parm94) was used for simulations and the parm96 force field [25], which differs only in the ϕ, ψ torsional potentials of the peptide unit, was used in the MM-PBSA free-energy analysis because we have found parm96 to be more robust in protein stability calculations [26].

Minimization

We used a single minimization protocol on all protein conformations: steepest descent for the first 10 cycles, followed by conjugate gradient until the RMS of the Cartesian elements of the potential energy gradient fell below 0.4 kcal/mol·Å. Minimizations were carried out in the gas phase, with a distance-dependent dielectric constant of $4r_{ij}$ and a cutoff for all nonbonded interactions of 25 Å.

Molecular Dynamics

We ran all production phase molecular-dynamics simulations with a 2.0 fs time step under the isothermal-isobaric ensemble (300 K and 1 atm) with explicit solvent; we used the TIP3P model [27] for water, periodic boundary conditions, the particle mesh Ewald (PME) method [28] for electrostatics, a 10 Å cutoff for Lennard-Jones interactions, and SHAKE [29] for restricting motion of all covalent bonds involving hydrogen atoms. We added water molecules around the proteins by using a 10 Å buffer from the edge of the periodic box. The temperature and pressure were maintained by the Berendsen coupling algorithm with τ coupling constants of 1.0. PME grid spacing was approximately 1.0 Å and was interpolated on a cubic B-spline, with the direct sum tolerance set to 10^{-5}. We removed the net center of velocity every 100 ps to correct for the small energy drainage that results from the use of SHAKE, discontinuity in the potential energy near the Lennard Jones cutoff value, and constant pressure conditions.

For equilibration, we solvated the minimized structures, minimized the water molecules alone until the rmsd was <0.1 kcal/mol·Å, and then slowly heated, while allowing the water to move unrestrained for 25 ps (with a 1.0 fs time step) in order to fill any vacuum pockets.

MM-PBSA

Coordinates from a trajectory were saved every 5 ps, and the MM-PBSA calculation was evaluated for each of them. Using an interior dielectric constant of 4, we approximated the MM-PBSA free energy of each snapshot as the sum of two terms, the internal energy of the protein (E_{MM}) and a solvation free energy (ΔG_{solv}). E_{MM} is the sum of an internal strain energy (E_{int}), a van der Waals energy (VDW), and an intrasolute electrostatic energy (EEL). ΔG_{solv} consists of the cost of submerging a discharged solute in solvent (Δsolv_NP) and the subsequent cost of adding the charges back to the solute (Δsolv_eel). Δsolv_NP is approximated as being linearly related to the solvent-accessible surface area (SASA): 5.42*SASA + 920 cal/mol. We adhered to the same Poisson-Boltzmann protocol as first described by Srinivasan et al. [30]; this protocol used DelPhi II [31] and most of its standard default parameters, together with PARSE atomic radii and Cornell et al. charges, to calculate Δsolv_eel. (Note, however, that because we did not factor in salt effects, the Poisson-Botlzmann equation reduces to simply the Poisson equation). The entropy of a given snapshot, which is mostly vibrational, can be calculated with normal mode analysis on a Newton-Raphson minimization. This, however, is the most time-intensive part of the MM-PBSA method on a per-snapshot basis. Given the results in our previous study [21], in which we found this term to be indistinguishable among the native state, the folding intermediate, and the unfolded state of HP-36, we did not perform this calculation in the current study. For a more detailed discussion of the MM-PBSA method, see the review by Kollman et al. [3].

Single-Point Minimization and Ensemble Average Calculations

When comparing the experimental and 150 ps structures with all the structures in the decoy set, we took each individual structure, performed minimization, and evaluated MM-PBSA. We used only a single value for the reported MM-PBSA, which we refer to as single-point minimization values. For the ensemble average values, we took the average of every tenth picosecond over a 150 ps molecular-dynamics simulation because we previously showed that this protocol provides the least expensive, yet statistically sufficient protocol for evaluating an ensemble average MM-PBSA [22].

NMR Structures

When using the term “the NMR structure,” we are referring to model 1 in each of the NMR ensembles. We used this as the representative for simulation purposes because it is more physically realistic than an average structure. The rmsd values, however, are always calculated in reference to the average NMR structure because it is most representative of the various geometries of the ensemble.

Fully Extended Conformations

In order to create a fully extended chain for our reference state, we selected all α-helical conformations because they were computationally efficient and well behaved. The other alternative, an extended β strand, experiences bends in the rod wherever a proline resides. Such bends prevent the extended state from being linearly shaped and lead to water box sizes that are immensely larger than those for the all α-helical conformations. We used flat-well restraints on the backbone ϕ and ψ torsion angles to keep the backbone in a helical conformation, with no energy penalty for $-180° < \phi < -60°$ and $-60° < \psi < -30°$, a parabolic side extending $\pm 20°$ with a 30 kcal/mol·rad^2 force constant, and linear sides, with slopes at the outer edge of the parabolas, extending beyond that.

Z Score

The Z score of a given value among a sample, Z_i, expresses how many standard deviations value *i* is away from the average value of the sample. Negative Z scores mean the value is less than the average. For example, in the four-state reduced decoy set, a Z score of -2.0 for a crystal structure would mean that the crystal structure had an energy that was 2.0 standard deviations lower than the average, which for a perfectly Gaussian distribution would mean that the native was more favorable than 97.5% of all the decoys.

Acknowledgments

We gratefully acknowledge research support from the National Institutes of Health (GM-29072) as well as supercomputer time through the National Science Foundation-supported National Center for Supercomputing Applications as well as the Advanced Biomedical

Computing Center at the National Cancer Institute-Frederick. This article is dedicated in memory of the late Peter Andrew Kollman, who passed away on May 25, 2001, and loved science so much that he did not succumb to cancer until a few days after having helped to complete this work.

Received: May 4, 2001
Revised: August 24, 2001
Accepted: August 31, 2001

References

1. Creighton, T.E. (1992). *Proteins: structures and molecular properties*, Second Edition. (New York: Freeman).
2. Rhodes, G. (1999). *Crystallography Made Crystal Clear*, Second Edition. (San Diego, CA: Academic Press).
3. Kollman, P.A., et al., and Cheatham, T.E. (2000). Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. Acc Chem Res *33*, 889–897.
4. Baker, D. (downloaded in June, 2000). Rosetta All Atom Decoy Set (http://depts.washington.edu/bakerpg/).
5. Simons, K. T., Bonneau, R., Ruczinski, I., and Baker, D. (1999). Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins Suppl. *3*, 171-176.
6. Levitt, M. (1999; downloaded in March, 2001). Decoys 'R' Us (http://dd.stanford.edu/ddownload.cgi?4state_reduced).
7. Park, B., and Levitt, M. (1996). Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. J Mol Biol *258*, 367–392.
8. Anfinsen, C.B. (1973). Principles that govern the folding of protein chains. Science *181*, 223–230.
9. Derman, A.I., and Agard, D.A. (2000). Two energetically disparate folding pathways of alpha-lytic protease share a single transition state. Nat Struct Biol *7*, 394–397.
10. Park, B.H., Huang, E.S., and Levitt, M. (1997). Factors affecting the ability of energy functions to discriminate correct from incorrect folds. J Mol Biol *266*, 831–846.
11. Lazaridis, T., and Karplus, M. (2000). Effective energy functions for protein structure prediction. Curr Opin Struct Biol *10*, 139–145.
12. Dominy, B.N., and Brooks, C.L. III. (2001). Identifying native-like protein structures using physics-based potentials. J Comp Chem, in press.
13. Gatchell, D.W., Dennis, S., and Vajda, S. (2000). Discrimination of near-native protein structures from misfolded models by empirical free energy functions. Proteins *41*, 518–534.
14. Bryant, S.H., and Altschul, S.F. (1995). Statistics of sequence-structure threading. Curr Opin Struct Biol *5*, 236–244.
15. Lazaridis, T., and Karplus, M. (1999). Discrimination of the native from misfolded protein models with an energy function including implicit solvation. J Mol Biol *288*, 477–487.
16. Dill, K.A. (1985). Theory for the folding and stability of globular proteins. Biochemistry *24*, 1501–1509.
17. Bryngelson, J.D., and Wolynes, P.G. (1987). Spin glasses and the statistical mechanics of protein folding. Proc Nat Acad Sci USA *84*, 7524–7528.
18. Dill, K.A., and Chan, H.S. (1997). From Levinthal to pathways to funnels. Nat Struct Biol *4*, 10–19.
19. Chiche, L., Gregoret, L.M., Cohen, F.E., and Kollman, P.A. (1990). Protein model structure evaluation using the solvation free energy of folding. Proc Nat Acad Sci USA *87*, 3240–3243.
20. Eisenberg, D., and McLachlan, A.D. (1986). Solvation energy in protein folding and binding. Nature *319*, 199–203.
21. Lee, M.R., Duan, Y., and Kollman, P.A. (2000). Use of MM-PB/SA in estimating the free energies of proteins: application to native, intermediates, and unfolded villin headpiece. Proteins *39*, 309–316.
22. Lee, M.R., Baker, D., and Kollman, P.A. (2001). 2.1 and 1.8 angstrom average C-alpha rmsd structure predictions on two small proteins, HP-36 and S15. J Am Chem Soc *123*, 1040–1046.
23. Case, D.A., et al., and Kollman, P.A. (1997). AMBER 5.0. San Francisco: University of California-San Francisco.
24. Cornell, W.D., et al., and Kollman, P.A. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc *117*, 5179–5197.
25. Kollman, P., Dixon, R., Cornell, W., Fox, T., Chipot, C., and Pohorille, A. (1997). The development/application of a 'minimalist' organic/biochemical molecular mechanic force field using a combination of *ab initio* calculations and experimental data. In *Computer Simulation of Biomolecular Systems*, Vol. 3, P. Wilkinson, P. Weiner, and W. Van Gunsteren, eds. (Dordrecht, The Netherlands: Kluwer Academic Publishers), pp. 83-96.
26. Lee, M.R., Tsai, J., Baker, D., and Kollman, P.A. (2001). Molecular dynamics in the endgame of protein structure prediction. J Mol Biol, in press.
27. Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., and Klein, M.L. (1983). Comparison of simple potential functions for simulating liquid water. J Chem Phys *79*, 926–935.
28. Darden, T., York, D., and Pedersen, L. (1993). Particle mesh Ewald: an N.log(N) method for Ewald sums in large systems. J Chem Phys *98*, 10089–10092.
29. Ryckaert, J.P., Ciccotti, G., and Berendsen, H.J.C. (1977). Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Comp Phys *23*, 327–341.
30. Srinivasan, J., Cheatham, T.E., Cieplak, P., Kollman, P.A., and Case, D.A. (1998). Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate-DNA helices. J Am Chem Soc *120*, 9401–9409.
31. Sharp, K.A., Nicholls, A., and Sridharan, S. (1998). Delphi II edit. (New York: Columbia University).