Application of MM/PBSA colony free energy to loop decoy discrimination: Toward correlation between energy and root mean square deviation

FEDERICO FOGOLARI¹ and SILVIO C.E. TOSATTO²

¹Dipartimento di Scienze e Tecnologie Biomediche, Università di Udine, 33100 Udine, Italy

²Dipartimento di Biologia and CRIBI Biotech Centre, Università di Padova, 35131 Padova, Italy

(RECEIVED July 23, 2004; FINAL REVISION December 20, 2004; ACCEPTED December 20, 2004)

Abstract

Accurate free energy estimation is needed in many predictive tasks. The molecular mechanics/Poisson-Boltzmann solvent accessible surface area (MM/PBSA) approach has proven to be accurate. However, the correlation between the estimated free energy and the distance (e.g., root mean square deviation [RMSD]) from the most stable conformation is hindered by the strong free energy dependence on minor conformational variations. In this paper, a protocol for MM/PBSA free energy estimation is designed and tested on several loop decoy sets. We show that further integration of MM/PBSA free energy estimator with the colony energy approach makes the correlation between the free energy and RMSD from the native structure apparent, for the test sets on which it could be applied. Our results suggest that (1) the MM/PBSA free energy estimator is able to detect native-like structures for most decoy sets, and (2) application of the colony energy approach greatly hampers the MM/energy strong dependence on minor conformational changes.

Keywords: loop decays; MM/PBSA; implicit solvent; Poisson-Boltzmann; colony energy

One common problem in fields like structural biology, as well as in computational and pharmaceutical chemistry, is the prediction of molecular conformations or complexes which have low free energy, i.e., that are thermodynamically stable (Ajay and Murcko 1995; Honig and Yang 1995; Vajda et al. 1997; Schapira et al. 1999; Kollman et al. 2000; Lazaridis and Karplus 2000).

Basically predicting a stable molecular conformation may be viewed as a two-step procedure where first a large number of plausible molecular models are generated, possibly involving some rough free energy estimation, and, second, their free energy is accurately evaluated (see, e.g., Bonneau and Baker 2001).

Accurate free energy estimation is difficult because the environment (usually aqueous) must be taken into account and because it is difficult to estimate entropic effects, which depend on the energy landscape.

The progress made in recent years in protein structure prediction based on free energy estimation (see, e.g., Lee et al. 2001) seems of wider impact than just that related to the protein predictors' community.

Designing good (relative) free energy estimators for protein conformation is particularly challenging because only the most stable structure is available from experiments. Higher free energy structures (often termed decoys) can be generated by computational methods, but their relative free energy is not available. Thus, predicted free energies cannot be compared with experimental values and the only possible test is to see whether the relative free energy of the decoys is indeed higher than that of the native structure.

The quality of the decoys is not known in advance but may be judged by how many free energy estimators fail to recognize the native structure. Decoy sets conversely may be used to test which free energy estimators perform consistently better than others (Park and Levitt 1996; Samudrala and Levitt 2000). A trade-off between performance and computational time is, however, expected.

Since the native structure is assumed to be at a thermodynamic minimum (Anfinsen 1973), the recognition of a native structure among well-constructed decoys is a prereq-

Reprint requests to: Federico Fogolari, Dipartimento di Scienze e Tecnologie Biomediche, Università di Udine, P.le Kolbe 4, 33100 Udine, Italy; e-mail: ffogolari@mail.dstb.uniud.it; fax: ++39-0432-494301.

Article and publication are at <http://www.proteinscience.org/cgi/doi/10.1110/ps.041004105>.

uisite of any (free) energy function that aims at reproducing protein folding thermodynamics. It should be clear that this task is, however, only loosely related to the problem of protein structure prediction where the native structure is not available and, at best, native-like structures are available in the pool of predicted models. For this reason the necessary stronger requirement for a good free energy estimator is that structures assigned the lowest free energy, in a decoy set containing at least one native-like structure, must be native-like.

The relationship between free energy and similarity to the native structure (measured for instance by backbone coordinate root mean square deviation [RMSD]) is complicated by the fact that small changes in atomic coordinates (e.g., of sidechain atoms) may imply rather large changes in various energy terms, e.g., covalent, van der Waals, and electrostatic terms. This fact led some groups (see, e.g., Dominy and Brooks 2002) to neglect some energy terms, more sensitive to poor local geometry.

Lazaridis and Karplus (2000) classified the approaches to free energy estimation in two broad classes, those based on statistical effective energy functions (SEEF) and those based on physical effective energy functions (PEEF). Evidence has been presented in the literature in favor of one or the other class of approaches whose effectiveness depends on the functions, protocols, and test sets employed. Recent results show that physical effective energy functions based on molecular mechanics force fields and accurate estimation of solvation effects are well suited for estimating the relative free energy of atomic detail molecular models (see, e.g., Gatchell et al. 2000; Lee and Kollman 2001; Feig and Brooks 2002; de Bakker et al. 2003; Hsieh and Luo 2004).

SEEFs employing more or less detailed representations of protein chains have also been developed and used successfully on many decoy sets (see, e.g., Zhou and Zhou 2002; Berrera et al. 2003 and references cited therein; Zhang et al. 2004a,b).

The advantage of PEEFs over SEEFs is that they should be transferable to any system (if the methods they employ are themselves transferable, as they usually are). If methods are accurate they should be able to discard energetically disfavored conformations that are not recognized by many successful SEEFs.

One disadvantage of PEEFs based on molecular mechanics forcefields is that they are extremely sensitive to model-building errors that lead to bad atomic contacts. If not removed by minimization steps or partial relaxation, these lead to higher conformational energy.

The large number of successful molecular dynamics simulations in explicit solvent demonstrate that accurate molecular mechanics forcefields are available and the intramolecular energy of a protein conformation may be estimated (at least in a relative sense) accurately (see, e.g., MacKerell et al. 1998). Although molecular mechanics en-

ergies have been used to discriminate plausible versus non-plausible protein conformations (for an early reference, see Novotny et al. 1984), they perform poorly (Wang et al. 1995) because of the following difficulties:

1. Solvent effects are of utmost importance and must be accurately estimated.
2. Entropic effects are not immediately computed from a single conformation.
3. Lennard-Jones and Coulomb energy terms dependence on interatomic distance is often an order of magnitude larger than the whole protein-folding energy. Therefore, minimization with approximate treatment of solvation effects may lead to inaccuracies.

Many models with varying degrees of accuracy have been proposed for estimation of solvent effects. A minimal, but efficient, model prescribes that solvation free energy is proportional to the solvent-accessible surface area (SASA) distinguishing polar and nonpolar surfaces (Eisenberg and MacLachlan 1986; Fraternali and van Gunsteren 1996). In general, it is widely accepted that solvation effects may be decomposed in a "hydrophobic" component and a "polar" or electrostatic component, and that the two components may be treated, at least approximately, independently (Honig et al. 1993; Roux and Simonson 1999). A large amount of experimental data on hydrocarbon solubility supports the notion that the free energy of solvation (transfer from vacuum or from a nonpolar medium to water) for nonpolar molecules is proportional to SASA, through a surface tension coefficient γ (Honig et al. 1993; Roux and Simonson 1999; but see also Blokzijl and Engberts 1993; Levy et al. 2003).

The polar part of the solvation free energy may be estimated in different analytical ways that approximate the computationally expensive solution of the Poisson-Boltzmann equation, which is the underlying model (Davis and McCammon 1991a; Honig and Nicholls 1995; Bashford and Case 2000; Fogolari et al. 2002).

The model based on molecular mechanics forcefields, Poisson-Boltzmann (PB) equation, and SASA (MM/PBSA) is thoroughly described in the Materials and Methods section.

Estimation of entropic effects requires that the ensemble of conformations in the proximity of the conformation under scrutiny is evaluated. Some approaches approximate the energy function close to the minimum with a harmonic oscillator for which entropy can be analytically computed (Srinivasan et al. 1998). Others assign conformational restrictions to torsional degrees of freedom (Doig and Sternberg 1995). A recently proposed approach uses the same ensemble of generated conformations (and their effective energy) to assess the number of conformations belonging to

the “colony” of which each conformation is a sample (Xiang et al. 2002). Although this approach has several advantages that are described later, it has not been widely applied.

It is likely that procedures that build a large number of structures do not properly minimize the energy of predictive models and therefore the effective energy is dominated by poor geometry. An energy minimization protocol and/or a short-time molecular dynamics simulation (Vorobjev et al. 1998; Vorobjev and Hermans 2001) should therefore be applied to any predictive model before subjecting the same to more accurate effective energy estimation (see, e.g., Feig and Brooks 2002). Since the energy minimization will be performed without explicit solvent molecules, some care must be taken to include solvent effects, at least approximately. Due to the approximated model that is necessarily adopted, the minimization leads to a minimum that may be quite different from both the starting model and the effective energy minimum. For these reasons, extensive minimization, depending obviously on the quality of starting models, should be avoided.

In this paper, we design a protocol for the estimation of the free energy of molecular conformations by combining the MM/PBSA and the colony energy (Xiang et al. 2002) approaches. For the test sets studied, we find a good correlation between the estimated free energy and the similarity to the native structure, even with a rather conservative choice of parameters. These results suggest that MM/PBSA free energy estimation (and perhaps other MM/implicit solvent methods) may be made more robust, i.e., less sensitive to minor conformational changes, by supplementing it with the colony free energy approach.

Results

To test the MM/PBSA energy estimation protocol, it has been first applied on

1. The Misfold dataset of Holm and Sander (1992), which includes 26 decoys for 23 protein structures. This served for gross testing of the protocol.
2. The best eight-residue loop decoy dataset of de Bakker et al. (2003), which includes 50 decoys for each of 32 eight-residue loops, from different proteins, optimized by minimizing the MM/GBSA energy. This dataset constitutes a rather difficult test set because decoys have been already optimized with an accurate methodology.

Then, the MM/PBSA colony free energy estimation protocol has been applied on two sets containing a large number of decoys, a prerequisite for colony energy approach application:

3. The abm_dataset (available at <http://dd.stanford.edu/>) (Samudrala and Levitt 2000), which includes four eight- and nine-residue decoy sets from an antibody. This test set allowed us to apply a general version of the colony energy approach because the set includes a large number of decoys;
4. A set of almost 1000 decoys generated by a fast divide-and-conquer method (Tosatto et al. 2002), which includes 20 decoys for all eight-residue stretches of a small protein.

Misfold dataset

A first test has been performed on the misfold decoy set (Holm and Sander 1992), mainly for comparison with other published reports and to check the protocols employed. Results (reported in Table 1) are consistent with the results of Petrey and Honig (2000) and Dominy and Brooks (2002), although free energy differences here are larger probably due to the different model and protocol used. Structures 1rei.pdb and 2ts1.pdb have been excluded from analysis because they are not monomeric or contain a break in chain connectivity as revealed by the program WHATIF (Vriend 1990).

The protocol employed here also recognizes the native structure of 1fdx.pdb versus the misfolded 1fdxonrxn.pdb, but the free energy difference is the second lowest one, consistent with previously reported results. Free energy is incorrectly estimated because two Fe–S clusters are not included in the calculation, as discussed by Petrey and Honig (2000). Free energy differences are in the range from -61 to -2264 kcal mol⁻¹ showing that the native structure is easily recognized.

It should be noted that for this decoy set molecular mechanics (MM) energy alone is able to discriminate all but one native structure (decoy 1ppton1cbh) and, with the protocol employed here, it performs slightly better than MM/GBSA energy. For almost all decoys, Coulombic energy is more favorable in the native structure than in the decoy, while PB solvation energy favors the decoy versus the native structure. Overall, the sum of the two contributions, i.e., of direct and solvation electrostatic interactions, is able to recognize the native structure in all but one case (1cbhon1ppt).

This quick test confirmed that the protocol designed to remove possible local poor geometry and at the same time take into account solvation effects is effective.

de Bakker et al. best eight-residue loop decoys

The decoy set of de Bakker et al. (2003) has been extensively studied by the same authors with excellent results, using MM/GBSA, as implemented in AMBER. Before dis-

Table 1. MM/PBSA energy and energy components for the Misfold data set

Decoy	MM/PBSA	Coul + solv	Coul	Solv	SASA	MM/GBSA	MM
1bp2on2paz	-822.8	-734.1	-926.8	192.7	1.7	-262.1	-1017.2
1cbhon1ppt	-61.5	8.0	-27.2	35.2	-9.2	-22.0	-87.5
1fdxon5rxn	-64.5	-38.3	3.9	-42.2	-9.0	-57.6	-13.3
1hipon2b5c	-494.8	-352.3	-665.6	313.3	-13.0	-172.5	-795.2
11hlon2ilb	-693.5	-567.3	-575.9	8.6	-14.5	-268.5	-687.7
1p2pon1rn3	-274.0	-288.7	-514.3	225.6	-10.9	-56.6	-488.6
1ppton1cbh	-150.6	-95.7	115.6	-211.3	5.6	-131.3	55.2
1rhdon2cyp	-1304.1	-984.5	-1036.5	52.0	4.5	-401.5	-1360.6
1rn3on1p2p	-459.2	-285.1	-481.7	196.5	-1.4	-141.8	-654.3
1sn3on2ci2	-381.0	-293.8	-630.8	337.0	-6.1	-132.0	-711.9
1sn3on2cro	-301.2	-215.5	-345.3	129.8	24.4	-120.8	-455.4
2b5con1hip	-551.0	-466.0	-623.5	157.5	18.1	-202.2	-726.6
2cdvon2ssi	-642.8	-575.2	-1044.1	468.9	-40.4	-222.6	-1071.2
2ci2on1sn3	-468.8	-367.7	-662.2	294.5	2.6	-199.5	-765.9
2ci2on2cro	-473.0	-400.6	-808.9	408.3	5.9	-112.7	-887.2
2croon1sn3	-301.0	-377.5	-737.9	360.4	158.6	7.5	-820.0
2croon2ci2	-405.8	-267.5	-247.2	-20.3	-67.8	-196.8	-317.6
2cypon1rhd	-2264.0	-1657.8	-1924.8	267.0	-21.1	-861.7	-2509.9
2i1bon1lh1	-486.8	-292.6	-549.6	257.0	20.2	-111.7	-764.0
2pazon1bp2	-481.4	-464.1	-512.9	48.9	153.2	14.6	-683.5
2ssion2cdv	-169.7	-108.1	-138.6	30.5	26.1	39.9	-226.4
2tmnon2ts1	-1853.9	-1365.7	-2459.7	1094.0	-68.3	-211.6	-2879.6
5padon1rei	-1621.8	-846.9	-1361.9	515.0	-65.2	-796.7	-2071.5
5rxnon1fdx	-249.2	-91.5	-235.0	143.5	0.0	-193.0	-392.7

The difference in energies (in kcal/mol) between the decoy and the native structures is reported. The decoy identity is reported in the first column followed by the differences in MM/PBSA energy (MM/PBSA) in the sum of Coulombic energy and PB electrostatic solvation energy (Coul + solv), Coulombic energy (Coul), in PB electrostatic solvation energy (solv), in solvent accessible surface area energy (SASA), MM/GBSA energy: (MM/GBSA), and in molecular mechanics energy (MM).

cussing results on this set, a few observations must be made. Models might be incomplete due to missing groups, similar to the structure of ferredoxin in the misfold set. Here, 25 of the 32 structures have been crystallized with heterogroups (different from water) present in the PDB file. These heterogroups are not represented in the analyzed models. Seventeen out of 32 structures have one or more *cis*-peptide bonds and 14 structures have one or more residues falling in unfavorable regions of the Ramachandran plot. The energetics of these structural features is not guaranteed to be accurately represented by forcefields that have been designed to reproduce small molecule crystal structures.

Related to this observation, a very recent report by Hsieh and Luo (2004) indicates that inaccuracies in the AMBER forcefield are related with the 1–4 electrostatic energy term, which is subject to large fluctuations. The same authors are developing a modified forcefield that is able to improve the accuracy of the MM/PBSA energy estimation.

Another issue that is worth noting is that the ranking of the native structure among decoys might not be an appropriate way to assess the quality of the free energy estimator, because for some decoy sets many or most structures have very low RMSDs from the native structure. In view of the approximations involved in the methodology, we should tolerate mis-ranking of the native structure compared with

very similar structures. Since the native structure is not available in a real prediction test, we should be more concerned with ranking near-native structures at the first places, rather than ranking the native structure first. This principle also conforms to the spirit of the colony energy approach.

Application of MM/PBSA free energy estimator to the 32 decoy sets, modeling eight-residue loops, ranks the native structure first in 18 sets, second in five other sets, within the first five in three other sets. Overall, the native structure is ranked among the lowest five energy structures in 81% of the sets.

When compared with the results of de Bakker et al. (2003), the performance of the MM/PBSA model is slightly worse than that of the AMBER/GBSA model. The protocol employed here is, however, different and these small differences in performance may be possibly attributed to the fact that, contrary to the cited work, we also allow the framework to move during minimization.

The superposition of loops on native structures performed for the lowest energy structure in each decoy set, however, shows that the RMSD is rather low for almost all best-ranked structures (<1.0 Å for 26 out of 32 decoy sets). A number of exceptions are found and closer inspection showed that for three of them (1phf-76, 1thw-18, and 1poa-71) unusual backbone conformations (falling in disallowed

Ramachandran regions) are found in the best energy minimized structures and not in the native minimized structure. These unusual conformations are not properly penalized by MM/PBSA free energy estimators. As a remedy, each backbone conformation falling in disallowed regions and in generously allowed regions were weighted 40 kcal mol⁻¹ and 5 kcal mol⁻¹, respectively. These figures were empirically found to work, but the problem of correct backbone energy estimation must be definitely solved in a more accurate way. The work of Hsieh and Luo (2004) is going in this direction. Besides ad hoc countermeasures, the presence of bad geometry or unusual conformations points out inaccuracies of the force field and nonrobustness of the minimization protocol.

When unfavorable conformations are weighted as mentioned above, results are similar as before, as far as ranking of the native structure is concerned but improve drastically as far as RMSD from the native structure is concerned. This

can be appreciated from results reported in Table 2. In 28 out of 32 decoy sets, the lowest free energy structure has a RMSD from native, computed by superimposing the modeled loop, lower than 0.7 Å. The highest RMSD among all 32 sets is, however, as low as 2.0 Å.

When the native structure is removed from each set, results are more adherent to a real prediction game. We note first that the number of models considered here for each target is limited to 50, and therefore the colony energy approach can not be safely applied on these datasets. Another problem is that even if there is a model with low RMSD from the native structure, side chains are not guaranteed to be properly modeled and therefore it is not obvious that the free energy will be low. Notwithstanding these issues, the number of the lowest free energy models that exhibit a RMSD from the native structure lower than 2.0 Å is 26, and in most cases the RMSD is consistently lower than the average for that set (Table 3).

Table 2. Results for de Bakker et al. decoy sets

Decoy set	Energy rank native	RMSD rank best energy	Global RMSD (Å)	Loop RMSD (Å)	Energy (kcal mol)
1351-84	7	8	0.37	0.62	-3569.90
1alc-34	1	1	0.34	0.16	-2061.91
1art-88	1	1	0.26	0.23	-9464.76
1btl-50	2	2	0.24	0.67	-7514.37
1cbs-55	26	3	0.27	0.43	-3557.35
1ddt-127	1	1	0.32	0.52	-8819.00
1fnd-262	4	8	0.23	0.23	-6105.90
1gky-72	1	1	0.21	0.09	-3784.48
1hfc-142	1	1	0.26	0.16	-3825.70
1iab-48	1	1	0.23	0.22	-3803.07
1nar-192	12	17	0.29	0.62	-6003.00
1oyc-80	2	21	0.29	0.53	-10369.00
1phf-76	1	1	0.24	0.16	-10919.80
1poa-71	5	47	0.83	1.31	-2591.10
1prn-150	1	1	0.22	0.19	-4811.50
1sbp-107	1	1	0.22	0.15	-7339.10
1thw-18	1	1	0.25	0.16	-4225.07
1tml-187	1	1	0.22	0.17	-5108.20
1tys-83	1	1	0.20	0.23	-6698.36
1xnb-99	1	1	0.21	0.13	-3309.00
2ayh-123	1	1	0.24	0.22	-2910.30
2cmd-270	1	1	0.20	0.18	-4923.60
2ctc-89	1	1	0.23	0.14	-5680.56
2dri-161	1	1	0.23	0.13	-5727.80
2exo-262	24	6	0.27	0.24	-7740.00
2fox-88	1	1	0.23	0.12	-2501.14
2ran-26	9	28	0.74	1.41	-9570.30
2sga-8	40	8	0.25	0.62	-2631.79
3grs-424	13	35	0.60	2.03	-7781.00
4enl-24	5	10	0.28	0.35	-8769.10
5p21-45	1	1	0.22	0.11	-5199.15
8dfr-65	12	34	0.60	1.73	-4046.10

The decoy set is reported in the first column followed by the ranking in energy of the native structure, the ranking in RMSD of the best energy structure the RMSD computed on C_α carbons for the best energy structure, the RMSD computed on loop C_α carbons for the best energy structure, and its MM/PBSA energy.

Table 3. Results for de Bakker et al. decoys sets not including the native structure

Decoy set	Loop RMSD best energy	Min RMSD (Å)	Mean RMSD (Å)	Max RMSD (Å)
1351-84	0.6	0.3	1.3 (0.6)	2.8
1alc-34	2.2	1.4	2.3 (± 0.3)	2.9
1art-88	2.3	0.3	1.4 (± 0.6)	2.5
1btl-50	0.7	0.4	1.5 (± 0.7)	2.8
1cbs-55	0.4	0.4	1.5 (± 0.8)	3.1
1ddt-127	0.5	0.5	1.6 (± 0.4)	2.0
1fnd-262	0.2	0.2	0.6 (± 0.3)	1.5
1gky-72	0.6	0.5	2.0 (± 0.7)	3.0
1hfc-142	0.4	0.2	0.7 (± 0.5)	2.3
1iab-48	1.9	1.0	1.6 (± 0.4)	2.6
1nar-192	0.6	0.2	0.8 (± 0.3)	1.5
1oyc-80	0.5	0.5	0.8 (± 0.3)	1.7
1phf-76	1.5	0.8	1.7 (± 0.4)	2.5
1poa-71	1.3	0.5	1.2 (± 0.4)	2.5
1prn-150	2.2	0.5	2.0 (± 0.4)	2.7
1sbp-107	2.3	0.6	2.4 (± 0.6)	3.2
1thw-18	1.0	0.8	3.2 (± 1.0)	4.2
1tml-187	1.1	0.5	2.7 (± 1.0)	4.0
1tys-83	1.9	0.4	2.1 (± 0.6)	3.2
1xnb-99	1.3	0.3	0.8 (± 0.3)	1.4
2ayh-123	1.4	0.8	1.9 (± 0.4)	2.7
2cmd-270	3.4	1.8	3.2 (± 0.4)	3.8
2ctc-89	0.2	0.2	1.9 (± 0.7)	3.2
2dri-161	1.7	1.0	2.1 (± 0.4)	2.8
2exo-262	0.3	0.2	0.5 (± 0.2)	0.9
2fox-88	1.1	0.3	1.8 (± 0.6)	3.3
2ran-26	1.4	0.9	2.4 (± 0.5)	3.2
2sga-8	0.6	0.3	0.8 (± 0.2)	1.1
3grs-424	2.0	1.3	2.0 (± 0.3)	2.6
4enl-24	0.4	0.3	1.0 (± 0.4)	2.4
5p21-45	0.2	0.2	0.5 (± 0.2)	1.1
8dfr-65	1.7	1.3	1.7 (± 0.3)	2.7

The decoy set is reported in the first column followed by the RMSD computed on loop C $_{\alpha}$ carbons for the best MM/PBSA energy structure; minimum, average, and maximum RMSDs computed on loop C $_{\alpha}$ carbons for the same decoy set.

abm_database set

The abm_database set of loop decoys deposited in the Decoys'r'us database (<http://dd.stanford.edu/>) (Samudrala and Levitt 2000), was tested next. This is a set of loop decoys for the d1.3 antibody (original PDB accession no. 1vfa) generated using a database approach for the main chain positions and a graph-theoretic algorithm for the side chains (Samudrala and Moult 1998). The size of the four loops modeled is 8 and 9 residues. This set is attractive because RMSDs span a wide range of values, including native-like conformations.

To show how difficult native loop structure recognition is in this set (and in general), we plotted the contact energy estimated by a successful SEEF versus the RMSD computed on loop C $_{\alpha}$ atoms, keeping framework residues fixed (Fig. 1). It should be noted that the same empirical contact

energies performed well on widely used decoy sets (Berrera et al 2003).

A slightly more time-consuming protocol has been applied for MM/PBSA free energy calculations, which was later replaced by the protocol described in the Materials and Methods section, with the same results. Fifty steepest descent and 50 conjugate gradients energy minimization steps were applied on the C $_{\alpha}$ restrained structure, followed by 150 steepest descent and 150 conjugate gradients minimization steps. For all conjugate gradient steps, the conjugate gradient cycle was set to 10. The plot of MM/PBSA energy versus the RMSD, computed on loop C $_{\alpha}$ atoms, after superposition of C $_{\alpha}$ atoms not in the loop, is reported in Figure 2

There is no strict correlation between the computed free energy and the RMSD because small differences in conformation may correspond to rather large energy changes. For instance, this is the case for loop 190, whose global RMSD from native, after minimization, is even lower than the RMSD of the native loop itself, after minimization. The structure, however, has very bad van der Waals contacts and, therefore, its final energy is more than 2000 kcal mol $^{-1}$ higher than for the minimized native structure. The lowest free energy structures, however, also possess the lowest RMSD from the native structure.

When free energy components are analyzed, it is seen that, for this decoy set, the native structure is also correctly discriminated by MM energy. In general, a correlation between MM energy and RMSD is found. The gap in energy between the native structure and the best decoy is larger for MM/PBSA than for MM energy (29.6 and 23.5 kcal mol $^{-1}$, respectively) and slightly larger than that found with MM/GBSA energy (6.8 kcal mol $^{-1}$).

Both MM/PBSA and MM energy rankings place the lowest RMSD structures among the lowest free energy ones. For instance, of the five lowest MM/PBSA and MM free energy structures none exceeds a RMSD (over loop C $_{\alpha}$ carbons) of 0.7 Å and 1.1 Å from the native structure, respectively.

Figure 1. Empirical contact energies for abm_dataset, computed as described in Berrera et al. (2003), vs. RMSD from the native structure. RMSD is computed on loop C $_{\alpha}$ atoms keeping the framework fixed. The horizontal line represents the energy of the native structure.

Figure 2. MM/PBSA energies for abm_dataset vs. RMSD from the native structure. RMSD is computed on loop C_{α} atoms after superposition of C_{α} atoms not belonging to the modeled loop.

MM/GBSA, on this particular decoy set, performs worse than both the MM and the MM/PBSA energy estimators. The PBSA free energy versus RMSD plot shows a weak anticorrelation. The solvent accessibility distribution shows a weak correlation with RMSD.

The opposing effects of direct Coulombic and solvation electrostatic terms, which have been noted before, are also found for this decoy set with correlation coefficient -0.84 .

The total electrostatic component of the free energy shows a weak correlation with RMSD. For this particular decoy set, the free energy is dominated by MM terms. Due to the large uncertainty on parameters employed for continuum calculation, we test the application of different scaling factors on the electrostatic contributions to the free energy (direct Coulombic and PB solvation components) and on the term proportional to solvent accessibility. In this way we determine a plausible range for both dielectric constant and surface tension coefficient, to be used in this approach.

The solvation free energy estimated by the PB equation scales approximately with the inverse of the dielectric constant, just like the direct Coulombic term. We can therefore multiply both terms, for consistency, by a range of constants and assess what values are feasible or unfeasible. For this particular decoy set, the surface tension coefficient can range between 0 and $1200 \text{ cal mol}^{-1} \text{ \AA}^{-2}$ with optimal separation between the lowest free energy native structure and the next lowest free energy decoy achieved for $0 \text{ cal mol}^{-1} \text{ \AA}^{-2}$. The dielectric constant can range between 1.0 and 5.0 with the largest energy separation, between native and next lowest energy structure, obtained at 1.0.

Although the separation in energy of the native structure from the best decoy may not be a good criterion, overall these observations show that for standard choices of force-field parameters and minimization protocols, the dielectric constant values usually employed for electrostatic calculations (between 1.0 and 4.0) are able to discriminate native structures from decoys and are therefore reliable. The assessment of reasonable values for the surface tension coefficient requires a different test set.

The colony energy approach proposed by Honig and co-workers (Xiang et al. 2002), modified as described in the Materials and Methods section, has been applied on this set. Pairwise RMSDs have been computed on all protein C_{α} carbons because different loops have been modeled. For this reason the computed pairwise RMSDs are extremely low, ranging from $0.2 \text{ \AA}$ to $1.25 \text{ \AA}$. The correlation between the RMSD from the native structure computed on all protein C_{α} atoms and on loop C_{α} atoms are approximately linearly dependent with a correlation coefficient of 0.994 . To prevent that all colony energies collapse toward the lowest energy, as discussed in the Materials and Methods section, a $0.2 \text{ \AA}$ threshold, t , lower than the lowest RMSD of minimized structures from native, has been chosen. The native structure was removed from the ensemble so that no information about RMSD from native is used to compute the colony energy reported in Figure 3. Only after colony energies are computed, they are plotted against the RMSD of the modeled loop from the native loop, computed after superposition of the remaining residues. The correlation of colony energy with RMSD from the native structure is therefore not artifactual and witnesses the quality of the method and protocol used.

ligd octamer decoys

The decoy set generated according to the procedure described in Materials and Methods is different from other sets in that all octamers of a small protein are modeled instead of just one loop. Similar to what is described above, we have added to the energy an unfavorable energy term for those structures containing unusual backbone conformations. Here, the MM term alone is not able to discriminate the native structure, although the latter is ranked second in energy. The performance of MM energy alone is rather poor and it assigns lowest energy to structures whose RMSD from the native structure is high (larger than $4.0 \text{ \AA}$). MM/PBSA energy improves the performance of MM energy alone, but four of the 10 lowest energy structures have RMSD from native larger than $2.0 \text{ \AA}$ (Fig. 4). MM/GBSA energy performs slightly better and assigns the 10 lowest energies

Figure 3. MM/PBSA colony energies for abm_dataset vs. RMSD from the native structure. RMSD is computed on loop C_{α} atoms after superposition of C_{α} atoms not belonging to the modeled loop. The threshold t is $0.2 \text{ \AA}$.

to structures whose RMSD from native is below 1.0 Å, although separation of the native structures from decoys is much smaller compared with MM/PBSA (0.5 kcal mol⁻¹ vs. 14.6 kcal mol⁻¹).

We have also tested here the scaling factors applied to electrostatic and SASA energy terms. For electrostatics, the scaling factor may range from 1.0 to 0.095 (corresponding roughly to dielectric constants in the range 1.0–11.0), with optimal discrimination for a value of 1.0, while the surface tension coefficient may vary between 0.0 cal mol⁻¹ and 700.0 cal mol⁻¹ Å⁻² with optimal discrimination achieved at about 250 cal mol⁻¹ Å⁻².

Anticorrelation is also found here between Coulombic and electrostatic solvation energy with correlation coefficient -0.87.

Application of the colony energy approach is somewhat less effective here, as far as the smooth appearance of the curve is concerned, because the set contains 20 decoys for all octamers in the sequence and there are therefore many uncorrelated structures. For this reason, it may happen that no low energy structure is present in the “colony.” The effect of the threshold value may be appreciated in this particular case. The choice of threshold 0.5 Å and 1.0 Å (Figs. 5–7) produces increasingly regularized curves.

In this case, even with the choice of a 0.2 Å threshold, the relationship between RMSD (computed on loop residues) and free energy becomes apparent. Even with such a conservative choice, the largest RMSD (computed on loop residues) from the native structure is 1.0 Å, among the 10 lowest free energy structures. It is worth noting that with all choices of threshold, the lowest free energy structure has an RMSD from the native structure (computed on the loop alone, as described above) <1.0 Å. When the threshold value 1.0 Å is used, best results are obtained with the lowest free energy structure showing an RMSD of 0.18 Å and approximately the 150 lowest free energy structures showing an RMSD from native <1.0 Å.

In this set, the colony energy approach improves the choice of a plausible predictive model significantly. Since the native structure is completely ignored in computing the colony energy for each model, the predictive properties of the approach should not be considered artifactual.

Conclusions

The MM/PBSA approach has been clearly and completely formalized by Kollman et al. (2000), although similar approaches to the estimation of free energies from molecular structures have been used by many other authors (see, e.g., Baginski et al. 1997; Vorobjev et al. 1998). Two problems related to free energy estimation from molecular structures are accurate estimation of solvent effects and entropic contributions.

Recent work of our and other groups has demonstrated that the approach based on molecular mechanics for estimation of intramolecular energy and on PB equation and SASA for estimating solvation effects is accurate (see, e.g., Kollman et al. 2000; Feig and Brooks 2002; Fogolari et al. 2003). Estimation of entropic effects is more difficult because it requires exploration of conformational space.

Moreover, the strong dependence of the estimated energy on small coordinate variations poses a question on the robustness of the approach.

The results described above show that estimation of a solute potential of mean force (intramolecular energy and solvation effects) by the MM/PBSA approach, using for instance the protocol implemented here, is overall accurate. Possible sources of inaccuracies appear to be related to the MM energy estimation of unusual backbone conformations.

As far as the MM/PBSA free energy estimation is concerned, our results compare well with similar works. In particular, the comparison with the work of de Bakker et al. (2003), who studied the same decoy sets using the AMBER/GBSA method, shows that the two approaches have very similar performances. The average RMSD from native of the best model (computed on all backbone atoms, after superposition of framework residues) is slightly better here (by ~0.1 Å), although the native structure is found among the five lowest free energy models in 75% of the sets (~15% less than using the AMBER/GBSA method).

Comparison with the results obtained using the AMBER/GBSA method suggests that the GBSA model is as accurate as the PBSA model, at least at the level of accuracy required by the decoy sets studied. Native structure for almost all decoys considered is found in a deep energy well as expected.

Although native-like structures for most cases have low free energy, it appears that application of the colony energy approach considerably improves the relationship between the RMSD from the native structure and estimated free energy. In particular, the approach has the beneficial effect to greatly hamper the dependence of the estimated free energy on conformational variations. The plots reported in this work, built without making use of information on the native structure, show a much smoother appearance than the corresponding raw MM/PBSA versus RMSD plots.

A similar behavior has been observed by Honig and co-workers (Xiang et al. 2002). Here, the application of the colony energy approach is even more effective due to the more detailed MM/PBSA free energy estimator.

In summary, our results suggest that the MM/PBSA approach with prior minimization using accurate estimation of solvent effects, coupled with the colony approach for estimating entropic effects and for smoothing free energy predictions, is well suited for evaluating predictive protein models, and for the analysis of protein folding simulations.

Materials and methods

MM/PBSA methodology

The free energy of each protein conformation (predictive model), with reference, say, to the unfolded state, may be computed using standard statistical thermodynamics, according to the framework reviewed by Gilson et al. (1997), in the context of ligand binding. If we denote by C the ensemble of conformations corresponding to the macroscopic state C (e.g., this can be the folded or the unfolded state), the free energy of conformation C may be written as

$$\Delta G^C = -kT \ln \frac{\int_C \exp\left(-\frac{U}{kT}\right) d\tilde{r}_1, \dots, d\tilde{r}_n, d\tilde{r}_{n+1}, \dots, d\tilde{r}_N}{\int \exp\left(-\frac{U}{kT}\right) d\tilde{r}_1, \dots, d\tilde{r}_n, d\tilde{r}_{n+1}, \dots, d\tilde{r}_N}$$

where U is the potential energy and indices $1, \dots, n$ run over solute coordinates and indices $n+1, \dots, N$ run over solvent coordinates. Now we may consider rewriting the above equation in terms of the protein potential of mean force $W(\tilde{r}_1, \dots, \tilde{r}_n)$, which is by definition (Hill 1956) such that:

$$\exp\left(-\frac{W(\tilde{r}_1, \dots, \tilde{r}_n)}{kT}\right) = \frac{V^n \int \exp\left(-\frac{U}{kT}\right) d\tilde{r}_{n+1}, \dots, d\tilde{r}_N}{\int \exp\left(-\frac{U}{kT}\right) d\tilde{r}_1, \dots, d\tilde{r}_n, d\tilde{r}_{n+1}, \dots, d\tilde{r}_N}$$

where V is the volume.

With this notation the free energy of the macroscopic state C , unless a constant term, is

$$\Delta G^C = -kT \ln \frac{\int_C \exp\left(-\frac{W(\tilde{r}_1, \dots, \tilde{r}_n)}{kT}\right) d\tilde{r}_1, \dots, d\tilde{r}_n}{\int \exp\left(-\frac{W(\tilde{r}_1, \dots, \tilde{r}_n)}{kT}\right) d\tilde{r}_1, \dots, d\tilde{r}_n}$$

We assume that each predictive model is a sample from a macroscopic state C . To estimate the free energy of the macroscopic state C we must be able to (1) reasonably estimate the potential of mean force $W(\tilde{r}_1, \dots, \tilde{r}_n)$; (2) treat the integral in a reasonable fashion.

It is often assumed that the potential of mean force $W(\tilde{r}_1, \dots, \tilde{r}_n)$ may be written as the sum of a solute energy term $U(\tilde{r}_1, \dots, \tilde{r}_n)$ and a solvation free energy term that is further split in a polar (electrostatic) and a nonpolar (hydrophobic) term (Honig et al. 1993; Roux and Simonson 1999).

$$W(\tilde{r}_1, \dots, \tilde{r}_n) = U(\tilde{r}_1, \dots, \tilde{r}_n) + \Delta G^{\text{polar}}(\tilde{r}_1, \dots, \tilde{r}_n) + \Delta G^{\text{non-polar}}(\tilde{r}_1, \dots, \tilde{r}_n)$$

The free energy of the macroscopic state C is often estimated as the sum of the potential of mean force for one or more sampled microscopic states and an entropic contribution arising from restrictions in conformational and/or translational and rotational freedom (see, e.g., Finkelstein and Janin 1989; Doig and Sternberg 1995):

$$\Delta G^C = W - T\Delta S$$

If we assume that each conformation is a sample from a basin of the same number of conformations possessing the same potential of mean force, i.e., that ΔS is the same for all different conformations sampled, then the potential of mean force may be used to discriminate the native (stable) versus unfolded or misfolded (non-stable) conformations. The rationale behind this choice is that folded conformations are usually more compact and therefore conformationally restricted than unfolded states. As a consequence, the entropic term will favor unfolded states. The potential of mean force component of the free energy should therefore be, a fortiori, lower for folded states than for unfolded states.

For folded or misfolded states, the entropic restrictions should be comparable and therefore the enthalpic component should be sufficient to discriminate stable versus less stable, but equally compact, states.

We adopted, where possible, the colony energy approach of Honig and coworkers (Xiang et al. 2002), described later.

The solute energy term $U(\tilde{r}_1, \dots, \tilde{r}_n)$ has been computed using CHARMM (version 27b2), a classical and well-tested molecular mechanics forcefield (Brooks et al. 1983; MacKerell et al. 1998). All structures have been energy minimized in a two-step fashion, except where indicated. First, all C_α coordinates have been fixed and the energy minimized with 20 steepest descent and 30 conjugate gradients steps to remove high energy spots, like van der Waals clashes. During this minimization, a distance-dependent dielectric constant has been employed ($\epsilon = 1/r$) with a cutoff of 12 Å. In the second step, the structure is totally unrestrained and a more accurate solvent model (the Generalized Born Solvent Accessible model) has been employed using the parameters suggested by Qiu et al. (1997), to reduce artifacts ensuing from missing solvent. First, 50 steepest descent minimization steps have been performed followed by 200 conjugate gradients minimization steps. After minimization, the energy was evaluated with no cutoff and with dielectric constant 1.0, in order to be consistent with the forcefield used (see the discussion below).

$\Delta G^{\text{polar}}(\tilde{r}_1, \dots, \tilde{r}_n)$ has been computed according to the PB theoretical framework (Marcus 1955; Sharp and Honig 1990; Zhou 1994; Fogolari and Briggs 1997) as the difference in free energy for the hypothetical charging process of the solute in vacuo and in ionic solvent. To get rid of self energy contributions, which are strongly dependent on grid mesh and positioning, two identical computations are performed differing only in the solvent dielectric constant and ionic strength (80 and 100 mM, respectively, for solvent calculation and 1 and 0 mM, respectively, for the in vacuo calculation). All calculations have been performed using the program UHBD (Madura et al. 1994, 1995).

The choice of 1.0 for inner dielectric constant requires a word of caution. In most studies involving the MM/PBSA approach or some of its variants, the inner dielectric constant is assumed to be in the range 2–4, typical of organic molecules. In applications exploring molecular dynamics within the MM/PBSA approach it was found that substantially higher dielectric constant values (4.0–17.0) must be employed to achieve reasonable molecular dynamics trajectories, thus posing serious problems of consistency between the MM and PB energy estimations (Fogolari et al. 2001, 2003; Lu et al. 2002). A clear discussion about the choice of dielectric constant has been given by Schutz and Warshel (2001). Simonson (1999) has accurately estimated the dielectric constant resulting from relaxation of protein dipoles and net charges. In the protocol presented here, structures are minimized and therefore some relaxation is taken into account. Moreover, we notice that Ramachandran maps generated with the protocol chosen here and using

dielectric constant values larger than 1.0 do not show the low energy region corresponding to typical backbone conformations. Thus, we decided to adopt the value of 1.0 for the dielectric constant more consistent with the MM energy estimation and correctly estimating the energy of backbone conformations for model compounds.

The following protocol has been used: First, the linear PB equation is solved on a large grid, with 40^3 grid points spaced 2.5 Å, centered on the protein to obtain boundary conditions for the following focusing steps. Then the same equation is solved on smaller and finer grids (33^3 points spaced 0.45 Å) centered on each amino acid center of geometry. In each of the focusing steps, the potential at each atomic position of the amino acid is stored and later used to estimate the electrostatic free energy of the protein, which is expressed for the linear PB equation as:

$$\Delta G = \sum_i \frac{q_i \phi_i}{2},$$

where q_i is the charge of the i^{th} atom and ϕ_i is the potential at the same atom. The linear PB equation is appropriate for moderately charged biomolecules (Fogolari et al. 1999). To speed up calculations, the inner dielectric constant was assigned to space regions within van der Waals radii. The solvent-accessible surface was generated by inflating all atomic radii by 1.4 Å and considering the inflated van der Waals surface.

Dielectric boundaries have been smoothed according to the procedure of Davis and McCammon (1991b).

$\Delta G^{\text{nonpolar}}(\vec{r}_1, \dots, \vec{r}_n)$ is taken to be proportional to the SASA A , that is, $\Delta G^{\text{nonpolar}}(\vec{r}_1, \dots, \vec{r}_n) = \gamma A$. Although a large range of values have been used in the literature for the surface tension coefficient γ , we have used a value of 20 cal mol⁻¹ Å⁻² lower than the 50 cal/mol Å⁻² value proposed by Nicholls et al. (1991) for proteins. The rationale for this choice is that, in the absence of solvent molecules, unbalanced intramolecular van der Waals forces provide a strong tendency toward collapse anyway. The employed surface tension value has been previously empirically determined by considering change in van der Waals energy upon folding of a small protein (Fogolari et al. 2003).

The colony energy approach

The problem of estimating the entropic term due to the solute is difficult to solve because the number of conformations accessible from each microscopic state depends on the local energy landscape, and this is not easily available. A simple approach is just to neglect entropic effects assuming that for compact structures they should be similar and therefore cancel out in comparison. This approach has been implicitly assumed in many studies. A few other approaches have been presented that deal explicitly with entropy and deserve being mentioned. A simple line of thought is to consider how many rotamers for each torsion angle can be explored in each state and consider the number of microstates compatible with each macrostate simply as the product of the number of rotamers. The resulting entropy is therefore:

$$\Delta S_{0 \rightarrow 1} = -kT \ln \prod_i \frac{n_i^1}{n_i^0} = -kT \sum_i \ln \frac{n_i^1}{n_i^0}.$$

Typically for side-chain torsion angles upon folding, an entropic restriction of

$$-kT \ln \frac{1}{3}$$

is assumed (Doig and Sternberg 1995).

Kollman et al. (2000) estimated the entropy of each conformation as the entropy computed for the energy minimized conformation in the harmonic approximation of the potential close to the minimum. This procedure faces the drawback of necessary extensive minimization, which is known to introduce artifacts and should employ some approximation of solvation effects.

A different approach, requiring molecular dynamics simulation, has been used by Baginski et al. (1997), who estimated the solute entropy via thermodynamic integration with reference to a model where all atoms are not interacting and harmonically constrained to the initial position.

Vorobjev et al. (1998) estimated solute entropy from the covariance fluctuation matrix computed on a molecular dynamics trajectory. This approach, however, also requires molecular dynamics simulation.

Honig and coworkers (Xiang et al. 2002) adopted a different approach that is particularly useful for free energy estimation of predictive models. The basic idea, applied to loop modeling, is that each conformation is a sample of a colony of similar conformations. If we assume for the time being that all these conformations are reachable within the colony, the free energy corresponding to the colony would be

$$\Delta G_i^C = -kT \ln \frac{\sum_{j \in C} \exp\left(\frac{-W(r_1^j, \dots, r_n^j)}{kT}\right)}{\sum_i \exp\left(\frac{-W(r_1^i, \dots, r_n^i)}{kT}\right)}$$

To estimate the contributions of mostly nonsampled conformations to the colony free energy, the same set of models under scrutiny is used. In practice, (1) it is assumed that the set of generated structures samples conformational space in a statistically significant way (hence the requirement of a large number of structures) and that the sum over all conformations defined as macrostate C is restricted to the sampled conformations; and (2) a weight function ranging from 1 (e.g., for identical backbone conformations) to 0 (for unrelated backbone conformations) is used rather than precisely defining which microstates should be included in the sum. The weight function empirically determined by Honig and coworkers (Xiang et al. 2002) takes the general form

$$w_{ij} = \exp\left(\frac{RMSD_{ij}^3}{t^3}\right)$$

where i is the index of the reference structure, j is the index of each other sampled conformation, and t is a threshold RMSD that defines the similarity between two conformations. With this weight function the free energy corresponding to the i^{th} conformation is:

$$\Delta G_i^C = -kT \ln \frac{\sum_j \exp\left(\frac{-W(r_1^j, \dots, r_n^j)}{kT} - \frac{rmsd_{ij}^3}{t^3}\right)}{\sum_j \exp\left(\frac{-W(r_1^j, \dots, r_n^j)}{kT}\right)}$$

Honig and coworkers (Xiang et al. 2002) used a value for t tailored to the loop modeling problem, in particular t was chosen equal to

Figure 4. MM/PBSA energies for ligd decoys dataset vs. RMSD from the native structure. RMSD is computed after superposition of loop C_α atoms.

$$\frac{1}{(6L)^3},$$

where L is the length of the modeled loop, corresponding, for a loop of eight residues, to a value of 3.6 Å. As already noted in the original paper, besides modeling entropic effects, the approach has the further beneficial effect of smoothing the energy versus RMSD plot. This is particularly important because often, due to minor errors in placing side-chain atoms, the computed energy may be very high. If a low energy structure has a significant weight in the sum over other structures, the colony energy of the reference structure will be lower. Smoothing the appearance of the plot allows to immediately pick up the reliability of a free energy estimator. In particular, if the free energy estimator does not consistently assign lower energy to native-like structures, the plot will be smooth, but it will not show the RMSD–energy correlation that is the ultimate goal of all free energy estimators. With these premises how should the parameter t be chosen? We suggest that the choice should roughly match or be less than the attempted accuracy of the prediction. In other words, t should be set equal to or less than the RMSD value we accept for defining two structures similar. For instance, if the goal is ab-initio protein prediction, perhaps a value of 4.0–6.0 Å should be appropriate. For accurate loop modeling, values in the range 1.0–2.0 Å should be appropriate.

For the sake of clarity let us consider two extreme cases: If a t value of 0 Å is chosen, then the weight is 0.0 for all conformations except for the reference conformation i , which has a weight of 1.0. The colony approach in this case is simply not applied to the data. If a very large t value is chosen, all conformations will receive the same weight close to 1.0 and all conformations will have the same colony energy. In this case the plot will display a single horizontal line.

Figure 5. MM/PBSA colony energies for ligd decoys dataset vs. RMSD from the native structure. The threshold is 0.2 Å. RMSD is computed after superposition of loop C_α atoms.

Figure 6. MM/PBSA colony energies for ligd decoys dataset vs. RMSD from the native structure. The threshold is 0.5 Å. RMSD is computed after superposition of loop C_α atoms.

Ligd octamer decoy generation

The ligd octamer decoys were generated with a fast divide-and-conquer method (Tosatto et al. 2002). A sliding window running from the N to the C terminus was placed on protein G immunoglobulin binding domain (PDB code: ligd), a small 61-residue protein containing both α-helices and β-strands. All eight residue fragments under the sliding window were individually predicted using the fast loop modeling method. A total of 500 conformations were constructed and ranked according to coarse geometric and energy criteria.

The top 20 solutions were retained as part of the decoy set. Since the loop modeling method does not predict the conformation of side-chain atoms, these were predicted using the SCWRL3 method (Canutescu et al. 2003). The whole procedure guarantees the generation of an ensemble of decoys that differ from the native structure in a single eight-residue fragment. Because of the selection process, all decoys are also ensured to have at least acceptable backbone geometry and interaction energies.

Acknowledgments

Prof. H. Molinari and Prof. G. Valle are gratefully acknowledged for useful discussions and comments on the manuscript. Dr. Mark de Pisto is gratefully acknowledged for providing the Rapper decoy sets. F.F. thanks Drs. G. Tecchioli and P. Zuccato of Exadron, the HPC Division of the Eurotech Group, for providing hardware and expert technical assistance. S.T. is funded by a “Rientro dei cervelli” grant from the Italian Ministry for Education, University, and Research. Part of the research was also funded by Telethon (Italy), grant number B057-I.

Figure 7. MM/PBSA colony energies for ligd decoys dataset vs. RMSD from the native structure. The threshold is 1.0 Å. RMSD is computed after superposition of loop C_α atoms.

References

- Ajay and Murcko, M.A. 1995. Computational methods to predict binding free energy in ligand-receptor complexes. *J. Med. Chem.* **38**: 4953–4967.
- Anfinsen, C.B. 1973. Principles that govern the folding of protein chains. *Science* **181**: 23–230.
- Baginski, M., Fogolari, F., and Briggs, J.M. 1997. Electrostatic and non-electrostatic contributions to the binding free energies of anthracycline antibiotics to DNA. *J. Mol. Biol.* **274**: 253–267.
- Bashford, D. and Case, D.A. 2000. Generalized born models of macromolecular solvation effects. *Ann. Rev. Phys. Chem.* **51**: 129–152.
- Berrera, M., Molinari, H., and Fogolari, F. 2003. Amino acid empirical contact energy definitions for fold recognition in the space of contact maps. *BMC Bioinformatics* **4**: 8.
- Blokzijl, W. and Engberts, J.B.F.N. 1993. Hydrophobic effects. Opinions and facts. *Angew. Chem. Int. Ed. Engl.* **32**: 1545–1579.
- Bonneau, R. and Baker, D. 2001. Ab initio protein structure prediction: Progress and prospects. *Ann. Rev. Biophys. Biomol. Struct.* **30**: 173–189.
- Brooks, B.R., Brucoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., and Karplus, M. 1983. Charmm: A program for macromolecular energy minimization and dynamics calculations. *J. Comput. Chem.* **4**: 187–217.
- Canutescu, A.A., Shelenkov, A.A., and Dunbrack, R.L. 2003. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.* **12**: 2001–2014.
- Davis, M.E. and McCammon, J.A. 1991a. Electrostatic in biomolecular structure and dynamics. *Chem. Rev.* **90**: 509–521.
- . 1991b. Dielectric boundary smoothing in finite difference solutions of the Poisson equation: An approach to improve accuracy and convergence. *J. Comp. Chem.* **12**: 909–912.
- de Bakker, P.I.W., De Pristo, M.A., Burke, D.F., and Blundell, T.L. 2003. Ab initio construction of polypeptide fragments: Accuracy of loop decoy discrimination by an all-atom statistical potential and the amber force field with the generalized born solvation model. *Proteins* **51**: 21–40.
- Doig, A.J. and Sternberg, M.J. 1995. Side-chain conformational entropy in protein folding. *Protein Sci.* **4**: 2247–2251.
- Dominy, B.N. and Brooks III, C.L. 2002. Identifying native-like protein structures using physics-based potentials. *J. Comput. Chem.* **23**: 147–160.
- Eisenberg, D. and McLachlan, A.D. 1986. Solvation energy in protein folding and binding. *Nature* **319**: 199–203.
- Feig, M. and Brooks III, C.L. 2002. Evaluating casp4 predictions with physical energy functions. *Proteins* **49**: 232–245.
- Finkelstein, A.V. and Janin, J. 1989. The price of lost freedom: Entropy of bimolecular complex formation. *Protein Eng.* **3**: 1–3.
- Fogolari, F. and Briggs, J.M. 1997. On the variational approach to the Poisson-Boltzmann free energies. *Chem. Phys. Lett.* **281**: 135–139.
- Fogolari, F., Zuccato, P., Esposito, G., and Viglino, P. 1999. Biomolecular electrostatics with the linearized Poisson-Boltzmann equation. *Biophys. J.* **76**: 1–16.
- Fogolari, F., Esposito, G., Viglino, P., and Molinari, H. 2001. Molecular mechanics and dynamics of biomolecules using a solvent continuum model. *J. Comput. Chem.* **22**: 1830–1842.
- Fogolari, F., Brigo, A., and Molinari, H. 2002. The Poisson-Boltzmann equation for biomolecular electrostatics: A tool for structural biology. *J. Mol. Recogn.* **15**: 377–392.
- . 2003. Protocol for MM/PBSA molecular dynamics simulations of proteins. *Biophys. J.* **85**: 159–166.
- Fraternali, F. and Van Gunsteren, W.F. 1996. An efficient mean solvation force model for use in molecular dynamics simulations of proteins in aqueous solution. *J. Mol. Biol.* **256**: 939–948.
- 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.
- Gilson, M.K., Given, J.A., Bush, B.L., and McCammon, J.A. 1997. The statistical-thermodynamic basis for computation of binding affinities: A critical review. *Biophys. J.* **72**: 1047–1069.
- Hill, T. 1956. *An introduction to statistical mechanics*. Dover Publications, New York.
- Holm, L. and Sander, C. 1992. Evaluation of protein models by atomic solvation preference. *J. Mol. Biol.* **225**: 93–105.
- Honig, B. and Nicholls, A. 1995. Classical electrostatic in biology and chemistry. *Science* **268**: 1144–1149.
- Honig, B. and Yang, A.-S. 1995. Free energy balance in protein folding. *Adv. Prot. Chem.* **46**: 27–58.
- Honig, B., Sharp, K., and Yang, A.-S. 1993. Macroscopic models of aqueous solution: Biological and chemical applications. *J. Phys. Chem.* **97**: 1101–1109.
- Hsieh, M.-J. and Luo, R. 2004. Physical scoring function based on amber force field and Poisson-Boltzmann implicit solvent for protein structure prediction. *Proteins* **56**: 475–486.
- Kollman, P.A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Duan, Y., Wang, W., Donini, O., et al. 2000. Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Acc. Chem. Res.* **33**: 889–897.
- Lazaridis, T. and Karplus, M. 2000. Effective energy functions for protein structure prediction. *Curr. Op. Struct. Biol.* **10**: 139–145.
- Lee, M.R. and Kollman, P.A. 2001. Free-energy calculations highlight differences in accuracy between x-ray and NMR structures and add value to protein structure prediction. *Structure* **9**: 905–916.
- Lee, M.R., Baker, D., and Kollman, P.A. 2001. 2.1 and 1.8 Å average c_α RMSD structure predictions on two small proteins, hp-36 and s15. *J. Am. Chem. Soc.* **123**: 1040–1046.
- Levy, R.M., Zhang, L.Y., Gallicchio, E., and Felts, A.K. 2003. On the nonpolar hydration free energy of proteins: Surface area and continuum solvent models for the solute-solvent interaction energy. *J. Am. Chem. Soc.* **125**: 9523–9530.
- Lu, B.Z., Chen, W.Z., Wang, C.X., and Xu, X.-J. 2002. Protein molecular dynamics with electrostatic force entirely determined by a single Poisson-Boltzmann calculation. *Proteins* **48**: 497–504.
- MacKerell Jr., A.D., Bashford, D., Bellott, M., Dunbrack Jr., R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., et al. 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**: 3586–3616.
- Madura, J.D., Davis, M.E., Gilson, M.K., Wade, R., Luty, B.A., and McCammon, J.A. 1994. Biological applications of electrostatics calculations and Brownian dynamics simulations. *Rev. Comp. Chem.* **5**: 229–267.
- Madura, J.D., Briggs, J.M., Wade, R., Davis, M.E., Luty, B.A., Ilin, A., Antosiewicz, J.A., Gilson, M.K., Bagheri, B., Ridgway Scott, L., et al. 1995. Electrostatics and diffusion of molecules in solution: Simulations with the University of Houston Brownian Dynamics program. *Comput. Commun. Phys.* **91**: 57–95.
- Marcus, R.A. 1955. Calculation of thermodynamic properties of polyelectrolytes. *J. Chem. Phys.* **23**: 1057–1068.
- Nicholls, A., Sharp, K.A., and Honig, B. 1991. Protein folding and association: Insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11**: 281–296.
- Novotny, J., Brucoleri, R., and Karplus, M. 1984. An analysis of incorrectly folded protein models. Implications for structure predictions. *J. Mol. Biol.* **177**: 787–818.
- 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.
- Petrey, D. and Honig, B. 2000. Free energy determinants of tertiary structure and the evaluation of protein models. *Protein Sci.* **9**: 2181–2191.
- Qiu, D., Shenkin, P., Hollinger, F., and Still, W. 1997. The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate born radii. *J. Phys. Chem.* **101**: 3005–3014.
- Roux, B. and Simonson, T. 1999. Implicit solvent models. *Biophys. Chem.* **78**: 1–20.
- Samudrala, R. and Levitt, M. 2000. Decoys 'r' us: A database of incorrect protein conformations to improve protein structure prediction. *Protein Sci.* **9**: 1399–1401.
- Samudrala, R. and Moulton, J. 1998. A graph-theoretic algorithm for comparative modelling of protein structure. *J. Mol. Biol.* **279**: 287–302.
- Schapira, M., Totrov, M., and Abagyan, R. 1999. Prediction of the binding energy for small molecules, peptides, and proteins. *J. Mol. Recogn.* **12**: 177–190.
- Schutz, C.N. and Warshel, A. 2001. What are the dielectric “constants” of proteins and how to validate electrostatic models? *Proteins* **44**: 400–417.
- Sharp, K.A., and Honig, B. 1990. Calculating total electrostatic energies with the non-linear Poisson-Boltzmann equation. *J. Phys. Chem.* **94**: 7684–7692.
- Simonson, T. 1999. Dielectric relaxation in proteins: Microscopic and macroscopic models. *Int. J. Quant. Chem.* **73**: 45–57.
- Srinivasan, J., Cheatham III, 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.
- Tosatto, S.C.E., Bindewald, E., Hesser, J., and Maenner, R. 2002. A divide and conquer approach to fast loop modeling. *Protein Eng.* **15**: 279–286.
- Vajda, S., Sippl, M., and Novotny, J. 1997. Empirical potentials and functions for protein folding and binding. *Curr. Op. Struct. Biol.* **7**: 222–228.
- Vorobjev, Y.N. and Hermans, J. 2001. Free energies of protein decoys provide insight into determinants of protein stability. *Protein Sci.* **10**: 2498–2506.
- Vorobjev, Y.N., Almagro, J.C., and 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* **32**: 399–413.
- Vriend, G. 1990. What if: A molecular modeling and drug design program. *J. Mol. Graph.* **8**: 52–54.
- Wang, Y., Zhang, H., Li, W., and Scott, R.A. 1995. Discriminating compact nonnative structures from the native structure of globular proteins. *Proc. Natl. Acad. Sci.* **92**: 709–713.
- Xiang, Z., Soto, S.C., and Honig, B. 2002. Evaluating free energies: The colony energy and its application to the problem of loop prediction. *Proc. Natl. Acad. Sci.* **99**: 7432–7437.
- Zhang, C., Liu, S., Zhou, H., and Zhou, Y. 2004a. Accurate and efficient loop selections by the dfire-based all-atom statistical potential. *Protein Sci.* **13**: 391–399.
- . 2004b. An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. *Protein Sci.* **13**: 400–411.
- Zhou, H.X. 1994. Macromolecular electrostatic energy within the nonlinear Poisson-Boltzmann equation. *J. Chem. Phys.* **100**: 3152–3162.
- Zhou, H. and Zhou, Y. 2002. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.* **11**: 2714–2726.