doi:10.1006/jmbi.2001.5032 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2001) **313**, 417–430

JMB

Molecular Dynamics in the Endgame of Protein Structure Prediction

Matthew R. Lee[1]*, Jerry Tsai[2], David Baker[2] and Peter A. Kollman[1]

[1]*Department of Pharmaceutical Chemistry, University of California San Francisco San Francisco, CA 94143, USA*

[2]*Department of Biochemistry University of Washington Seattle, WA 98195, USA*

**Corresponding author*

In order adequately to sample conformational space, methods for protein structure prediction make necessary simplifications that also prevent them from being as accurate as desired. Thus, the idea of feeding them, hierarchically, into a more accurate method that samples less effectively was introduced a decade ago but has not met with more than limited success in a few isolated instances. Ideally, the final stages should be able to identify the native state, show a good correlation with native similarity in order to add value to the selection process, and refine the structures even further. In this work, we explore the possibility of using state-of-the-art explicit solvent molecular dynamics and implicit solvent free energy calculations to accomplish all three of those objectives on 12 small, single-domain proteins, four each of alpha, beta and mixed topologies. We find that this approach is very successful in ranking the native and also enhances the structure selection of predictions generated from the Rosetta method.

Keywords: MM-PBSA; proteins; scoring; ranking; refinement

Introduction

Approaches for predicting three-dimensional protein structure based on amino acid sequence, ranging from *ab initio* to comparative modeling, all make considerable approximations in order to contend with the otherwise intractable number of possible conformations. Commonly in *ab initio* methods, a simplified energy potential is used together with a reduced representation of the protein, in which case side-chains are often represented by centroids, hydrogen atoms are usually omitted, and only a few discrete torsional angles are allowed. Comparative modeling methods also rely on many of the same approximations, albeit primarily on the non-homologous regions. These simplifications, while beneficial in that they filter out the majority of unrealistic and improbable structures, limit the degree of accuracy that can be obtained. Even over the homologous regions of a comparative modeling effort, the exact native structure of any sequence inevitably differs from its nearest structural neighbor template, particularly in localized areas that may allow for small global superposition differences, despite large, local deviations that can not be corrected without a more accurate representation of the protein and the energy potential, as well as sufficient sampling.

The solution for overcoming the limiting simplifications is not to remove them from the outset, but rather to add in the detail when necessary, because introducing a higher level of accuracy to the energy potential makes for a more rugged surface that is more difficult to sample, thereby restricting the distance in conformational space that can be sampled on a practical time scale. Thus, the current structure prediction methods must draw the tertiary structure sufficiently close to the correct structure, within a ''radius of convergence'', before all-atom detail with a continuum torsional space may be capable of improving them further.

The first attempts at using all-atom models as the final stage of a hierarchical approach took place a decade ago,[1,2] before control simulations were even capable of maintaining the native state, at a time when the computer power required for even short simulations was very demanding. These investigators applied their methods to the GCN4 leucine zipper, which has a very simple coiled-coil homodimer topology consisting of two 33-mer

Present address: M. R. Lee, Lion Bioscience, Inc., 9880 Campus Point Drive, San Diego, CA 92121, USA.

Abbreviations used: MM-PBSA, molecular mechanics-Poisson-Boltzmann Surface Area; ε_{int}, interior dielectric constant; SASA, solvent-accessible surface area; VDW, van der Waals; eel-tot, total electrostatics.

E-mail address of the corresponding author: matthew.lee@lionbioscience.com

0022-2836/01/020417–14 $35.00/0

monomers, of which all 33 residues were α-helical. In the end, they obtained ~1 Å backbone RMSD structures, but only with the help of α-helical constraints applied to every residue. While having brought forth the enticing idea of hierarchical protein structure prediction, these studies were only successful because they knew the correct structure to begin with and used native constraints to severely reduce the conformational search. Samudrala *et al*.[3] later attempted a hierarchical approach by building all-atom models from a subset of off-lattice predictions on a set of 13 proteins and applying minimization alone, leading to the correct global topology in six of the cases. However, in this study, it was not demonstrated, and is unlikely, that, the final stage of this hierarchical effort added any value to the initial off-lattice models, since minimization affords extremely limited conformational sampling at best. More recently, with advances in simulation methods, most notably being accurate means for treatment of long-range electrostatics[4] that allows for maintenance of native protein structures,[5] our group used an enhanced sampling protocol called ''locally enhanced sampling'',[6] which has been shown to lower energy barriers using a mean-field approach, that drove a 3.7 Å 29-mer protein structure with an incorrectly packed β-sheet to a 2.2 Å conformation with the correct topology.[7] Even more recently, we ran nanosecond time scale, state-of-the-art molecular dynamics simulations, with accurate long-range electrostatics and explicit solvent, on initial structure predictions for the 36-mer HP-36 and the 65-mer S15 alpha proteins, not only improving some of the model predictions to sub-2.0 Å C^{α} RMSD structures, but also demonstrating that the highest resolution models also had the best predicted molecular mechanics-Poisson-Boltzmann Surface Area (MM-PBSA) free energies among a handful of other models with less native similarity.[8]

In the current work, we further explore the promise of using explicit solvent molecular dynamics simulations together with MM-PBSA for the endgame of structure predictions on 12 other small single-domain proteins, four alpha, four beta and four mixed. The three main objectives are: (1) identification the native state; (2) improved filtering over the previous stage by providing better correlation with native similarity; and (3) refinement of the structures.

Results and Discussion

Conformational families

For each of the 12 proteins, 30 Rosetta model predictions were compared: the centers of the five most highly populated clusters, and among the remaining Rosetta predictions, the five with the best C^{α} RMSD predictions, and the 20 with the most favorable Rosetta energy scores. We equilibrated each of these Rosetta models and the experimental structures in a box of TIP3P[9] water with a 10 Å buffer and ran one ns production phase trajectories, for a total of 372 explicit solvent one ns simulations. After having clustered the resulting trajectories, using a 2.5 Å C^{α} RMSD cutoff (see Methods), we observe that the Rosetta model predictions had an average of 1.8 conformational families over the course of the nanosecond simulation; more specifically, the alpha proteins averaged 1.5, the beta proteins 2.4, and the mixed proteins 1.6. In comparison, 11 of the 12 trajectories on experimental structures had only a single conformational family, with the lone exception, 1gab, spending 90% of the time in the initial conformational family that had a slightly more favorable MM-PBSA free energy (Table 1). The ensemble-

Table 1. Native state stability

Protein[a]	Residues	$\langle RMSD \rangle_{init}$[b]	$\langle RMSD \rangle_{2nd}$	$\langle Q \rangle_{init}$[b]	$\langle Q \rangle_{2nd}$
Alpha					
1gab	47	1.7	2.3	85.4	84.8
1utg	62	1.5		91.1	
1uxd	43	1.3		86.2	
1pou	70	2.0		84.4	
Beta					
1sro	66	1.9		76.9	
1qyp	42	2.3		72.3	
1vif	48	0.9		89.8	
2cdx	54	2.6		74.1	
Mixed					
1leb	63	1.5		88.5	
2ptl	60	1.9		79.8	
5icb	72	1.6		87.0	
5znf	25	1.6		86.9	

11 of the 12 native state one ns trajectories did not leave the initial conformational family, most of which had average C^{α} RMSD values under 2.0 A.

1gab spent 90% of the simulation in the initial family.

[a] 1utg, 1vif and 5icb are X-ray crystal structures. The remaining nine are NMR structures.

[b] In the NMR cases, the average NMR structure was used as the reference, as it best represents the whole ensemble.

average values for the MM-PBSA as a function of two native similarity metrics, C^{α} RMSD on the left and percentage of native contacts on the right, are plotted for each conformational family in Figures 1-3.

Native states

The proteins for this study were selected only on the basis of size and topology, without regards to the experimental method used for the structure determination, leading to a total of nine NMR and three X-ray cases. Table 1 shows that, for the most part, control simulations led to very stable native states having average C^{α} RMSDs values under 2.0 Å and on average a percentage of native contacts (Q-values) greater than 80%. Among the exceptions, the NMR model for 1pou seemed to have a $\langle C^{\alpha}$ RMSD$\rangle$ on the high end, although it still had a very good Q-value of 84%. The three beta proteins with NMR structures, 1sro, 1qyp and 2cdx, had RMSD's on the high end as well as Q-values on the low end, when being compared to their respective average NMR structures. For 2cdx, the one with the greatest deviation from the NMR models, the snapshots from the 1 ns trajectory showed an average pair-wise C^{α} RMSD of 1.36 Å from one another with a standard deviation of 0.36 Å, and consisted of a single conformational family. Similarly, for 1sro and 1qyp, the average pair-wise RMSD values were 1.42 and 1.32 Å, with standard deviations of 0.41 and 0.43 Å, respectively, and they too populated single conformational families throughout their simulations. These findings are in agreement with a separate study,[10] in which we suggest that the approximate treatment of solvent used in solving NMR structures causes them to be less reliable than crystal structures.

MM-PBSA parameters

In the MM-PBSA free energy method, there are a few parameters that one cannot derive from ''first principles.'' Perhaps the greatest difficulty lies in deciding what interior dielectric constant (ε_{int}) is most appropriate. On the one hand, because the atomic point charges in our force field have been derived based on high level quantum mechanical charges with a dielectric constant of 1, we may be justified in using $\varepsilon_{int} = 1$. On the other hand, the experimental dielectric constant in proteins is $\sim$4. Thus, the choice of ε_{int} may be system dependent, with larger dielectric constants than 1 likely to be appropriate in some instances.

Another uncertainty is in deciding which parameters to use for describing the backbone torsional potentials. Because the original Cornell *et al.* force field (PARM94),[11] which was parameterized on a set of dipeptides, was shown to slightly favor α-helical conformations on a training set of tetrapeptides,[12] the torsions for phi and psi had been modified in response to high level *ab initio* calculations on the alanine tetrapeptide, which led to a significantly better agreement between molecular mechanical and quantum mechanical relative free energies on the tetrapeptide training set, giving rise to the PARM96[13] force field. However, it is still not clear that one is more generally the better choice for proteins, particularly in the post-processing stage of MM-PBSA calculations.

The non-polar component of the solvation free energy is a third area in which one may explore multiple values or functional forms. In principle, this term should account for all of the non-electrostatic contributions associated with solvating a molecule, primarily including the entropically unfavorable cost of cavity formation and the always attractive dispersion interactions between solute and solvent. Since it has been reasoned that both of the primary factors involved in this term are roughly proportional to the solvent-accessible surface area (SASA), as found in alkanes, MM-PBSA and other methods that calculate the solvation free energy with a continuum solvent[14–17] use a small positive linear γ coefficient to scale this term as a function of SASA, which assumes that the relative weighting of the unfavorable cavitation is stronger than that of the attractive dispersion. An alternative approach, long used by Cramer's and Truhlar's groups, has been to calculate atomic surface tensions that depend on properties such as atom type and nearest-neighbor recognition,[18] which does not always lead the non-polar solvation free energy to be positive. Recently, Pitera & van Gunsteren[19] demonstrated the importance of considering all solute-solvent van der Waals (VDW) interactions including those buried in the protein interior, indicating that solvent excluded volume may be more appropriate than surface area in relating to the favorable aspect of non-electrostatic solvation free energies. While we continue to make the linear approximation, we explore the effect of using different γ coefficients.

Studies applying MM-PBSA to binding free energies[20] and relative free energies of stability on proteins[8,21] have been successful using values between 1 and 4 for ε_{int}, the PARM96 force field, and
a γ coefficient between 5 and 7 cal mol^{-1} Å^{-2} (1 cal = 4.184 J). Figures 1-3 graphically depict the results using our standard values: $\varepsilon_{int} = 4$, PARM96 and $\gamma = 5.42$ cal mol^{-1} Å^{-2} and Tables 2-4 show the effects of changing ε_{int}, the dihedral component of the force field, and γ on the ability to rank the native structure and on the strength of relationship with C^{α} RMSD, which we discuss below.

Native rank

Table 2 shows the native rank of the conformational families containing the equilibrated experimental structures, according to its VDW and total electrostatics (eel_tot) components, and according to MM-PBSA, using various permutations of the three parameters mentioned above.

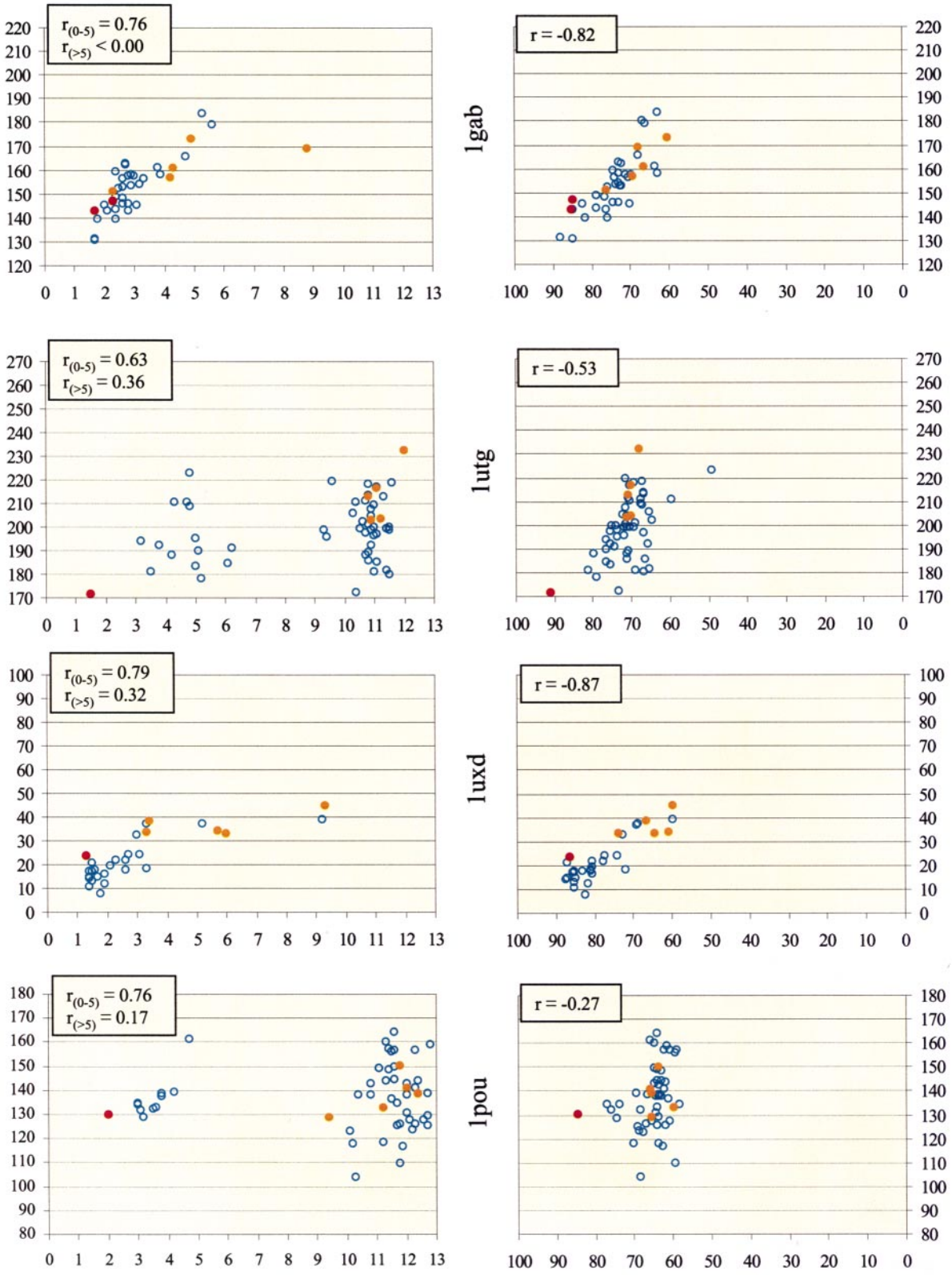

Figure 1. Alpha proteins. For each protein, simulations on 31 initial conformations were carried out: the experimental structure, the top five Rosetta predictions, and 25 other Rosetta predictions. For each conformation, the ensemble-average MM-PBSA was plotted as a function of either C^{α} RMSD (left panels) or Q (right panels), % of native contacts, for every conformational family with a life of more than 100 ps. The conformational families containing the native state are illustrated as closed red circles, those containing the five most favorable Rosetta initial structures as closed orange circles, and those having started from the remaining Rosetta structures as open blue circles. Correlation coefficients were determined using every available data point. On the left panels, it is calculated separately for all points below a 5 Å C^{α} RMSD limit (see the text) and for those above the limit.

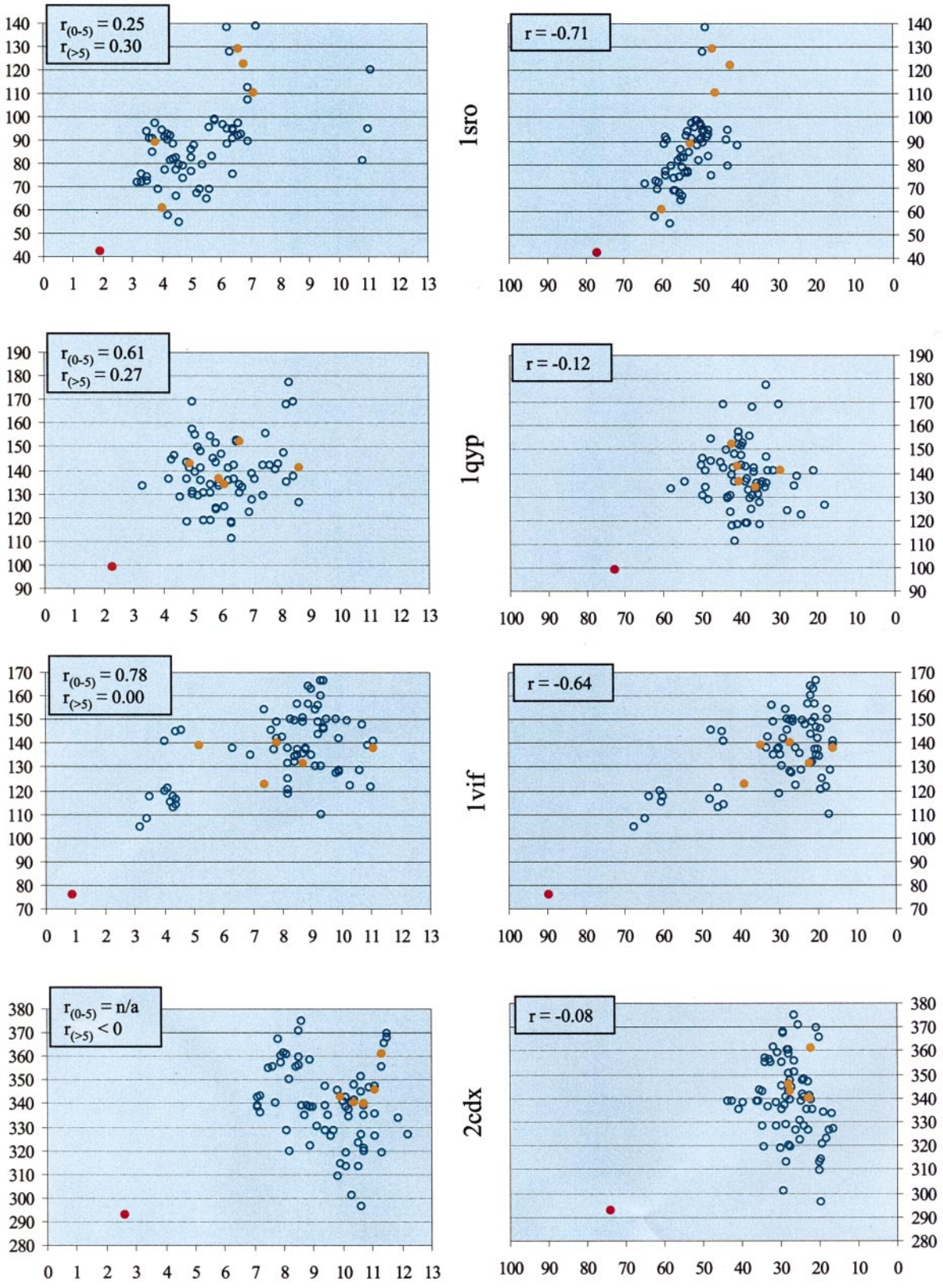

Figure 2. Beta proteins. See the legend of Figure 1.

With the standard set, MM-PBSA predicts the native family as most energetically favorable in eight of the 12 proteins (Table 2). In the alpha proteins 1gab and 1uxd, the average C^{α} RMSD of conformational families lower in free energy is only 2.08 and 1.80 Å, respectively, with the lowest

Table 2. Native rank

Protein	VDW[a]	eel_tot[b]	MM-PBSA				n
			Standard set[c]	$\varepsilon_{int} = 1$[d]	PARM94[e]	$\gamma = 54.2$[f]	
Alpha							
1gab	9	25	7 (2.08 Å)	12 (2.42 Å)	8 (2.14 Å)	10 (2.24 Å)	39
1utg	4	18	1	4 (9.00 Å)	1	2 (10.40 Å)	55
1uxd	2	35	23 (1.80 Å)	27 (1.90 Å)	10 (1.56 Å)	17 (1.70 Å)	36
1pou	18	1	17 (11.01 Å)	1	10 (10.24 Å)	1	53
Beta							
1sro	2	37	1	1	4 (4.27 Å)	1	70
1qyp	1	65	1	1	2 (8.60 Å)	1	71
1vif	1	16	1	1	1	1	73
2cdx	8	53	1	1	32 (10.16 Å)	3 (10.45 Å)	77
Mixed							
1leb	1	48	1	4 (7.07 Å)	3 (7.95 Å)	1	53
2ptl	1	46	5 (3.70 Å)	13 (3.84 Å)	4 (3.67 Å)	1	54
5icb	1	49	1	9 (5.74 Å)	4 (6.33 Å)	1	53
5znf	9	26	1	1	1	1	38
Weighted avg.[a]	4.46	36.23	4.11	5.11	7.20	2.69	

In parenthesis are the average RMSD values of the conformational families having lower energies than the native. With the standard set of parameters, MM-PBSA ranked the experimental structures as best in eight of 12, with two (lgab and 1uxd) of the four false negatives amidst other structures that are arguably part of the native state as well.

[a] Van der Waals energy.

[b] Total electrostatic energy, using $\varepsilon_{int} = 4$: intra-solute Coulombic + $\Delta G_{solv.pol}$.

[c] Standard set is $\varepsilon_{int} = 4$, PARM96, and $\gamma = 5.42$ cal/mol Å^2.

[d] Standard set, except for ε_{int}.

[e] Standard set, except for the force field.

[f] Standard set, except for γ.

[g] Weighted according to n (see Methods).

energy 1gab Rosetta structures having as good a RMSD as the NMR model, and with one having even more native contacts (from the average NMR structure) than the NMR model (Figure 1). In the mixed protein 2ptl, only three of the 54 conformational families had a lower predicted free energy, all three of which were low RMSD structures. Only in the case of the four-helix bundle 1pou do the standard parameters decidedly fail, where nearly half of the 53 conformational families scored better than native. VDW alone performs worse than the standard set MM-PBSA, predicting the native family as most favorable in only five of the same eight that the standard set MM-PBSA did and no others. Interestingly, eel_tot predicts the native as best in only a single instance, 1pou, the protein that the standard set had the most difficulty with, and otherwise ranks very inadequately. Along those lines, using the lower $\varepsilon_{int} = 1$ allows MM-PBSA to correctly rank native in 1pou, while really only worsening three others, the alpha protein 1utg and the two mixed proteins 1leb and 5icb. The PARM94 force field, which has been suggested to unduly favor α-helices,[12] performs similarly to PARM96 on the alpha proteins, but worse on the beta and mixed proteins. Finally, amplifying the γ coefficient, which would more heavily weight the unfavorable cavitation term's dependence on SASA, also allows MM-PBSA to correctly rank 1pou and 2ptl, but slightly upsets the correct ranking of 1utg and 2cdx, with a net effect of ranking a bit better than the standard set.

While others have reportedly demonstrated near 100% success in discriminating native from decoys,[17,22] the decoys have been far less challenging than those investigated in the current work. While the Park & Levitt set[23] spans a wide range of RMSD values, the deviations from native are achieved by relatively minor perturbations of backbone dihedrals from the experimental structure. The other source of decoys in these works, structures threaded onto incorrect sequences, have many unpaired and incorrectly paired tertiary contacts, as well as locally unsatisfied terms, such as helical residues represented as strands. In contrast, the Rosetta decoys for any particular protein consist of widely varying topologies, each containing plausible local structure with optimized tertiary contacts, and are thus considerably more challenging, as corroborated in a study by Gatchell *et al.*[24]

Correlation with native similarity

In order for any energy function to be useful for structure prediction, it must exhibit a good association with native similarity, not just correctly rank the native structure among a set of decoys. Moreover, in a successful hierarchical approach, the final stage must be more effective at correlating with native similarity than the initial structure prediction methods. In this study, we examine the linear correlation coefficient between C^{α} RMSD and the various energies as above, but only for structural families that were less than 5 Å from the

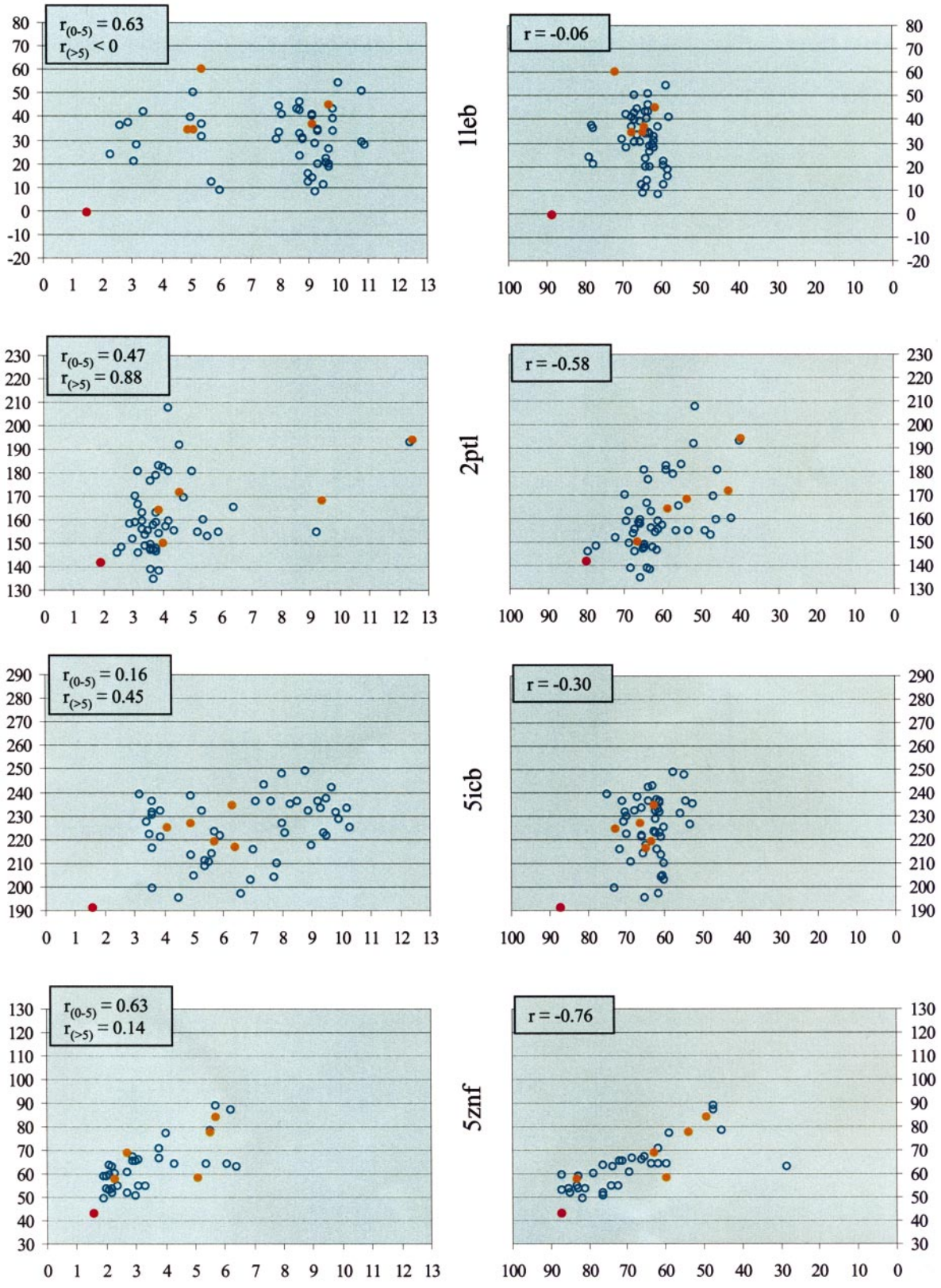

Figure 3. Mixed proteins. See the legend of Figure 1.

experimental structures. We impose this 5 Å limit, because we suggest, in a separate work,[10] based both on the notion of a globally convex free energy landscape and on data of large decoy sets, that the relationship between C^{α} RMSD and an effective free energy such as MM-PBSA is only linear near

the native state, that the relationship weakens dramatically beyond 5 Å C^{α} RMSD.

In Table 3, we summarize our findings, which show that the standard parameter set MM-PBSA correlates with C^{α} RMSD as well as any of the other terms or MM-PBSA parameter permutations. Somewhat surprisingly, VDW does as well as the much more computationally demanding entire effective free energy function itself, even though it did not rank native as well. Eel_tot shows virtually no correlation, which causes the MM-PBSA with $\varepsilon_{int} = 1$ to have a lesser association. The PARM94 force field performs similarly to PARM96 and the higher γ coefficient seem to have no effect on the strength of association between MM-PBSA and C^{α} RMSD.

Part of the reason why the relationship of our free energy with RMSD falls off beyond a certain point stems from RMSD not being the best way of describing native similarity. For instance, a single hinge motion between two domains can lead to very large RMSD values, despite native similarity being otherwise very high. Thus, *Q*-values, the percentage of formed native contacts, provide another way of judging how similar a given conformation is to the reference point, and should not lose their association with a free energy as readily as RMSD. This notion is supported by Figures 1-3, where the relationship of MM-PBSA is roughly more well behaved with respect to *Q* than RMSD in three (1gab, 1utg and 1uxd) of the four alpha proteins, in two (1sro and 1vif) of the four beta proteins, and in all four of the mixed proteins. Furthermore, Table 4 shows that MM-PBSA correlates with the *Q*-value over all structural families to the same extent as it does with RMSD below the 5 Å mark.

Although it is useful to have a scoring function that relates to native similarity, the more relevant issue, in the context of hierarchical protein structure prediction, is whether or not MM-PBSA provides a better filter than Rosetta at selecting the most promising predictions. Because Rosetta does not rely entirely on its energy score in identifying its most favored conformations, it is not appropriate to calculate a correlation coefficient between its score and C^{α} RMSD. Instead, to evaluate whether or not MM-PBSA or VDW is advantageous over Rosetta in scoring its predictions, we compare the $\langle C^{\alpha}$ RMSD$\rangle$ of the best five conformations in Table 5, this being the centers of the five most highly populated clusters generated from Rosetta and the five lowest energy structures according to MM-PBSA or VDW alone. Under each of the three scoring functions, for each protein, we show the average C^{α} RMSD values of the five deemed best, from both their initial and their ensemble-averages. As can be seen from Table 5 and from Figures 1-3, MM-PBSA (using the standard parameter set) improves the structure selection process, as does VDW. Among the alpha proteins, three of the four benefited from MM-PBSA, with the initial structures being on average 2-3 Å better than the five chosen by Rosetta, most notably in the cases of 1gab and 1uxd, where the Rosetta selections were on average ~5 Å, compared to ~2 Å from MM-PBSA and VDW. In addition, for 1gab and 1uxd,

Table 3. Strength of association with C^{α} RMSD

			MM-PBSA				
Protein	VDW[a]	eel_tot[b]	Standard set[c]	$\varepsilon_{int} = 1$[d]	PARM94[e]	$\gamma = 54.2$[f]	*n*
Alpha							
1gab	0.77	0.07	0.77	0.53	0.82	0.81	35
1utg	0.64	−0.31	0.63	0.02	0.49	0.60	11
1uxd	0.78	−0.52	0.74	0.49	0.76	0.77	31
1pou	0.82	−0.40	0.76	0.75	0.67	0.34	11
Beta							
1sro	0.08	0.05	0.25	0.29	−0.04	0.22	35
1qyp	0.57	−0.35	0.61	0.62	0.70	0.65	15
1vif	0.81	0.12	0.78	0.78	0.77	0.80	14
2cdx							0
Mixed							
1leb	0.50	−0.90	0.63	0.47	0.31	0.44	8
2ptl	0.46	−0.05	0.39	0.35	0.41	0.47	46
5icb	0.24	−0.27	0.16	−0.04	0.23	0.20	17
5znf	0.48	−0.08	0.63	0.63	0.47	0.58	29
Weighted avg.[g]	0.53	−0.18	0.55	0.44	0.49	0.54	

These values represent the Pearson product-moment correlation coefficients among families <5 Å from the experimental structure. As seen in the Figures, the linear relationship falls apart for structures that are very dissimilar from the native; the 5 Å mark appears to generally demarcate the limit.

[a] Van der Waals energy.

[b] Total electrostatic energy: intra-solute Coulombic + $\Delta G_{solv.pol}$.

[c] Standard set is $\varepsilon_{int} = 4$, PARM96, and $\gamma = 5.42$ cal mol^{-1} Å^{-2}.

[d] Standard set, except for ε_{int}.

[e] Standard set, except for the force field.

[f] Standard set, except for γ.

[g] Weighted according to *n* (see Methods).

Table 4. Strength of association with Q

Protein	VDW[a]	eel_tot[b]	MM-PBSA				n
			Standard set[c]	$\varepsilon_{int} = 1$[d]	PARM94[e]	$\gamma = 54.2$[f]	
Alpha							
1gab	−0.76	0.43	−0.82	−0.71	−0.83	−0.80	39
1utg	−0.43	−0.06	−0.53	−0.30	−0.50	−0.43	55
1uxd	−0.88	0.00	−0.87	−0-81	−0.85	−0.89	35
1pou	−0.18	−0.18	−0.27	−0.38	−0.28	−0.28	53
Beta							
1sro	−0.65	0.03	−0.71	−0.75	−0.66	−0.72	70
1qyp	−0.58	−0.04	−0.64	−0.64	−0.70	−0.65	71
1vif	−0.58	−0.53	−0.64	−0.64	−0.70	−0.65	73
2cdx	0.09	0.11	−0.08	−0.02	0.30	0.01	77
Mixed							
1leb	−0.20	0.11	−0.06	0.00	−0.32	−0.13	53
2ptl	−0.47	0.20	−0.58	−0.47	−0.42	−0.60	54
5icb	−0.27	0.14	−0.30	−0.17	−0.15	−0.31	53
5znf	−0.82	0.35	−0.76	−0.75	−0.84	−0.78	38
Weighted avg.[g]	−0.44	0.02	−0.50	−0.45	−0.46	−0.49	

These values, as in Table 3, represent the Pearson product-moment correlation coefficients. As Q likely describes native similarity better than RMSD, particularly among the dissimilar structures, no cutoff is used to evaluate a scoring functions' correlation with Q.

[a] Van der Waals energy.

[b] Total electrostatic energy: intra-solute Coulombic + $\Delta G_{solv,pol}$.

[c] Standard set is $\varepsilon_{int} = 4$, PARM96, and $\gamma = 5.42$ cal mol^{-1} Å^{-2}.

[d] Standard set, except for ε_{int}.

[e] Standard set, except for the force field.

[f] Standard set, except for γ.

[g] Weighted according to n (see Methods).

the molecular dynamics simulations improved the RMSD even further by an additional ~0.5 Å. In the beta proteins, the selection process also benefited from the more accurate MM-PBSA in all four proteins, albeit marginally with 2cdx, which was only 0.5 Å better. Finally, among the mixed proteins, the selection process improved substantially in half the cases, with 2ptl and 5znf structures having C^{α} RMSD values that were roughly 3 Å lower, but those from 1leb and 5icb were marginally worse by one and 0.5 Å, respectively. Using the VDW energy alone as the scoring function allowed for the same qualitative areas of improvement in filtering, although the extent of improvement was slightly less. In summary, the average initial C^{α} RMSD among the five chosen by Rosetta was 7.14 Å, that from VDW was 5.90 Å, and that from MM-PBSA was 5.53 Å.

Refinement

The final objective in the endgame of hierarchical protein structure prediction entails improving the native quality of the initial predictions. As we have found when refining two small alpha proteins,[21] there are two aspects of refinement: (1) relaxation to allow for very small domain shifts and correction of locally unfavorable geometries, which have minimal barriers and occur within ~50 ps of molecular dynamics time, due to the more accurate free energy surface in molecular mechanics and (2) transitions over energy barriers into new conformational families that may have more favorable free energies and more native similarity. Thus, to isolate the two potentialities, each trajectory was clustered into conformational families with a 2.5 Å cutoff, as mentioned above.

For looking at possible refinement in the form of relaxation, we examine the initial C^{α} RMSD values in comparison to the average RMSD values of the very first conformational family (Table 6). We further split the data into close, medium and distant bins, 0-2.5 Å from the experimental structure, 2.5-5.0 Å, and >5.0 Å, respectively, because we believe that the closer the structure is to the native to begin with, the greater the likelihood that conformational changes will be favorable. On average, only the relaxation of structures in the close and medium bins of the alpha proteins improved the native similarity, but only slightly.

Not all of the trajectories contained more than a single conformational family, but among those that did, Table 7 compares the ensemble-average C^{α} RMSD values of the initial and second conformational families, again further split into similarity bins. While we did not see any bin in which conformational changes led to more native families, we would only expect this to happen with any regularity in the close similarity bin, where there were very few transitions that are probably not statistically relevant.

Conclusions

While, in principle, progressive improvement in detail should allow for more accurate protein structure prediction, this has not been shown to be the case, except in a few isolated proteins, largely due

Table 5. Ability to filter decoys

Protein	Rosetta[a]		MM-PBSA[b]		VDW[b]	
	Init.[c]	Avg.[d]	Init.[c]	Avg.[d]	Init.[c]	Avg.[d]
Alpha						
1gab	4.56	4.90	2.45	2.08	1.97	1.98
1utg	10.87	11.20	8.36	8.32	9.96	9.80
1uxd	5.31	5.52	1.82	1.52	1.70	1.58
1pou	11.45	11.34	10.88	11.08	10.88	11.06
Beta						
1sro	6.12	5.66	4.34	4.56	4.72	4.48
1qyp	6.51	6.42	5.78	5.86	6.30	6.58
1vif	8.18	8.04	4.76	4.92	6.46	6.46
2cdx	10.03	10.68	9.56	10.26	9.64	10.14
Mixed						
1leb	6.68	6.84	7.44	7.88	8.30	8.54
2pt1	6.43	6.86	3.48	3.46	3.50	3.92
5icb	5.33	5.48	5.96	5.86	5.62	5.44
5znf	4.26	4.26	1.58	2.38	1.78	2.48
Avg.	7.14	7.27	5.53	5.68	5.90	6.04

For each protein, a method's five most favorable are chosen out of the set of 30. MM-PBSA and VDW outperform Rosetta in this task.

[a] The average values of the five most highly populated Rosetta structures.

[b] The average values of the five lowest energy structures (standard parameter set).

[c] Initial C^{α} RMSD values.

[d] Ensemble-average C^{α} RMSD values from first conformational families if more than one. Note that this is an average of an average.

to the very short radius of convergence afforded by the more accurate, detailed methods such as molecular mechanics that include all-atom accuracy along with energy potentials based on first principles, and due to the structure predictions not being of high enough resolution, lying outside of the radius of convergence. Improvements in the initial stages of protein structure prediction by a

Table 6. Relaxation of initial conformations

Protein	0-2.5 Å			2.5-5.0 Å			>5.0 Å		
	Init[a]	$\langle RMSD \rangle_{init}$[b]	n	Init[a]	$\langle RMSD \rangle_{init}$[b]	n	Init[a]	$\langle RMSD \rangle_{init}$[b]	n
Alpha									
1gab	2.26	2.23	9	3.35	3.11	20	8.10	8.80	1
1utg			0	3.88	4.42	5	10.36	10.52	25
1uxd	1.86	1.70	20	3.25	3.38	6	6.73	7.00	3
1pou	2.30	3.00	1	3.68	3.63	4	11.43	11.57	25
Beta									
1sro			0	4.13	4.18	19	7.47	7.20	10
1qyp			0	3.99	4.54	8	6.49	6.57	21
1vif			0	3.62	4.02	5	9.02	8.93	23
2cdx			0			0	9.10	9.61	29
Mixed									
1leb	2.50	2.45	2	3.35	3.58	4	8.12	8.31	25
2pt1	2.46	3.22	7	3.62	3.84	22	9.03	9.27	3
5icb			0	3.81	4.02	4	7.53	7.62	24
5znf	1.50	2.37	21	3.84	4.50	8	6.50	6.40	1
Weighted avgs.[c]									
alpha	1.99	1.90	30	3.45	3.40	35	10.61	10.78	54
beta			0	4.02	4.25	32	8.22	8.36	83
mixed	1.79	2.57	30	3.66	3.97	38	7.87	8.02	53
all	1.89	2.24	60	3.70	3.86	105	8.80	8.95	190

Conformational families for each protein are grouped into three bins, based on their initial C^{α} RMSD, to see if initial extent of native similarity relates to the propensity for relaxation closer to the native state. In the majority of cases, relaxation in the all-atom representation did not improve native similarity.

Values reported in this table are the mean values among all members in the bin.

[a] Initial C^{α} RMSD.

[b] Ensemble-average of the initial conformational family.

[c] Weighted according to n (see Methods).

Table 7. Transitions from initial conformations

Protein	0-2.5 Å			2.5-5.0 Å			>5.0 Å		
	$\langle RMSD \rangle_{init}$[a]	$\langle RMSD \rangle_{2nd}$[b]	n	$\langle RMSD \rangle_{init}$[a]	$\langle RMSD \rangle_{2nd}$[b]	n	$\langle RMSD \rangle_{init}$[a]	$\langle RMSD \rangle_{2nd}$[b]	n
Mixed									
1leb			0	3.20	3.40	1	8.59	8.58	19
2ptl	3.20	3.55	4	3.86	4.00	14	10.95	10.86	3
5icb			0	4.16	4.16	5	7.60	7.75	19
5znf	2.77	3.63	2	5.43	5.88	3			0
Alpha									
1gab	2.65	3.25	2	3.43	3.83	3	8.80	9.00	1
1utg			0	4.42	4.26	5	10.55	10.47	15
1uxd	2.13	2.23	3	3.40	3.30	1	9.30	9.20	1
1pou	3.00	3.10	1	3.67	3.90	3	11.73	11.27	16
Beta									
1sro			0	4.24	4.48	14	6.80	6.78	9
1qyp			0	4.80	4.96	7	6.52	6.78	18
1vif			0	4.01	4.14	5	8.92	8.87	20
2cdx			0			0	9.44	9.67	24
Weighted avg.[c]	2.75	3.15	12	4.17	4.32	61	8.88	8.90	145

Trajectories for each protein that underwent a structural transition are grouped into three bins based on their initial C^{α} RMSD, in order to see what effect transitions had on native similarity. No improvement appears to accompany these structural changes.

Values reported in this table are the mean values among all members in the bin.

[a] Ensemble-average of the initial conformational family.

[b] Ensemble-average of the second conformational family.

[c] Weighted according to n (see Methods).

small handful of methods, most notably the Rosetta method, and development of accurate methods for evaluating the relative free energies of stability, have provided us with the opportunity to demonstrate a successful hierarchical collaboration.

Native rank, filtering and refinement are the three main objectives for the endgame of protein structure prediction. Among the 12 proteins, each with a distinctive topology, the methods presented here handle the first of these goals very well, correctly placing native as first in eight of the examples. In the remaining four, the lower energy structures had average C^{α} RMSD values of only $\sim$2 Å in two of the proteins, which is considered to lie within the narrow range of natural fluctuation around the native state under physiological conditions,[25] and 3.7 in another. In the fourth protein, using either a lower interior dielectric constant of unity or a higher γ coefficient for the non-polar solvation free energy lead to a corrected native rank. The methods presented here also perform adequately as a filtering mechanism in an absolute sense, and substantially better than Rosetta in a relative sense. The third objective, despite our success on the HP-36 villin headpiece and ribosomal S15 protein,[21] is one in which we do not succeed; this does not imply that molecular dynamics made structures worse, only that it did not improve them.

That we were unable to really refine the best structures came as somewhat of a disappointment, but not entirely as a surprise. The nature of refinement found in both S15 and HP-36 included small helical domain shifts into more tightly packed structures. While our energy function is inadequate to improve the structures, we do feel that perhaps one nanosecond, explicit solvent simulations are too short for more systematic refinement of close structures. In order to overcome this limitation, apart from trying to simply run longer simulations, methods that improve the sampling may provide the solution. Locally enhanced sampling was effective on CMTI, which has three disulfide bridges over 29 residues, as mentioned above, but application of this approach on proteins less stable than the disulfide-rich CMTI led to unstable control simulations on the native structure (unpublished results), presumably due to the additional entropy of the method which altered the free energy surface. But because locally enhanced sampling still stands out as a promising method for overcoming large energy barriers, particularly when used locally rather than globally, one might envision application of this mean-field approach directed at those regions with greater known uncertainty in the beginning stages of a hierarchical structure prediction, such as the intervening sequences between predicted secondary structural elements. Alternatively, implicit solvent simulations (R. Luo, L. David & M. K. Gilson, unpublished results)[17,26-28] provide another potential approach for improving sampling, both in terms of the length of simulation that can be accomplished and in terms of the more rapid conformational changes that accompany the absence of solvent viscosity.

Apart from the lack of success in the refinement aspect, the methods presented in this work still performed admirably in ranking the native and selecting better structures than Rosetta. With the automation software used in this work, along with

the increasingly greater computational power that continues to emerge, the methods described for the final stages of structure prediction are much more accessible to the structure prediction community than only a few years ago.

Materials and Methods

The AMBER 5 suite of programs[29] was used for all molecular mechanics simulations. The PARM94 all-atom force field[11] was used for the molecular dynamics simulations and both the PARM94 and PARM96[13] force fields, the latter of which differs only in the ϕ, ψ torsional potentials of the peptide unit, was used in the MM-PBSA free energy analysis.

Rosetta structure prediction

Rosetta builds protein structures from fragments with similar amino acid sequences using a fragment insertion-simulated annealing method for searching conformational space and a simple side-chain centroid based energy/scoring function which favors hydrophobic burial, strand pairing, and other low resolution features of native protein structures.[30]

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 one atmosphere pressure) with explicit solvent, using the TIP3P model[9] for water, periodic boundary conditions, the particle mesh Ewald (PME) method[4] for electrostatics, a 10 Å cutoff for Lennard-Jones interactions, and the use of SHAKE[31] for restricting motion of all covalent bonds involving hydrogen atoms. Water molecules were added around the proteins using a 10 Å buffer from the edge of the periodic box. The temperature and pressure were maintained by the Berendsen coupling algorithm[32] using a ô coupling constants of 1.0. PME grid spacing was ~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.

To cluster the molecular dynamics trajectories, we defined conformational families as being those with C^{α} RMSD values of <2.5 Å from the first structure in the family, with the first snapshot lying >2.5 Å from the first member of the initial family deemed as the first structure of the second conformational family. On those families that were not populated for $\geqslant$100 ps, we did not calculate ensemble-averages and did not consider them in any of the results we report in this study.

MM-PBSA

Coordinates from a trajectory were saved every 5 ps, and the MM-PBSA calculation evaluated on each of them. The MM-PBSA free energy of each snapshot is approximated as the sum of two terms, using an interior dielectric constant of 4: 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 VDW energy, and an intra-solute electrostatic energy (EEL). ΔG_{solv} consists of the cost of submerging a discharged solute in solvent ($\Delta solv$_NP) and the subsequent cost of adding the charges back to the solute ($\Delta solv$_eel). $\Delta solv$_NP is approximated as being linearly related to the SASA: γ*SASA + 920 cal/mol. We adhered to the same Poisson-Boltzmann protocol as described,[14] which used DelPhi II[33] and most of its standard default parameters, together with PARSE atomic radii and Cornell *et al.* charges,[11] to calculate $\Delta solv$_eel. 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 bases. Given the results in our previous study,[8] where 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.*[20]

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, as it is more physically realistic than an average structure. The RMSD values, however, are always calculated in reference to the average NMR structure, as it is most representative of the various geometries of the ensemble.

Q-values

A contact is defined as any two residues containing atoms $\leqslant$3.5 Å apart. A contact map is generated for the actual experimental structure of X-ray crystals and for the average NMR structure of NMR ensembles. The *Q*-value represents the percentage of contacts in the native contact map that are also found in the conformation being evaluated, with the exact same topologies being required in both the reference and target configurations.

Weighted averages

In Tables 2, 3, 5 and 6, we report a weighted average according to n, which is calculated as follows:

$$weighted\ avg. = \frac{\sum_{i=1}^{N} in_i}{N}$$

n_i is the number of samples in ensemble i, and N is the total number of samples among all ensembles in the bin.

Automation

The bottleneck in running molecular dynamics simulations and MM-PBSA calculations on a single protein conformation lies in the computer time. However, when dealing with larger numbers, human intervention and data analysis assume that role. For this work, in which we simulated nanosecond length simulations, analyzed, and post-processed the MM-PBSA on 372 different struc-

tures, a set of programs with the Perl scripting language was written to automate the process of not only handling large numbers of protein structures, but also of routinely converting from standard pdb to AMBER file format, equilibrating in solvent, running production phase simulations and calculating the MM-PBSA free energies, with a minimal working knowledge of the AMBER suite of programs, thereby making the MM-PBSA calculations for protein stability much more accessible to protein modelers (Appendix A in thesis by M. R. Lee[34]). The simulations scale with complete efficiency up to the number of computer processors available, by running simulations in coarse grain parallel. The majority of simulations in this work were run on six separate four-processor Compaq Alpha ES40 machines, which when combined with the automation software, allowed for 24 independent simulations running simultaneously.

Acknowledgments

This article is dedicated to the late Peter Andrew Kollman, who passed away on May 25, 2001, and whose passion for science, gifted mind, and gracious heart contributed immeasurably to the intellectual development of M.R.L. and to the advancement of our scientific community.

P.A.K. was grateful to the NIH for research support (GM-29072) and M.R.L. thanks the Advanced Biomedical Computing Center of the National Cancer Institute at Frederick for computer time.

References

1. Vieth, M., Kolinski, A., Brooks, C. L., III & Skolnick, J. (1994). Prediction of the folding pathways and structure of the GCN4 leucine zipper. *J. Mol. Biol.* **237**, 361-367.
2. Nilges, M. & Brunger, A. T. (1991). Automated modeling of coiled coils - applicatino to the GCN4 dimerization region. *Protein Eng.* **4**, 649-659.
3. Samudrala, R., Xia, Y., Huang, E. & Levitt, M. (1999). Ab initio protein structure prediction using a combined hierarchical approach. *Proteins: Struct. Funct. Genet.* **Suppl. 3**, 194-198.
4. Darden, T., York, D. & Pedersen, L. (1993). Particle mesh Ewald: an N.log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089-10092.
5. Fox, T. & Kollman, P. A. (1996). The application of different solvation and electrostatic models in molecular dynamics simulations of ubiquitin: how well is the X-ray structure ''maintained''?. *Proteins: Struct. Funct. Genet.* **25**, 315-334.
6. Roitberg, A. & Elber, R. (1991). Modeling side-chains in peptides and proteins - application of the locally enhanced sampling and the simulated annealing methods to find minimum energy conformations. *J. Chem. Phys.* **95**, 9277-9287.
7. Simmerling, C., Lee, M. R., Ortiz, A. R., Kolinski, A., Skolnick, J. & Kollman, P. A. (2000). Combining MONSSTER and LES/PME to predict protein structure from amino acid sequence: application to the small protein CMTI-1. *J. Am. Chem. Soc.* **122**, 8392-8402.
8. Lee, M. R., Duan, Y. & 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: Struct. Funct. Genet.* **39**, 309-316.
9. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926-935.
10. Lee, M. R. & Kollman, P. A. (submitted). Free energy calculations highlight differences in accuracy between X-ray and NMR structures and add value to protein structure prediction. *Structure* **00**, 000-000.
11. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M. *et al*. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117**, 5179-5197.
12. Beachy, M. D., Chasman, D., Murphy, R. B., Halgren, T. A. & Friesner, R. A. (1997). Accurate ab initio quantum chemical determination of the relative energetics of peptide conformations and assessment of empirical force fields. *J. Am. Chem. Soc.* **119**, 5908-5920.
13. Kollman, P., Dixon, R., Cornell, W., Fox, T., Chipot, C. & 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* (Wilkinson, P., Weiner, P. & Van Gunsteren, W., eds), vol. 3, pp. 83-96, Elsevier, Amsterdam.
14. Srinivasan, J., Cheatham, T. E., Cieplak, P., Kollman, P. A. & Case, D. A. (1998). Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate - DNA helices. *J. Am. Chem. Soc.* **120**, 9401-9409.
15. Still, W. C., Tempcyzk, A., Hawley, R. C. & Hendrickson, T. (1990). Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **112**, 6127-6129.
16. Dominy, B. N. & Brooks, C. L., III (2001). Identifying native-like protein structures using physics-based potentials. *J. Comp. Chem.* In the press.
17. Lazaridis, T. & Karplus, M. (1999). Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J. Mol. Biol.* **288**, 477-487.
18. Giesen, D. J., Gu, M. Z., Cramer, C. J. & Truhlar, D. G. (1996). A univeral organic solvation model. *J. Org. Chem.* **61**, 8720-8721.
19. Pitera, J. W. & Van Gunsteren, W. (2001). The importance of solute-solvent van der Waals interactions with interior atoms of biopolymers. *J. Am. Chem. Soc.* **123**, 3163-3164.
20. Kollman, P. A., Massova, I., Reyes, C., Kuhn, B., Huo, S. H., Chong, L. *et al*. (2000). Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.* **33**, 889-897.
21. Lee, M. R., Baker, D. & 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.
22. Vorobjev, Y. N., Almagro, J. C. & Hermans, J. (1998). 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: Struct. Funct. Genet.* **32**, 399-413.

23. Park, B. & Levitt, M. (1996). Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. *J. Mol. Biol.* **258**, 367-392.
24. Gatchell, D. W., Dennis, S. & Vajda, S. (2000). Discrimination of near-native protein structures from misfolded models by empirical free energy functions. *Proteins: Struct. Funct. Genet.* **41**, 518-534.
25. Brooks, C. L., III, Karplus, M. & Pettitt, B. M. (1988). Proteins: a theoretical perspective of dynamics, structure and thermodynamics. In. *Advances in Chemical Physics*, vol. 71, pp. 1-259, J. Wiley, New York.
26. Dominy, B. N. & Brooks, C. L., III. (1999). Development of a generalized Born model parameterization for proteins and nucleic acids. *J. Comp. Chem.* **103**, 3765-3773.
27. Tsui, V. & Case, D. A. (2000). Molecular dynamics simulations of nucleic acids with a generalized Born solvation model. *J. Am. Chem. Soc.* **122**, 2489-2498.
28. Liu, Y. & Beveridge, D. L. (2001). Exploratory studies of *ab initio* protein structure prediction: multiple copy simulated annealing, AMBER energy functions and a generalized Born/Solven accessibility solvation model. *Proteins: Struct. Funct. Genet.* In the press.
29. Case, D. A., Pearlman, D. A., Caldwell, J. A., Cheatham, T. E., Ross, W. S. *et al*. (1997). *AMBER 5.0*, University of California, San Francisco, San Francisco.
30. Simons, K. T., Bonneau, R., Ruczinski, I. & Baker, D. (1999). Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins: Struct. Funct. Genet.* **Suppl. 3**, 171-176.
31. Ryckaert, J. P., Ciccotti, G. & 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.
32. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684-3690.
33. Sharp, K. A., Nicholls, A. & Sridharan, S. (1998). Delphi II edit, Columbia University, NY.
34. Lee, M. R. (2001). Using molecular dynamics for high resolution protein structure prediction. PhD thesis in Pharmaceutical Chemistry, University of California, San Francisco.

Edited by B. Honig

(Received 9 May 2001; received in revised form 17 August 2001; accepted 17 August 2001)