

Discrimination of the Native from Misfolded Protein Models with an Energy Function Including Implicit Solvation

Themis Lazaridis¹ and Martin Karplus^{1,2*}

¹Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford St Cambridge, MA 02138, USA

²Laboratoire de Chimie Biophysique, ISIS, Institute le Bel, Université Louis Pasteur 67000 Strasbourg, France

An essential requirement for theoretical protein structure prediction is an energy function that can discriminate the native from non-native protein conformations. To date most of the energy functions used for this purpose have been extracted from a statistical analysis of the protein structure database, without explicit reference to the physical interactions responsible for protein stability. The use of the statistical functions has been supported by the widespread belief that they are superior for such discrimination to physics-based energy functions. An effective energy function which combined the CHARMM vacuum potential with a Gaussian model for the solvation free energy is tested for its ability to discriminate the native structure of a protein from misfolded conformations; the results are compared with those obtained with the vacuum CHARMM potential. The test is performed on several sets of misfolded structures prepared by others, including sets of about 650 good decoys for six proteins, as well as on misfolded structures of chymotrypsin inhibitor 2. The vacuum CHARMM potential is successful in most cases when energy minimized conformations are considered, but fails when applied to structures relaxed by molecular dynamics. With the effective energy function the native state is always more stable than grossly misfolded conformations both in energy minimized and molecular dynamics-relaxed structures. The present results suggest that molecular mechanics (physics-based) energy functions, complemented by a simple model for the solvation free energy, should be tested for use in the inverse folding problem, and supports their use in studies of the effective energy surface of proteins in solution. Moreover, the study suggests that the belief in the superiority of statistical functions for these purposes may be ill founded.

© 1999 Academic Press

Keywords: protein folding; implicit solvation; molecular dynamics; threading

*Corresponding author

Introduction

Due to the increasing number of protein sequences that are being determined and the need to be able to predict their structures, there

is an intense effort at present to devise potential energy functions that can distinguish native from misfolded proteins. Use of such functions is based on the thermodynamic hypothesis which states that a native protein has the conformation or set of very similar conformations that minimize the free energy of the system. The equilibrium probability distribution of protein conformations, $p(q)$, is given by (e.g. Gibson & Scheraga, 1969; Karplus & Shakhnovich, 1992; Lazaridis & Karplus, 1999):

$$p(q) = \frac{\exp(-W/kT)}{\int \exp(-W/kT)dq} \quad (1)$$

Present address: T. Lazaridis, Department of Chemistry, City College of New York, Convent Ave & 138th St. New York, NY 10031, USA.

Abbreviations used: MD, molecular dynamics; NDK, nucleotide diphosphate kinase; RMSD, root-mean-square deviation; EEF1, effective energy function 1.

E-mail address of the corresponding author: marci@brel.u-strasbg.fr

where the effective energy $W(q)$ of conformation q can be written:

$$W(q) = E^{\text{intra}}(q) + \Delta G^{\text{sol}}(q) \quad (2)$$

Here $E^{\text{intra}}(q)$ is the intraprotein energy and $\Delta G^{\text{sol}}(q)$ is the solvation free energy. A stable protein under physiological conditions populates a narrow range of conformations (within about 2 Å RMS), that together make up the native state (Brooks *et al.*, 1988). The effective energy surface in the neighborhood of the native state is a complex multiminimum surface, as has been demonstrated by experiments (Frauenfelder *et al.*, 1991) and simulations (Caves *et al.*, 1998; Elber & Karplus, 1987; Levitt, 1983; Noguti & Go, 1989). Since the entropic contribution to the free energy of the native state can be approximated by that of a single or a few conformations (Karplus *et al.*, 1987) relative to the multitude of conformations in the unfolded state, the native state minimum must be very deep, i.e. there must be a large gap between the energy of the native state and the average non-native state whose structure differs significantly from it (Bryngelson & Wolynes, 1987; Karplus & Shakhnovich, 1992; Moult, 1997; Šali *et al.*, 1994). Any function that is useful for representing the effective energy of proteins must satisfy the condition that the native state is lowest in energy and that there is a sizeable energy gap.

Much of the recent effort in the search for such effective energy functions has focused on the analysis of known protein structures supplemented by assumptions concerning the form of the potential function (Bowie *et al.*, 1991; Casari & Sippl, 1992; DeBolt & Skolnick, 1996; Koehl & Delarue, 1994; Reva *et al.*, 1997). In many cases, such as the work described by Bowie *et al.* (1991), the term "effective energy function" is interpreted rather broadly since it depends only on the "environment" of each atom. Although there has been considerable success in using such empirical potentials in fold recognition (Marchler-Bauer *et al.*, 1997), particularly if no additions or deletions are allowed, they do not account for distortion of the covalent structure or steric crowding. In a recent comparison (Marti-Renom *et al.*, 1998), for example, several empirical potentials failed to recognize a badly misfolded structure of potato carboxypeptidase inhibitor. Levitt and co-workers (Park *et al.*, 1997; Park & Levitt, 1996) have tested a number of empirical functions and found none of them to be completely satisfactory in distinguishing incorrect conformations from the native structure. Others focused on discrimination of structures in the vicinity of the native structure, rather than grossly misfolded structures (Wang *et al.*, 1995; Williams *et al.*, 1992).

One reason for the recent emphasis on statistically based criteria is the widespread belief that "molecular mechanics" potential energy functions of the type used in simulations (e.g. Brooks *et al.*,

1983) cannot distinguish between native and misfolded structures. A study by Novotny *et al.* (1984), which was a precursor to modern threading studies (Finkelstein, 1997), is most frequently cited to support this thesis. In that study two proteins with the same number of residues but different folds were considered and the sequence of one was "threaded" onto the fold of the other. This created two pairs of correct and incorrect folds. After side-chains were built onto the incorrect models, the energy calculated with the CHARMM potential was minimized to remove bad contacts and then compared to the energy of the correct structure. The conclusion of the study, which seems not to have been fully understood, was that the calculated energy after mild minimization to eliminate bad contacts (RMS shift of 0.5 to 0.9 Å) was a "reasonable energy" for a native protein of the size of the test system, i.e. from the absolute energy by itself it was not obvious that the misfolded structure was wrong. However, if comparisons of the energies of the native and misfolded structure were made, the calculations did, in fact, discriminate the native fold, in spite of many statements to the contrary (e.g. Eisenberg & McLachlan, 1986).

Since there seems to be little, or actually no evidence that "physics-based" energy functions (Moult, 1997) are "worse" than statistically based functions, it seems worthwhile to make a more extended test of a function of the former type. As is evident from equation (2), it is the effective energy which includes solvation, that should be used both in minimization of the structure and in calculation of its energy. One specific problem that can arise in vacuum calculations is that conformations with the polar groups inside and the non-polar groups on the surface ("reverse proteins") are more stable because interactions between polar groups are strong and there is no desolvation free energy cost for their burial in the protein interior (Novotny *et al.*, 1984). This could happen in non-polar environments, such as membranes, but it is opposite of what is observed in protein structures that are stable in aqueous solution; i.e. the polar groups tend to be on the surface and it is the non-polar groups that are buried.

Explicit modeling of the solvent, although very useful in molecular dynamics (MD) simulations (Brooks *et al.*, 1988), is not suitable for protein structure prediction because of the large amount of computer time required to survey a range of conformations. For this purpose, one needs a function for the solvation free energy, ideally analytical with analytical derivatives, that can be calculated rapidly. Wesson & Eisenberg (1992), for example, combined empirical atomic solvation parameters with the polar hydrogen CHARMM 19 potential energy function (Brooks *et al.*, 1983; Neria *et al.*, 1996) and performed dynamics simulations on melittin. Stouten *et al.* (1993) presented a model based on contacts rather than accessible surface areas, combined it with the GROMOS energy function (van Gunsteren & Berendsen, 1987), and per-

formed simulations on BPTI. They found the model to be a significant improvement over vacuum simulations, although the observed RMSD from the crystal structure was somewhat larger than in explicit water simulations. Fraternali & van Gunsteren (1996) developed an empirical solvation potential adjusted so as to reproduce the experimental radius of gyration of proteins in MD simulations. Friesner and co-workers (Humphreys *et al.*, 1995; Monge *et al.*, 1995) used the AMBER force-field with the Generalized Born model (Still *et al.*, 1990) to evaluate protein structures and found that the resulting function sometimes showed non-native structures to have a lower effective energy than the native structure. Augspurger & Scheraga (1996) recently simplified the hydration shell model (Kang *et al.*, 1987) by including only double overlap terms in the calculation of the hydration shell volumes, which leads to enhanced computational efficiency. They compared the solvation free energy calculated by this model to that obtained by the Poisson-Boltzman equation (Augspurger & Scheraga, 1997), but did not report extensive tests of the combined energy function. This is true for most physics-based effective energy functions presented to date. They have not been tested for their ability to discriminate native from non-native structures in the spirit of Novotny *et al.* (1984) or for their ability to give stable native states in room temperature molecular dynamics simulations. Recently, Vorobjec *et al.* (1998) performed a limited Novotny-type test on nine out of the 22 proteins of the EMBL set (see below) with a hybrid approach. They generated an ensemble of conformations starting with both the native and the misfolded conformation by molecular dynamics simulations in explicit solvent, and evaluated these conformations using a molecular mechanics energy function complemented by three solvation terms: one for the cavity formation free energy, one for the protein-solvent dispersion interactions, and one for electrostatic polarization, evaluated by continuum electrostatics methods. They found that the average effective energy was always lower for the native structure. A stricter test of the proposed solvation model would be to use it in the generation of the ensemble of conformations, as well as in their evaluation.

We have recently developed an effective energy function (EEF1; Lazaridis & Karplus, 1997, 1999) based on the polar hydrogen form of the CHARMM potential energy function (Brooks *et al.*, 1983; Neria *et al.*, 1996) complemented by a theoretically based solvation free energy model. EEF1 has been shown to give stable native structures for a series of proteins when used for MD simulations at room temperature, reasonable energies for unfolded conformations, and unfolding pathways in agreement with explicit water simulations (Lazaridis & Karplus, 1997). This suggests that EEF1 may be sufficiently accurate to be used for distinguishing native from non-native states for a given sequence. Here, we report the results of

“Novotny” pairwise threading tests for a series of proteins with this effective energy function; for comparison we do the same test for the vacuum CHARMM 19 potential energy function. The threading tests are performed on the set of native misfolded pairs created by Holm & Sander (1992), a subset of which was used by Vorobjec *et al.* (1998), as mentioned above. Since these tests work well, we extend the tests to the small protein CI2 threaded into the fold of eight proteins of similar size. Finally, we apply EEF1 to a large number of decoys for six proteins prepared by Park & Levitt (1996), and to a set of CASP1 homology models.

Results

Table 1 shows the results obtained for the Holm-Sander misfolded structures with the CHARMM 19 vacuum potential. The total energy, the van der Waals and electrostatic components, as well as the RMSD after dynamics, are reported. With respect to the energy minimized structures, we see that CHARMM discriminates the correct conformation in most cases; there are three exceptions: 1ppt (avian pancreatic polypeptide), 1sn3 (scorpion neurotoxin), and 3b5c (cytochrome *b5*). After MD simulations are performed, the CHARMM 19 energy of the correct structure is higher than that of the incorrect model in most cases. The largest contribution to the difference usually derives from the electrostatic energy. The poor performance of the vacuum force-field is due primarily to the ionic side-chains. The side-chains with opposite charges in vacuum can have large electrostatic stabilization energies and these become more negative when they approach each other during the MD simulation. Neutralization of the ionic side-chains and use of a distance-dependent dielectric improves the performance substantially (data not shown).

The results obtained with EEF1 are shown in Table 2. The total energy, the van der Waals, the electrostatic, and solvation contributions are listed, as well as the RMSD after the MD simulation. The electrostatic term is calculated with neutralized side-chains and a distance-dependent dielectric constant (see Methods). The total energy, both after minimization and after molecular dynamics is always lower for the native conformation. However, the energy gap is usually significantly larger for the energy-minimized structures, since dynamics allows more effective relaxation of the misfolded structures. The exception of 1fdx *versus* 1fdxon5rxn for the energy-minimized structures (the misfolded structure is 1 kcal/mol lower in energy) is not a failure of the model, since ferredoxin (1fdx) has two iron-sulfur clusters which are an integral part of its structure and were omitted in the calculations.

Tables 1 and 2 also show the RMS deviation from the starting structure after the dynamics simulation. For the CHARMM 19 vacuum potential the deviation is often larger for the correct than

Table 1. Energies of native-misfolded pairs with CHARMM 19

	After 300 ABNR steps	After 50 ps MD + 300 ABNR	RMSD
	Total <i>E</i> (–vdW, –elec)	Total <i>E</i> (–vdW, –elec)	
1bp2	–7013 (799, 6595)	–7858 (785, 7529)	4.33
1bp2on2paz	–6837 (730, 6562)	–8160 (857m 7783)	4.18
1cbh	–1595 (186, 1498)	–1680 (196, 1583)	1.84
1cbhon1ppt	–1489 (154, 1422)	–1718 (175, 1625)	2.62
ifdx	–2311 (228, 2286)		
1fdxon5rxn	–2167 (221, 2129)		
1hip	–4515 (512, 4282)		
1hipon2b5c	–4260 (433, 4149)		
1lh1	–8413 (1022, 7840)	–9617 (1071, 9086)	4.15
1lh1on1ilb	–8291 (956, 7928)	–9602 (1057, 9104)	3.26
1p2p	–6890 (801, 6522)	–7834 (842, 7465)	3.64
1p2on1rn3	–6629 (745, 6333)	–7848 (859, 7487)	3.79
1ppt	–1600 (212, 1503)	–2220 (252, 2126)	6.68
1ppton1cbh	–1801 (185, 1803)	–2211 (260, 2126)	3.32
1rei	–10,319 (1509, 9583)	–11,422 (1664, 10,491)	2.37
1reion5pad	–10,051 (1327, 9478)	–11,809 (1543, 11,012)	4.91
1rhd	–15,925 (2133, 14,823)	–18,067 (2343, 16,949)	3.41
1rhdon2cyp	–14,886 (1935, 14,175)	–18,705 (2286, 17,784)	4.43
3rn3	–6550 (820, 6097)	–7613 (848, 7180)	3.46
1rn3on1p2p	–6377 (718, 6074)	–8490 (815, 7149)	3.08
1sn3	–3673 (389, 3491)	–4144 (398, 3994)	1.67
1sn3on2ci2	–3337 (328, 3227)	–4484 (401, 4323)	3.65
1sn3on2cro	–3691 (343, 3573)	–4406 (389, 4292)	2.79
3b5c	–4784 (568, 4505)	–5880 (613, 5593)	4.7
2b5con1hip	–4972 (478, 4873)	–5974 (605, 5732)	3.92
2cdv	–6248 (549, 6018)	–7546 (706, 7264)	4.26
2cdvon2ssi	–5796 (556, 5632)	–7717 (622, 7584)	3.84
2ci2	–4101 (431, 3901)	–4527 (440, 4346)	2.39
2ci20n1sn3	–3664 (372, 3565)	–4611 (407, 4507)	3.74
2ci2on2cro	–3625 (376, 3514)	–4624 (412, 4507)	3.84
2cro	–3479 (423, 3266)	–4057 (423, 3860)	3.55
2croon1sp3	–3099 (374, 2973)	–4101 (405, 4001)	5.1
2croon2ci2	–3335 (375, 3206)	–4032 (431, 3870)	3.87
2cyp	–16,939 (2281, 15,658)	–19,458 (2393, 18,252)	3.63
2cypion1rhd	–16,313 (1965, 15,678)	–20,233 (2282, 19,313)	4.41
2ilb	–9035 (1065, 8468)	–10,077 (1136, 9525)	3.24
2ilbon1lh1	–8347 (932, 7966)	–10,612 (1120, 10,097)	5.62
2paz	–6928 (776, 6555)	–7684 (835, 7312)	2.71
2pazon1bp2	–6612 (665, 6417)	–7843 (836, 7510)	5.24
2ssi	–4867 (609, 4576)	–5502 (677, 5193)	3.32
2ssion3cdv	–4488 (488, 4348)	–5475 (647, 5163)	5.65
2tmn	–16,645 (2419, 15,101)		
2tmnon2ts1	–15,278 (1950, 14,304)		
2ts1	–17,957 (2377, 16,569)	–21,101 (2548, 19,818)	4.28
2ts1on2tmn	–17,362 (2149, 16,582)	–22,038 (2428, 20,997)	5.04
5pad	–10,946 (1551, 10,047)	–11,922 (1569, 11,105)	2.83
5padon1rei	–9799 (1340, 9319)	–12,136 (1600, 11,397)	4.37

The first entry in each group is the native structure (four letter PDB code) and the other entries are the misfolded structures (AonB means the sequence of A threaded onto the structure of B). Total energies (kcal/mol) are given after minimization and after dynamics plus minimization. In parentheses are the van der Waals and electrostatic components. The last column is the RMSD (in Å) from the initial structure after the MD simulation.

for the incorrect structure. This never happens for EEF1, except for 2cdv (cytochrome c3), which has four heme molecules that were omitted in the simulation. Moreover, the RMSD with EEF1 after dynamics for native proteins is usually significantly smaller than that obtained with CHARMM 19 and the magnitude of the deviations are small (in the range 1.5 to 3 Å). The largest deviations are observed for 2cdv (see above), 1rei (Bence-Jones immunoglobulin variable portion), which is a V-shaped dimer fragment of a full immunoglobulin, and 2ts1 (tyrosyl t-RNA synthetase).

The results of the threading test for CI2 are given in Table 3. For all comparisons (minimized energy, dynamics average, dynamics plus minimiz-

ation) the native state is significantly lower in energy than any of the misfolded (threaded) structures. As before, the MD simulations result in a significant reduction in the energy difference between the misfolded structures and the native structure. The smallest difference is observed for ubiquitin (about 35 kcal/mol). This suggests that molecular dynamics makes threading tests much more stringent and should be used whenever possible. Unfortunately, many of the potentials used in threading cannot be used for dynamics simulations because they are discontinuous or employ a representation of a protein that is incomplete (e.g. they do not include a van der Waals term).

Table 2. Energies of native-misfolded pairs with EEF1

	After 300 ABNR steps	After 50 ps MD + 300 ABNR	RMSD
	Total W (–vdW, –elec, –solv)	Total W (–vdW, –elec, –solv)	
1bp2	–3746 (881, 1988, 1163)	–3913 (812, 2223, 1176)	2.21
1bp2on2paz	–3566 (810, 1881, 1193)	–3896 (806, 2197, 1185)	4.31
1cbh	–977 (194, 563, 295)	–1010 (193, 605, 291)	1.99
1cbhon1ppt	–937 (160, 554, 303)	–992 (152, 602, 313)	4.40
1fdx	–1323 (252, 770, 460)		
1fdxon5rxn	–1324 (249, 956, 460)		
1hip	–2400 (543, 1319, 758)		
1hipon2b5c	–2306 (480, 1301, 760)		
1lh1	–4365 (1123, 2301, 1268)	–4517 (1091, 2507, 1259)	2.12
1lh1on1ilb	–4152 (1020, 2212, 1324)	–4442 (1073, 2462, 1292)	3.19
1p2p	–3742 (887, 1994, 1187)	–3938 (857, 2197, 1180)	1.63
1p2on1rn3	–3632 (812, 1879, 1242)	–3912 (789, 2182, 1232)	4.73
1ppt	–1075 (222, 557, 395)	–1141 (209, 634, 390)	2.95
1ppton1cbh	–997 (224, 529, 382)	–1135 (217, 646, 377)	3.71
1rei	–5839 (1590, 3238, 1728)	–6386 (1544, 3668, 1736)	4.24
1reion5pad	–5674 (1380, 3189, 1718)	–6220 (1362, 3736, 1690)	4.67
1rhd	–8524 (2274, 4451, 2610)	–8993 (2258, 5043, 2513)	2.85
1rhdon2cyp	–8046 (2122, 4301, 2527)	–8701 (2080, 4914, 2533)	5.26
3rn3	–3824 (856, 2064, 1192)	–3990 (827, 2294, 1171)	2.31
1rn3on1p2p	–3596 (762, 1950, 1195)	–3914 (744, 2251, 1183)	3.98
1sn3	–1888 (417, 963, 659)	–1945 (428, 1039, 640)	1.44
1sn3on2ci2	–1823 (382, 930, 672)	–1939 (406, 1067, 628)	2.82
1sn3on2cro	–1775 (394, 909, 642)	–1924 (395, 1062, 626)	2.62
3b5c	–2707 (640, 1322, 928)	–2803 (643, 1444, 906)	2.35
2b5con1hip	–2543 (576, 1258, 923)	–2716 (550, 1456, 932)	3.54
2cdv	–3184 (595, 1595, 1213)	–3368 (622, 1849, 1138)	4.61
2cdvon2ssi	–3081 (638, 1570, 1151)	–3348 (642, 1836, 1119)	3.27
2ci2	–1931 (473, 987, 642)	–1994 (449, 1070, 643)	1.79
2ci2on1sn3	–1761 (441, 908, 628)	–1935 (460, 1064, 606)	4.01
2ci2on3cro	–1801 (446, 917, 617)	–1910 (401, 1051, 638)	2.74
2cro	–1965 (478, 1028, 603)	–2021 (441, 1129, 615)	2.66
2croon1sn3	–1854 (423, 973, 638)	–1994 (421, 1134, 613)	5.88
2croon2ci2	–1857 (411, 963, 648)	–2011 (420, 1126, 623)	2.90
2cyp	–9055 (2480, 4654, 2660)	–9399 (2342, 5118, 2688)	2.27
2cyp on1rhd	–8370 (2206, 4381, 2704)	–9155 (2210, 5103, 2670)	5.10
2ilb	–4572 (1133, 2350, 1468)	–4838 (1115, 2688, 1425)	2.77
2ilbon1lh1	–4439 (1052, 2384, 1372)	–4812 (1063, 2763, 1347)	3.58
2paz	–3487 (855, 1885, 1058)	–3604 (846, 2021, 1042)	1.89
2pazon1bp2	–3247 (765, 1754, 1055)	–3466 (739, 1973, 1061)	4.18
2ssi	–2743 (634, 1497, 880)	–2932 (636, 1679, 871)	2.46
2ssion2cdv	–2613 (527, 1444, 912)	–2852 (559, 1671, 876)	8.18
2tmn	–9498 (2589, 5080, 2521)		
2tmnon2ts1	–8923 (2116, 4834, 2704)		
2ts1	–9838 (2537, 5073, 2959)	–10,215 (2451, 5525, 2964)	3.29
2ts1on2tmn	–9071 (2441, 4777, 2801)	–9932 (2243, 5495, 2997)	6.39
5pad	–6150 (1625, 3304, 1765)	–6412 (1560, 3644, 1764)	2.58
5padon1rei	–5629 (1414, 3083, 1839)	–6327 (1435, 3653, 1804)	15.79

See legend to Table 1 for the definitions and general description.

Table 3. Threading of the CI2 sequence

Target model	PDB	Nres	Min ^a	Dynav ^b	Dynmin ^c	RMSD ^d	Rg ^e
CI2	2ci2	64	–1783	–1372	–1923 (445, 1053, 612)	1.80	11.23
CI2 (correct Trp24)			–1926	–1392	–1944 (437, 1073, 612)	1.36	11.08
Eglin c (not aligned)	leg1	70	–1682	–1330	–1880 (361, 1038, 651)	5.62	12.25
Eglin c (aligned)			–1749	–1357	–1914 (425, 1058, 610)	3.27	11.44
Tendamistat	1hoe	74	–1699	–1337	–1888 (357, 1057, 649)	6.86	12.7
Ubiquitin	1ubq	76	–1725	–1376	–1930 (445, 1093, 579)	2.18	10.97
G B2	1pgx	66	–1775	–1362	–1906 (407, 1070, 602)	4.74	12.6
Spectrin SH3	1shg	57	–1746	–1362	–1911 (404, 1078, 613)	3.25	12.8
Neurotoxin	2sn3	65	–1701	–1302	–1853 (384, 998, 646)	6.07	12.05
Cro repressor	2cro	65	–1728	–1346	–1903 (437, 1058, 593)	2.86	10.9
CTF of L7/L12	1ctf	68	–1775	–1341	–1899 (416, 1046, 607)	2.87	11.2
Native CI2			–1905	–1413	–1963 (549, 1064, 607)	1.05	11.32

^a Energy after minimization (kcal/mol).

^b Average energy over last 2 ps of the MD simulation at 300 K.

^c Quenched energy at the end of MD. In parenthesis minus the van der Waals, electrostatic, and solvation contributions.

^d Backbone RMSD (in Å) from starting structure (from crystal structure in the case of CI2 itself).

^e Final radius of gyration (in Å) after MD.

For CI2 built with extended side-chains, the energy is significantly higher than native CI2. The short MD simulations do not overcome the barriers associated with side-chain rearrangement, so that after dynamics many side-chains remain incorrect. One example of a misoriented side-chain is Trp24, which forms part of the hydrophobic core. In the native structure its six-membered ring points into the core, whereas in the reconstructed structure the six-membered ring points outwards. This disrupts the packing in the hydrophobic core. The second entry in Table 3 describes what happens if we place only the Trp side-chain in its correct orientation "by hand", minimize, and then run an MD simulation. Energy minimization does not make much difference (-1926 versus -1923 kcal/mol), but upon MD the energy drops by about 20 kcal/mol, half of the energy difference relative to the native state. Also, the RMSD from the native decreases from 1.8 to 1.36 Å. The energy of the reconstructed CI2 structure (that with the incorrect orientation of Trp24) is not much different from the energies of the misfolded structures. This implies that the energy difference alone is not sufficient to determine whether the structure is "close" or "far" from the native state. Apparently, the energy function can pick out a native structure from those that are non-native, but does not provide a simple way of determining "native-like character", e.g. there is no correlation between RMSD and the energy difference (see also Marti-Renom *et al.*, 1998). This behavior also gives some insight into the cooperativity of protein folding. Once certain side-chains are displaced from the correct orientation, the energy is nearly equivalent to many other unfolded or misfolded conformations. This makes clear the difficulty of protein structure predictions, since it shows how easy it is to "miss" the native state unless all the essential elements are correct.

In the case of eglin c, an incorrect alignment with CI2 produces large deviations from the initial structure; the deviation is smaller when the alignment is correct (see Table 3). This suggests that the present approach could be used in detecting incorrect alignments in "inverse folding" studies. The ubiquitin model is also of interest. Its energy is the lowest of the misfolded structures (much of which is due to very favorable electrostatic interactions, even with the solvation model) and exhibits small RMSD shifts after the MD simulation. Examination of this model reveals that it has an "ionic" core instead of a hydrophobic core, i.e. several ionic groups form "salt bridges" in the interior. This also explains why the solvation free energy of this model is so small in magnitude; there are only very few external pseudo-ionic side-chains that contribute to the solvation free energy.

The results for the decoys by Park & Levitt (1996) are shown in Figure 1. The test in this case involves only energy-minimized structures due to the large number of structures considered. These energy minimizations produce very small RMS

deviations from the starting structure, typically of the order of 0.5 Å. Therefore, the results are plotted in terms of the original RMSD. In all cases, the native structure (shown at zero RMSD) has lower effective energy than any of the decoys. The gap between the minimized crystal structure and the lowest in energy decoy varies between 5 kcal/mol for 3icb (Figure 1(f)) and 92 kcal/mol for 1sn3 (Figure 1(c)). The effective energy shows a tendency to increase as a function of RMSD. This is regarded as a desirable property, but there is no physical requirement for such a correlation. Significant scatter exists in the plots and it tends to increase as the RMSD increases (in these plots a few points with very high energies were omitted to improve the resolution for the bulk of the points). Examination of the energy components shows that the native state is favored by all internal terms (non-bonded as well as bonded), whereas the solvation term may favor the correct or the incorrect conformation. The electrostatic and van der Waals nonbonded terms usually are most important. For comparison, Figure 2 shows the same test for 2cro performed with the vacuum CHARMM 19 energy function, where there are several decoys with energies lower than the crystal structure. As in the results for the Holm-Sander set, this function discriminates against most but not all of the decoys.

The results for the CASP1 homology models are shown in Table 4. The minimized crystal structure has lower effective energy than the models in all cases, except for the first two models for nucleotide diphosphate kinase (NDK). These models are 0.5 and 1.4 Å backbone RMSD from the crystal structure. The first model is favored by both the van der Waals and the electrostatic term, and the second one is highly favored by the van der Waals term. Interestingly, a recently developed statistical potential also failed to discriminate the crystal structure of this protein (Samudrala & Moult, 1998). That EFF1 does not always discriminate with respect to structures in the vicinity of the crystal structure is not surprising considering the fact that the lowest energy structures found during molecular dynamics simulations of native proteins with this function are always 1-2 Å from the crystal structure. Also, different crystal and NMR structures of the same protein often have RMS deviations of the same order of magnitude (Smith *et al.*, 1994). Since proteins at ambient temperatures sample a range of conformations around the crystal structure with an RMS of 1 Å or more, it is not obvious that even nature's "function" would pass the test described here.

Discussion

Compared to statistical database potentials, physical effective energy functions have many advantages. First, they have a sound theoretical basis, whereas the theoretical basis of database

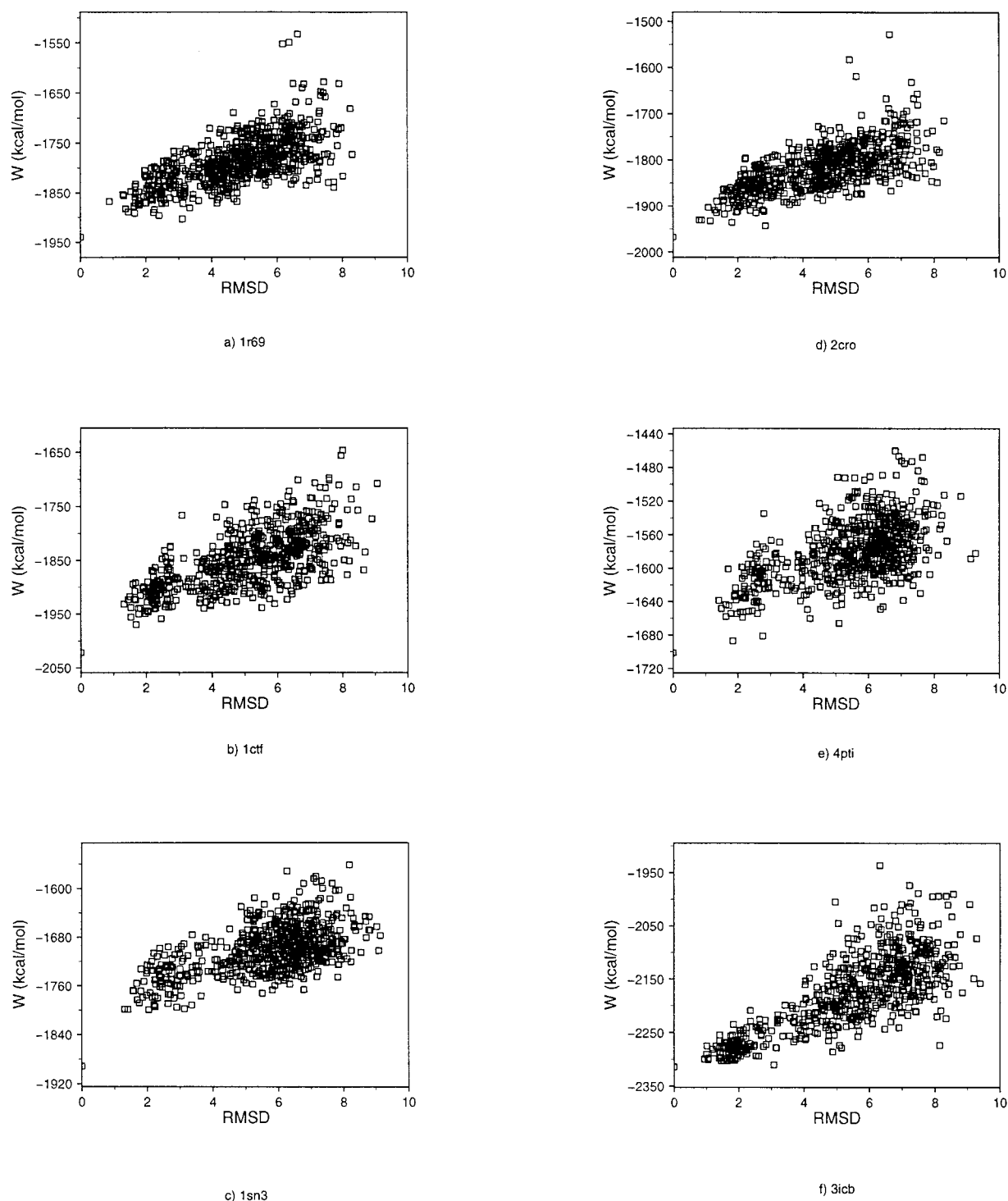


Figure 1. Effective energy of the Park & Levitt decoys compared to that of the crystal structure.

potentials is still being debated. Secondly, because they are "complete" energy functions with analytic first derivatives, they can be used in energy minimization and dynamics for studying the energetics unfolded, misfolded, and partially folded states, in addition to the native state. This can be very useful; for example, with MD one can test whether the structure examined is a stable (local) energy minimum. In the tests reported here we often observed that an incorrect structure unfolded during an MD

simulation. The disadvantage of molecular mechanics energy functions is their higher computational cost; database potentials are more efficient for filtering a large number of putative folds. However, this is no longer significant for the atom-based (rather than residue-based) statistical functions that have been introduced recently (Samudrala & Moulton, 1988). Although, in principle, statistical database potentials could be developed for MD, in practice no such function exists at present.

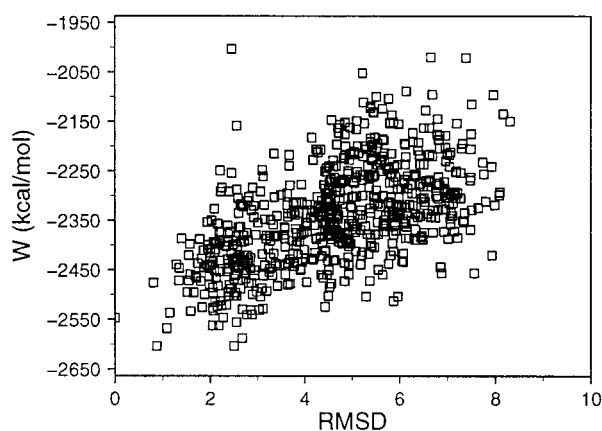


Figure 2. Vacuum energy of the Park & Levitt decoys for 2cro compared to that of the crystal structure.

Compared to other molecular mechanics energy functions with implicit solvation EEF1 is probably the most extensively tested so far. It is an all-purpose function, not limited to a particular application. One significant advantage of the present solvation model over a number of others is that it does not involve the calculation of accessible surface areas. This allows it to be only 50% more expensive than the corresponding vacuum simulations. Accessible surface-based methods add significant computational expense without being any more rigorous or accurate than the contact-based method used in EEF1.

The "Novotny" test is a necessary but not sufficient condition for the validity of an effective energy function. Although the number of misfolded conformations considered here was large, the results do not "prove" the validity of the energy function. In other work we examined thousands of conformations generated by high temperature unfolding molecular dynamics simulations and subsequent low temperature quenching and obtained a funnel-like effective energy surface for the protein C12 (Lazaridis & Karplus, 1997). This is consistent with the idea that EEF1 has a global minimum within 1 or 2 Å of the

experimental native structure. Of course more exhaustive searches of conformational space are required to prove this point; such studies, as well as a more quantitative characterization of solvation effects, will be made in the future.

In conclusion, the series of tests reported here demonstrate that EEF1 can discriminate the native structure from a series of misfolded structures, and suggests that such physics-based energy functions can be useful in protein structure evaluation and prediction. Of particular importance is the result that such applications can be made taking into account relaxation of the misfolded structures by molecular dynamics. Moreover, comparison with published data of corresponding tests with statistical energy functions suggest that EEF1 is at least as good, and perhaps better. We hope that the present study has laid to rest the widely held belief that, at the present stage of development, statistical energy functions are superior to physics-based energy functions for evaluating protein structures. In the future, both types of functions should be used and it may well turn out that one or the other is better for certain applications, e.g., residue-based statistical functions may be better for evaluating low-resolution models.

Methods

The details of the effective energy function are given elsewhere (Lazaridis & Karplus, 1997, 1999). The solvent contribution to the effective energy (potential of mean force) uses the assumption that the solvation free energy of a macromolecule ΔG^{solv} is the sum of contributions from its constituent groups:

$$\Delta G^{\text{solv}} = \sum_i \Delta G_i^{\text{solv}}; \quad \Delta G_i^{\text{solv}} = \Delta G_i^{\text{ref}} - \sum_j f_i(r_{ij})V_j$$

where ΔG_i^{ref} is the solvation free energy of group i in a reference compound, V_j is the volume of group j and $f_i(r_{ij})$ is the solvation free energy density of group i at distance r_{ij} . Values of ΔG_i^{ref} are obtained from the work reported by Privalov & Makhatadze (1993), with slight modifications. The groups considered correspond to the atom types in the CHARMM 19 polar hydrogen potential energy function (e.g. CH₃, CH₂, CH, aromatic CH, amide N, carbonyl carbon, etc.). The solvation free

Table 4. The CASP1 homology models

CRABPI		EDN		MCHPR		NDK	
Model	W	Model	W	Model	W	Model	W
Crystal	-4316	Crystal	-4020	Crystal	-2605	Crystal	-4395
ABAGYAN	-4173	KOEHL	-3862	ABAGYAN	-2592	ABAGAYAN	-4433
MOULT1	-4179	MOULT	-3775	BIOSYM	-2568	KOEHL	-4405
MOULT2	-4210	SALI1	-3806	KOBAYASHI	-2262	SALI	-4378
SALI	-4209	SALI2	-3823	KOEHL1	-2569	VIHENEN	-4353
VINALS1	-4018	SAQI1	-3729	KOEHL2	-2554	VRIEND	-4341
VINALS2	-4175	SAQI2	836	MOSENKIS	-2559	WEBER1	-4288
VINALS	-3971	VINALS1	-3653	MOULT	-2564	WEBER2	-4318
WEBER1	-4131	VINSALS2	-3747	VRIEND	-2567		
WEBER2	-4124	VINALS3	-3844	WEBER	-2537		
		WEBER	-3639				

All energies in kcal/mol. The quantity W is defined in equation (2).

For details concerning the proteins and their crystal structures see <http://prostar.carb.nist.gov>

energy density is given by a Gaussian function ($f_i 4\pi r_{ij}^2 = \alpha_i \exp(-x_{ij}^2)$) with $x_{ij} = (r_{ij} - r_{\min,i})/\lambda_i$ where $r_{\min,i}$ is the van der Waals radius of group i and λ_i is a correlation length (3.5 Å for most groups). The value of α_i is obtained from the solvation data and the requirement that the solvation free energy of deeply buried groups be approximately zero. Ionic side-chains are neutralized and a distance-dependent dielectric constant is used for the electrostatic interactions.

In performing threading tests, the quality of the constructed "decoys" is at least as important as the quantity (Park *et al.*, 1997; Park & Levitt, 1996). There is no point considering the vast numbers of non-compact conformations that can be generated, since it can be assumed that almost all native proteins are compact and many energy functions would be able to distinguish them from the native state, i.e. the alternate folds considered should be as "native-like" as possible. As the work of Novotny *et al.* (1984) showed for empirical energy functions of the CHARMM type, simple evaluations of the energy for an X-ray or a misfolded structure are not sufficient. To obtain a meaningful result, energy minimizations or better MD simulations should be performed to anneal the protein conformations prior to application of the effective energy function (Novotny *et al.*, 1988).

We first performed the threading tests on a set of native-misfolded pairs created by Holm & Sander (1992) available on the WWW (<http://prostar.carb.nist.gov> or <http://www.ebi.ac.uk/~holm>). Disulfide bonds were built where appropriate. In all cases, non-bonded metal ions and prosthetic groups were omitted. Since the native conformation that is given is the crystal structure whereas the misfolded conformation was minimized for 500 steepest descent steps with GROMOS (van Gunsteren & Berendsen, 1987), direct comparison of the energies of the two would not be appropriate. We minimized both models for 300 ABNR steps (Brooks *et al.*, 1983) with the function to be tested (EEF1 or CHARMM 19). Subsequently, the structures were subjected to 50 ps of MD simulations, again with the functions to be tested. The simulations started at 50 K and the temperature was increased by 50 K every 2 ps until a final temperature of 300 K was reached; the simulations continued for 40 ps in a microcanonical ensemble. No MD simulation was performed for 1fdx (has two Fe₄-S₄ clusters), 1hip (has one Fe₄-S₄ cluster) and 2tmn (has four Ca and one Zn atom) due to lack of solvation parameters for these metals.

The next "Novotny" test involved threading the small protein CI2 into the fold of eight proteins of similar size; the proteins are listed in Table 3. This protein was chosen because its stability and unfolding have been studied with the EEF1 potential (Lazaridis & Karplus, 1997). The procedure for the calculations was similar to the one used for the Holm-Sander structures. The backbone coordinates of the experimental structure of each of the eight proteins was used; if the target protein was longer, the additional C terminal residues were omitted. In the one case that the target protein was shorter (1shg) the remaining part of the CI2 backbone was built in an extended conformation. The CI2 side-chains were then built onto the backbone in an extended conformation and the energy was minimized for 300 ABNR steps. For comparison, we include CI2 with the side-chains built in the same way. This procedure was sufficient for all cases except for 1ctf and 2cro, where an aromatic side-chain was built such that other protein bonds went through the ring. Since energy minimization does not remove such an interaction, the aromatic ring was moved out of

the way "by hand" using QUANTA (Molecular Simulations, Inc.). After energy minimization, the MD simulation procedure described above was used; at the end of the simulation the structure was minimized with 300 ABNR steps. For eglin c, which has the same fold as CI2, two studies were conducted: one with the correct alignment, aligning Glu20 of eglin c with Leu20 of CI2 and deleting residue 72A of eglin c, and one with an alignment in which the first residue of the truncated CI2 (Leu20) corresponded to the first eglin c residue (Thr15).

The third decoy set was the "asilomar" set from J. Moults Web site (<http://prostar.carb.nist.gov>). These are homology models submitted at the first CASP protein structure prediction meeting. A few models were not considered because of significant sequence mismatches between the models and the crystal structure. In case of limited sequence mismatches, the mismatching residue in the model was replaced by that in the crystal structure and the missing atoms were built in ideal geometry and extended conformation. As above, all models and the crystal structures were subjected to a 300 step energy minimization using EEF1.

The final set of decoys was the "4state_reduced" set from M. Levitt's Web site (<http://dd.stanford.edu>). These are all-atom versions of the best among thousands of models created by Park & Levitt (1996) using their four-state off-lattice model, energy minimized for 2000 steps. Approximately 650 models are available for each of seven proteins but only six were considered here. The seventh, 4rxn (rubredoxin) contains an iron-sulfur cluster which would have complicated the energy evaluation. All models and the crystal structure for these six proteins were minimized with ABNR for 300 steps using EEF1. No dynamics simulations were done for this set due to the large number of structures. Several statistical energy functions have been tested on this set of decoys (Park *et al.*, 1997). Two of them were found to be reasonably good but did not rank the crystal structure first in all cases studied.

Acknowledgments

This work was supported by a grant from the National Science Foundation. T.L. was a Burroughs Wellcome PMMB Postdoctoral fellow. We thank the people who created the decoys used in this work, as well as those who created and maintain the CARB and Stanford Web sites from where they were obtained. We also thank a referee for urging a test of EEF1 on a broader set of decoys than that used originally; it was as a result of that comment that the Park & Levitt and CASP1 decoy sets were added.

References

- Augsburger, J. D. & Scheraga, H. A. (1996). An efficient, differentiable hydration potential of peptides and proteins. *J. Comp. Chem.* **17**, 1549-1558.
- Augsburger, J. D. & Scheraga, H. A. (1997). An assessment of the accuracy of the RRIGS hydration potential: comparison to solutions of the Poisson-Boltzmann equation. *J. Comp. Chem.* **18**, 1072-1078.
- Bowie, J. U., Luthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164-170.

- Brooks, B. R., Brucoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). CHARMM: a program for macromolecular energy minimization and dynamics calculations. *J. Comput. Chem.* **4**, 187-217.
- Brooks, C. L., III, Karplus, M. & Pettitt, B. M. (1988). Proteins: a theoretical perspective of dynamics, structure, and thermodynamics. *Advan. Chem. Phys.* **71**, 1-259.
- Bryngelson, J. D. & Wolynes, P. G. (1987). Spin glasses and the statistical mechanics of protein folding. *Proc. Natl Acad. Sci. USA*, **84**, 7524-7528.
- Casari, G. & Sippl, M. J. (1992). Structure-derived hydrophobic potential. *J. Mol. Biol.* **224**, 725-732.
- Caves, L. S. D., Evanseck, J. D. & Karplus, M. (1998). Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin. *Protein Sci.* **7**, 649-666.
- DeBolt, s. E. & Skolnick, J. (1996). Evaluation of atomic level mean force potentials *via* inverse folding and inverse refinement of protein structures. *Protein Eng.* **9**, 637-655.
- Eisenberg, D. & McLachlan, A. D. (1986). Solvation energy in protein folding and binding. *Nature*, **319**, 199-203.
- Elber, R. & Karplus, M. (1987). Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin. *Science*, **235**, 318-321.
- Finkelstein, A. V. (1997). Protein structure: what is it possible to predict now?. *Curr. Opin. Struct. Biol.* **7**, 60-71.
- Fraternali, F. & 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.
- Frauenfelder, H., Sligar, S. G. & Wolynes, P. G. (1991). The energy landscapes and motions of proteins. *Science*, **254**, 1598-1603.
- Gibson, K. D. & Scheraga, H. A. (1969). Minimization of polypeptide energy V. Theoretical aspects. *Physiol. Chem. Phys.* **1**, 109-126.
- Holm, L. & Sander, C. (1992). Evaluation of protein models by atomic solvation preference. *J. Mol. Biol.* **225**, 93-105.
- Humphreys, D. D., Friesner, R. A. & Berne, B. J. (1995). Simulated annealing of a protein in a continuum solvation by multiple-time-step molecular dynamics. *J. Phys. Chem.* **99**, 10674-10685.
- Kang, Y. K., Nemethy, G. & Scheraga, H. A. (1987). Free energies of hydration of solute molecules 1. Improvement of the hydration shell model by exact computations of overlap volumes. *J. Chem. Phys.* **91**, 4105-4109.
- Karplus, M. & Shakhnovich, E. (1992). Protein folding: theoretical studies of thermodynamics and dynamics. In *Protein Folding* (Creighton, T. E., ed.), pp. 127-195, Freeman, New York.
- Karplus, M., Ichiye, T. & Pettitt, B. M. (1987). Configurational entropy of native proteins. *Biophys. J.* **52**, 1083-1085.
- Koehl, P. & Delarue, M. (1994). Polar and nonpolar environments in the protein core: implications for folding and binding. *Proteins: Struct. Funct. Genet.* **20**, 264-278.
- Lazaridis, T. & Karplus, M. (1997). "New view" of protein folding reconciled with the old through multiple unfolding simulations. *Science*, **278**, 1928-1931.
- Lazaridis, T. & Karplus, M. (1999). Effective energy function for proteins in solution. *Proteins: Struct. Funct. Genet.* **35**, 132-152.
- Levitt, M. (1983). Molecular dynamics of native protein II. Analysis and nature of motion. *J. Mol. Biol.* **168**, 621-657.
- Marchler-Bauer, A., Levitt, M. & Bryant, S. H. (1997). A retrospective analysis of CASP2 threading predictions. *Proteins: Struct. Funct. Genet., Suppl.* **1**, 83-91.
- Marti-Renom, M., Stote, R., Querol, E., Aviles, F. & Karplus, M. (1998). Refolding of potato carboxypeptidase inhibitor by molecular dynamics simulations with disulphide bond constraints. *J. Mol. Biol.* **284**, 145-172.
- Monge, A., Lathrop, E. J. P., Gunn, J. R., Shenkin, P. S. & Freisner, R. A. (1995). Computer modeling of protein folding: conformational and energetic analysis of reduced and detailed protein models. *J. Mol. Biol.* **247**, 995-1012.
- Moult, J. (1997). Comparison of database potential and molecular mechanics force fields. *Curr. Opin. Struct. Biol.* **7**, 194-199.
- Neria, E., Fischer, S. & Karplus, M. (1996). Simulation of activation free energies in molecular systems. *J. Chem. Phys.* **105**, 1902-1921.
- Noguti, T. & Go, N. (1989). Structural basis of hierarchical multiple substates of a protein. I. Introduction. *Proteins: Struct. Funct. Genet.* **5**, 97-103.
- Novotny, J., Brucoleri, R. & Karplus, M. (1984). An analysis of incorrectly folded protein models. Implications for structure predictions. *J. Mol. Biol.* **177**, 787-818.
- Novotny, J., Rashin, A. A. & Brucoleri, R. E. (1988). Criteria that discriminate between native proteins and incorrectly folded models. *Proteins: Struct. Funct. Genet.* **4**, 19-30.
- Park, B. H. & Levitt, M. (1996). Energy functions that discriminate X-ray and near-native fold from well-constructed decoys. *J. Mol. Biol.* **258**, 367-392.
- Park, B. H., Huang, E. S. & Levitt, M. (1997). Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J. Mol. Biol.* **266**, 831-846.
- Privalov, P. L. & Makhatadze, G. I. (1993). Contribution of hydration to protein folding thermodynamics II. The entropy and Gibbs energy of hydration. *J. Mol. Biol.* **232**, 660-679.
- Reva, B. A., Finkelstein, A. V., Sanner, M. F. & Olson, A. J. (1997). Residue-residue mean-force potentials for protein structure recognition. *Protein Eng.* **10**, 865-876.
- Šali, A., Shakhnovich, E. & Karplus, M. (1994). How does a protein fold?. *Nature*, **369**, 248-251.
- Samudrala, R. & Moult, J. (1998). An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.* **275**, 895-916.
- Smith, L. J., Redfield, C., Smith, R. A. G., Dobson, C. M., Clore, G. M., Gronenborn, A. M., Walter, M. R., Naganbushan, T. L. & Wlodawer, A. (1994). Comparison of 4 independently determined structures of human recombinant interleukin-4. *Nature Struct. Biol.* **1**, 301-310.
- Still, W. C., Tempczyk, A., Hawley, R. C. & Hendrickson, T. (1990). Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **112**, 6127-6129.
- Stouten, P. F. W., Frommel, C., Nakamura, H. & Sander, C. (1993). An effective solvation term based on

- atomic occupancies for use in protein simulations. *Mol. Simul.* **10**, 97-120.
- van Gunsteren, W. F. & Berendsen, H. J. C. (1987). *GROMOS*, University of Groningen, The Netherlands.
- Vorobjev, Y. N., Amagro, J. C. & Hermans, J. (1998). Discrimination between native and intentionally misfolded conformations of proteins: ES/IS. *Proteins: Struct. Funct. Genet.* **32**, 399-4132.
- Wang, Y., Zhang, H., Li, W. & Scott, R. A. (1995). Discriminating compact nonnative structures from the native structure of globular proteins. *Proc. Natl Acad. Sci. USA*, **92**, 709-713.
- Wesson, L. & Eisenberg, D. (1992). Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Sci.* **1**, 227-235.
- Williams, R. L., Vila, J., Perrot, G. & Scheraga, H. A. (1992). Empirical solvation models in the context of conformational energy searches: application to BPTI. *Proteins: Struct. Funct. Genet.* **14**, 110-119.

Edited by A. R. Fersht

(Received 8 March 1999; accepted 10 March 1999)