Free energy determinants of tertiary structure and the evaluation of protein models

DONALD PETREY AND BARRY HONIG

Department of Biochemistry and Molecular Biophysics, Columbia University, Medical Hughes Medical Institute,
630 West 168th Street, New York, New York 10032

(RECEIVED July 18, 2000; FINAL REVISION September 8, 2000; ACCEPTED September 8, 2000)

Abstract

We develop a protocol for estimating the free energy difference between different conformations of the same polypeptide chain. The conformational free energy evaluation combines the CHARMM force field with a continuum treatment of the solvent. In almost all cases studied, experimentally determined structures are predicted to be more stable than misfolded “decoys.” This is due in part to the fact that the Coulomb energy of the native protein is consistently lower than that of the decoys. The solvation free energy generally favors the decoys, although the total electrostatic free energy (sum of Coulomb and solvation terms) favors the native structure. The behavior of the solvation free energy is somewhat counterintuitive and, surprisingly, is not correlated with differences in the burial of polar area between native structures and decoys. Rather, the effect is due to a more favorable charge distribution in the native protein, which, as is discussed, will tend to decrease its interaction with the solvent. Our results thus suggest, in keeping with a number of recent studies, that electrostatic interactions may play an important role in determining the native topology of a folded protein. On this basis, a simplified scoring function is derived that combines a Coulomb term with a hydrophobic contact term. This function performs as well as the more complete free energy evaluation in distinguishing the native structure from misfolded decoys. Its computational efficiency suggests that it can be used in protein structure prediction applications, and that it provides a physically well-defined alternative to statistically derived scoring functions.

Keywords: conformational free energy; continuum electrostatics; fold specificity; misfolded proteins; model evaluation; protein stability

Protein structure prediction requires a method of comparing two three-dimensional models of a particular sequence and determining which model is “closer” to the native structure. This is frequently accomplished through the use of a scoring function that is related in some way to the conformational free energy of a protein. A minimal test of the accuracy of such scoring functions is that they can be used to distinguish an experimentally determined protein structure from any incorrect or nonnative conformation of the same amino acid chain. Several sets of such nonnative conformations or “decoys” have been constructed for this purpose. Among the most commonly used are the EMBL deliberately misfolded set (Holm & Sander, 1992), the CASP1 set, and the Park and Levitt set (Park & Levitt, 1996). The EMBL set consists of pairs of native and misfolded structures created by taking the native sequence and modeling it onto the backbone of a structurally dissimilar protein with the same number of residues (Novotny et al., 1984). The CASP1

set consists of experimentally determined native structures and homology models submitted to CASP1 (Mosimann et al., 1995). The Park and Levitt set consists of experimentally determined structures and incorrect conformations constructed by defining “hinge regions” between the native secondary structure elements and allowing these hinge regions to occupy a small number of states in ϕ/ψ space. Conformations that are not “protein-like” are then removed, resulting in ~ 650 decoys for each native structure.

Scoring functions used for the evaluation of protein models generally fall into two broad categories (see Lazaridis & Karplus, 2000, for a recent review). “Statistical” effective energy functions, termed SEEFs by Lazaridis and Karplus (2000), are based on the observed properties of amino acids in known structures and have been widely used in fold recognition and homology modeling applications. Physical effective energy functions (PEEFs) are based on a direct evaluation of the conformational free energy of a protein. Although early studies demonstrated the difficulties involved in using PEEFs to evaluate models (Novotny et al., 1984), recent work has demonstrated that such a direct evaluation of the conformational free energy can be at least as successful as statistically based scoring functions in distinguishing the native structure of a protein from an incorrectly folded decoy, although generally at

Reprint requests to: Barry Honig, Department of Biochemistry and Molecular Biophysics, Columbia University, 630 West 168th Street, New York, New York 10032; e-mail: bh6@columbia.edu.

Abbreviations: FDPB, finite-difference Poisson-Boltzmann; PEEF, physical effective energy function.

greater computational cost. Janardhan and Vajda (1998) combined the CHARMM19 force field (Brooks et al., 1983) and the atomic solvation parameters of Eisenberg and McLachlan (1986) to develop a PEEF that achieved excellent discrimination of the native conformations in the cases they studied. Vorobjev et al. (1998) evaluated conformational free energy differences by averaging the Coulomb, van der Waals, covalent, solvation, and cavity formation energies over a sampling of conformations obtained from molecular dynamics simulations. They applied this protocol to a subset of the EMBL deliberately misfolded set of decoys and obtained 100% discrimination. Lazaridis and Karplus (1998) use the CHARMM19 (Brooks et al., 1983) force field with a novel solvation model to evaluate the free energy difference between native structures and misfolded models in a number of decoy sets. They achieve 100% discrimination of native structures from misfolded models in the EMBL deliberately misfolded set, a set of structures obtained from threading a small protein onto other templates, and the Park and Levitt set. They were also quite successful for the CASPI set. The PEEF used by Lazaridis and Karplus assumed that all ionizable residues were in their neutral state.

A distinct advantage of such physically derived functions is that they are based on well-defined physical interactions, thus making it easier to learn and to gain insight from their performance. In this paper, we develop a protocol to calculate the conformational free energy that is based on a continuum solvent approach developed previously, and which was used to identify the determinants of secondary structure stability (Yang & Honig, 1995a, 1995b; Yang

et al., 1996) and apply it to the decoy sets described above. The calculations are not efficient enough to be applied to large sets of possible structures, for example, to many alternate alignments of a sequence to a structural template. Rather, our goal was to ensure that the energy evaluation was as complete as possible, including the effects of ionized side chains. By evaluating the individual terms that contribute to the conformational free energy, we are able to identify those terms that are most effective in discriminating native from misfolded structures. A somewhat surprising finding is that, in most cases, Coulomb interactions involving charged side chains favor the native conformation. On this basis, a simple electrostatic scoring function was developed that, when used in conjunction with a simple hydrophobic term, provides discrimination of native from decoy structures that is as effective as that obtained using the more complete free energy evaluation. Our results suggest that coulombic interactions may play an important role in determining a protein's structure. Despite a number of papers that argue that this is indeed the case (Spasov et al., 1994, 1997), electrostatic interactions have been largely ignored in discussions of the factors that determine protein topology.

Results

Conformational free energies of decoys and native states

Tables 1–3 contain comparisons of the relative free energies of native and incorrect structures. Table 1 shows the results for the

Table 1. Differences between native and decoy conformations in the EMBL deliberately misfolded set^a

Native/decoy	All-atom energy function						Empirical measures				
	ΔG_{vdW}	ΔG_{cov}	ΔG_{coul}	ΔG_{solv}	ΔG_{elec}	ΔG	ΔA_{hp}	ΔA_{pol}	ΔE_{coul}	ΔE_{hp}	ΔE
1bp2/2paz	-28	58	-813	134	-679	-648	-163	155	-753	-402	-1,155
1cbh/1ppt	-39	7	19	14	33	-3	-301	-276	-31	-18	-49
1fdx/5rxn	-34	51	87	-48	39	53	-569	-139	0	-234	-234
1hip/2b5c	-62	3	-478	155	-323	-388	-895	24	-526	-387	-913
1lh1/2ilb	-79	-34	-702	7	-695	-812	-1,112	319	-442	-916	-1,358
1p2p/1m3	-4	134	-446	109	-337	-211	-212	-346	-446	-460	-906
1ppt/1cbh	17	-46	7	-130	-123	-151	-57	402	67	-139	-72
1rhd/2cyp	-137	129	-836	-20	-856	-864	-1,104	1,368	-981	-1,747	-2,728
1m3/1p2p	-112	-46	-331	70	-261	-421	-134	-182	-329	-448	-777
1sn3/2ci2	-51	18	-564	165	-399	-434	-255	-292	-502	-58	-560
1sn3/2cro	-30	-38	-154	-10	-164	-232	-51	-109	-193	-13	-206
2b5c/1hip	-68	-26	-542	67	-475	-570	-374	285	-312	-341	-653
2cdv/2ssi	18	-40	-964	274	-690	-712	1,038	146	-881	-110	-991
2ci2/1sn3	-30	-25	-562	164	-398	-453	-202	247	-482	-301	-783
2ci2/2cro	-15	-16	-587	185	-402	-433	43	132	-588	-238	-826
2cro/1sn3	-56	-24	-573	154	-419	-503	-372	-3	-519	-341	-860
2cro/2ci2	-69	13	-305	39	-266	-327	-769	-130	-221	-321	-542
2cyp/1rhd	-178	-223	-1,664	96	-1,568	-1,973	-1,884	672	-1,236	-1,549	-2,785
2ilb/1lh1	-129	-68	-382	151	-231	-433	-1,705	839	-643	-1,000	-1,643
2paz/1bp2	-145	-7	-161	-35	-196	-355	-1,117	234	-88	-661	-749
2ssi/2cdv	-79	61	-279	127	-152	-177	-1,836	-183	-230	-508	-738
2tmn/2ts1	-371	-13	-1,296	368	-928	-1,330	-3,981	-254	-1,409	-1,108	-2,517
2ts1/2tmn	-102	-275	-1,081	27	-1,054	-1,432	-1,047	1,514	-981	-1,590	-2,571

^aEnergies are expressed in kcal/mol and areas in Å². Differences are taken with the decoy as the reference state (i.e., $\Delta X = X_{native} - X_{decoy}$). G_{vdW} = van der Waals energy; G_{cov} = bond stretching + bond bending + torsional energy; G_{coul} = Coulomb energy ($\epsilon = 1$); G_{solv} = solvation energy; $\Delta G_{elec} = G_{coul} + G_{solv}$; A_{hp} = accessible hydrophobic area; A_{pol} = accessible polar area; E_{coul} = coulombic component of simplified PEEF; E_{hp} = hydrophobic component of simplified PEEF; E = total simplified PEEF.

Table 2. Differences between native and decoy conformations in CASP1 set^a

Native/decoy	All-atom energy function						Empirical measures				
	ΔG_{vdW}	ΔG_{cov}	ΔG_{coul}	ΔG_{solv}	ΔG_{elec}	ΔG	ΔA_{hp}	ΔA_{pol}	ΔE_{coul}	ΔE_{hp}	ΔE
CRABPI/ABAGYAN	-66	-53	-381	111	-270	-392	-339	-259	-440	-368	-808
CRABPI/MOULT1	-38	-81	-231	30	-201	-319	-238	-300	-162	-64	-226
CRABPI/MOULT2	-61	-38	-134	24	-110	-210	-172	-173	-24	-146	-170
CRABPI/VINALS1	-16	-94	-441	42	-399	-508	-413	247	-194	-402	-596
CRABPI/VINALS2	-10	-65	174	-135	39	-35	29	409	178	-329	-151
CRABPI/VINALS3	-22	-162	-807	185	-622	-805	-643	293	-688	-241	-929
CRABPI/WEBER2	-52	-146	14	-140	-126	-323	-177	766	110	-335	-225
EDN/KOEHL	-87	55	-94	-80	-174	-211	-266	182	63	-332	-269
EDN/MOULT	-102	-121	-102	-17	-119	-345	-224	-248	-20	-154	-174
EDN/SALI1	-121	39	-349	66	-283	-370	-410	-245	-212	-112	-324
EDN/SALI2	-100	47	-423	92	-331	-388	-435	-92	-266	-104	-370
EDN/SAQI1	-164	-91	128	-142	-14	-272	-181	201	159	-400	-241
EDN/SAQI2	-1,445	-2,952	-53	-43	-96	-4,500	-1,548	-239	3	-86	-83
EDN/VINALS1	-61	-137	-222	-47	-269	-468	-331	565	-101	-227	-328
EDN/VINALS2	-43	-148	-61	-94	-155	-345	-197	785	-43	-253	-296
EDN/VINALS3	-46	-71	152	-128	24	-95	-23	393	165	-191	-26
MCHPR/ABAGYAN	-19	-1	-238	118	-120	-141	-140	53	-182	-72	-254
MCHPR/BIOSYM	-20	13	-21	23	2	-6	-19	-330	-28	-50	-78
MCHPR/KOBAYASHI	-121	-94	-658	239	-419	-643	-549	-540	-392	-314	-706
MCHPR/KOEHL1	-31	36	-354	132	-222	-218	-254	-243	-238	-108	-346
MCHPR/KOEHL2	-40	55	-295	87	-208	-194	-249	-124	-192	-184	-376
MCHPR/MOSENKIS	-27	-7	-259	115	-144	-180	-173	-319	-224	-36	-260
MCHPR/MOULT	-33	-10	-199	83	-116	-159	-150	-209	-209	-52	-261
MCHPR/VRIEND	-61	-25	-327	121	-206	-294	-269	-73	-331	-104	-435
MCHPR/WEBER	-23	-31	-67	19	-48	-101	-70	70	-73	-57	-130
NUE/KOEHL	48	20	-254	1	-253	-184	-203	623	-137	-82	-219
NUE/SALI	8	23	-305	90	-215	-184	-206	-13	-241	-3	-244
NUE/VRIEND	-44	-58	-300	28	-272	-372	-314	33	-156	140	-16
NUE/WEBER1	-84	-124	-194	30	-164	-372	-248	423	-67	-44	-111
NUE/WEBER2	-44	-102	-327	10	-317	-461	-359	427	-223	53	-170

^aEnergies are expressed in kcal/mol and areas in Å². Differences are taken with the decoy as the reference state (i.e., $\Delta X = X_{native} - X_{decoy}$). G_{vdW} = van der Waals energy; G_{cov} = bond stretching + bond bending + torsional energy; G_{coul} = Coulomb energy ($\epsilon = 1$); G_{solv} = solvation energy; $\Delta G_{elec} = G_{coul} + G_{solv}$; A_{hp} = accessible hydrophobic area; A_{pol} = accessible polar area; E_{coul} = coulombic component of simplified PEEF; E_{hp} = hydrophobic component of simplified PEEF; E = total simplified PEEF.

EMBL deliberately misfolded set for which 22 out of 23 of the native/decoy pairs are discriminated correctly. The only error occurred for the pair ferredoxin/rubredoxin (Protein Data Bank (PDB) codes 1fdx and 5rxn) where the two Fe-S clusters in the core of this protein are not included in the calculations. Also, the major source of destabilization in the native structure is the interaction between D53 and E54, which have their side-chain carboxyls within a van der Waals radius of one another. This energetically unfavorable side-chain conformation is stabilized by crystal packing in the native structure, which places the N-terminal directly in contact with D53 and E54. When side chains are rebuilt using the program Modeller as we discuss below, the unfavorable interaction between D53 and E54 is reduced and 100% discrimination is achieved. It may be the case that one of these residues is neutral in the native protein or that they are farther away than indicated by the crystal structure.

Table 2 shows the results for the CASP1 set for which all of the native/decoy pairs are discriminated correctly. The large energy differences between the native and decoy structure in the case of edn_saqi2 are due to improper positioning of the side chains, which

could not be corrected by the minimization procedure (a phenyl ring is penetrated by the side chain of a methionine).

Table 3 shows the results for the “pseudo-native” set for which all of the native/decoy pairs are discriminated correctly. This set was generated by removing side chains from the native and decoy structures in the EMBL deliberately misfolded set and rebuilding the side chains using the program Modeller (Sali & Blundell, 1993). The average root-mean-square deviation (RMSD) between the structures in the pseudo-native set and their counterparts in the EMBL deliberately misfolded set was 1.9 Å. Use of the pseudo-native set provides a more meaningful test of our methods in that in real modeling applications the side-chain conformations have to be constructed on a particular backbone.

Individual contributions to free energy differences

Tables 1–3 also report the values of the various terms that contribute to the calculated conformational free energy differences. These include bond stretching, bond bending, and torsional energy terms (ΔG_{cov}), the van der Waals energy (ΔG_{vdW}), the Coulomb

Table 3. Differences between native and decoy conformations in EMBL deliberately misfolded set^a

Native/decoy	All-atom energy function						Empirical measures				
	ΔG_{vdW}	ΔG_{cov}	ΔG_{coul}	ΔG_{solv}	ΔG_{elec}	ΔG	ΔA_{hp}	ΔA_{pol}	ΔE_{coul}	ΔE_{hp}	ΔE
1bp2/2paz	45	-42	-286	-161	-447	-438	-40	526	-341	-233	-574
1cbh/1ppt	-46	-28	-6	63	57	-20	-289	-243	-36	-10	-46
1fdx/5rxn	-34	-8	-20	45	25	-19	-547	85	-65	-194	-259
1hip/2b5c	-28	-12	-318	195	-123	-167	-929	213	-408	-328	-736
1lh1/2i1b	-58	23	-454	-8	-462	-502	-911	486	-222	-747	-969
1p2p/1rn3	-15	18	-199	130	-69	-67	-162	-87	-244	-334	-578
1ppt/1cbh	38	7	186	-354	-168	-120	158	524	208	-95	113
1rhd/2cyp	-103	-62	-643	112	-531	-698	-645	1,411	-876	-1,466	-2,342
1rn3/1p2p	-57	-70	-335	281	-54	-184	-349	136	-341	-348	-689
1sn3/2ci2	-27	-13	-232	142	-90	-131	-143	-230	-254	6	-248
1sn3/2cro	-15	-67	3	-83	-80	-160	74	-2	-100	9	-91
2b5c/1hip	-18	12	-292	-130	-422	-428	-458	655	-165	-182	-347
2cdv/2ssi	7	-118	-371	209	-162	-274	831	170	-396	-114	-510
2ci2/1sn3	-10	-45	-269	153	-116	-172	-216	422	-217	-224	-441
2ci2/2cro	-11	-67	-287	217	-70	-148	105	290	-389	-183	-572
2cro/1sn3	-23	-13	-548	321	-227	-265	-420	165	-406	-225	-631
2cro/2ci2	-22	-20	-304	192	-112	-154	-519	73	-223	-244	-467
2cyp/1rhd	-67	-140	-1,476	165	-1,311	-1,522	-1,583	1,120	-1,032	-1,150	-2,182
2i1b/1lh1	-19	-63	161	-205	-44	-124	-1,215	1,093	-167	-830	-997
2paz/1bp2	-77	-51	-245	128	-117	-252	-1,137	294	-216	-448	-664
2ssi/2cdv	-109	0	-250	264	14	-103	-1,495	-241	-168	-563	-731
2tmn/2ts1	-158	-33	-1,715	1,265	-450	-655	-3,271	-59	-1,748	-740	-2,488
2ts1/2tmn	13	-207	-679	-368	-1,047	-1,238	-562	2,399	-667	-927	-1,594

^aEnergies are expressed in kcal/mol and areas in Å². Differences are taken with the decoy as the reference state (i.e., $\Delta X = X_{native} - X_{decoy}$). G_{vdW} = van der Waals energy; G_{cov} = bond stretching + bond bending + torsional energy; G_{coul} = Coulomb energy ($\epsilon = 1$); G_{solv} = solvation energy; $\Delta G_{elec} = G_{coul} + G_{solv}$; A_{hp} = accessible hydrophobic area; A_{pol} = accessible polar area; E_{coul} = coulombic component of simplified PEEF; E_{hp} = hydrophobic component of simplified PEEF; E = total simplified PEEF.

energy (ΔG_{coul}), the solvation energy (ΔG_{solv}), and the total electrostatic energy ($\Delta G_{elec} = \Delta G_{coul} + \Delta G_{solv}$). Interestingly, each term on its own generally favors the native state except for ΔG_{solv} , which tends to favor the decoy, and ΔG_{cov} , which, by itself, is not a particularly good criterion for identifying the native state for the examples in the CASP1 decoy set. Moreover, in almost all cases, the native structure would have been calculated to be more stable than the decoy even if ΔG_{cov} had been excluded.

The van der Waals energy is found to consistently favor the native structures, by itself correctly identifying 21 out of 23 native structures in the EMBL deliberately misfolded set, 19 out of 23 native structures in the pseudo-native set, and 28 out of 30 native structures in the CASP1 set. It is not surprising that native structures are better packed than decoys, but it is perhaps more interesting that it is possible to achieve better packing with the native backbone, as indicated by the results for the pseudo-native set shown in Table 3. This might suggest that an estimate of the van der Waals energy, or the quality of packing, should provide an efficient means of distinguishing correct from misfolded structures. The problem, as is evident from the different values for ΔG_{vdW} in Tables 1 and 3, is that in real applications it is necessary to add side chains to a given backbone, and the quality of the packing will depend on the algorithm used to add side chains. Moreover, minimization is usually required to relieve bad van der Waals contacts making the inclusion of such a term computationally expensive.

As is evident from Tables 1–3, the electrostatic free energy favors the native state in most cases. Quite surprisingly, this is due

in large part to the Coulomb term, which consistently favors the native structure. Twenty out of 23 native structures in the EMBL deliberately misfolded set, 20 out of 23 native structures in the pseudo-native set, and 26 out of 30 native structures in the CASP1 set are correctly identified as more stable based on the Coulomb term alone. In contrast, the solvation term tends to favor the incorrect structure, even in situations where the decoy buries more polar area than the native conformation. It is also surprising that there appears to be no correlation between buried polar area and the tendency of the solvation term as calculated by the FDBP method to favor the native or decoy structure. In fact, as shown in Tables 1–3 (see the pair 2cyp/1rhd in Table 1, for example), a native structure can have both more buried hydrophobic area and less buried polar area and still have a less favorable solvation energy.

The fact that the Coulomb energy tends to favor the native structure implies that these structures have a more favorable charge distribution than that of the decoys. To understand why the solvation free energy has the opposite effect, we consider the electrostatic free energy of two simple charge distributions in a spherical cavity (Fig. 1). The four charges shown in the figure are buried at the same depth from the surface in Figures 1A and 1B, but the charge distributions are different. As shown in the figure, the charge distribution with the more favorable Coulomb energy has the smaller solvation free energy. This can be understood in simple terms by noting that the charges in Figure 1A produce a zero dipole moment, while those in Figure 1B have a finite dipole moment. The dipole in Figure 1B interacts more strongly with the solvent than

Fig. 1. Simple system used to model the process of changing the native charge distribution (State A) to the decoy charge distribution (State B). G_{coul} is coulomb energy and G_{solv} is the solvation energy.

the quadrupole of Figure 1A, thus accounting for the more favorable solvation term. It is reasonable to expect that the picture provided by Figure 1 should be generally valid because a favorable Coulomb energy implies a favorable distribution of positive and negative charges, which, in turn, should produce a lower dipole moment than a charge distribution that is not optimized. As is evident from Tables 1–3, the expected behavior is observed in most but not all cases.

In addition to the terms calculated directly, we have also determined the ability of a traditional measure of protein stability, buried hydrophobic area (ΔA_{hp} in Tables 1–3), to distinguish native from decoy structures. Note that this term is not calculated explicitly in Equation 1, which is based on a thermodynamic cycle involving the gas phase and thus does not include a term that corresponds the transfer of nonpolar atoms from the alkane to the aqueous phase. In both the EMBL deliberately misfolded set and the pseudo-native set, the buried hydrophobic area term alone correctly identifies 21 out of 23 of the native structures. This is to be expected, but it is more surprising that 29 out of the 30 native structures in the CASP1 set are also identified by this term. This suggests the possibility that alignments used in homology modeling might be improved by optimizing buried hydrophobic area.

A simple empirical PEEF

The analysis in the previous section suggests that both buried hydrophobic area and the Coulomb energy are effective means of identifying the native conformation. A scoring function that is essentially the sum of the two might thus be expected to perform better than either term on its own. Our goal was to develop such a scoring function that can be applied to a large number of conformations without requiring the time-consuming processes of adding hydrogen atoms to crystallographically determined structures and extensive energy minimizations required, for example, to relax steric clashes. As described in Materials and methods, we have devised such a function, which includes a fast gas phase analysis of the Coulomb energy ΔE_{coul} and a contact-based measure of buried hydrophobic area ΔE_{hp} similar to that used by Huang and Levitt (Huang et al., 1995). The effect of these terms and their sum are indicated in Tables 1–3. Their sum ΔE correctly identified all

23 native structures in the EMBL set (better than our full free energy evaluation), 22 out of 23 native structures in the pseudo-native set, and all of the native structures in the CASP1 set. It is evident from the tables that the two terms in the scoring function complement each other. In cases where one term alone is ambiguous or incorrect, the other term compensates. For example, the hydrophobic term for the pairs 1sn3/2ci2, 1sn3/2cro, and 2cdv/2ssi weakly favors the native structure, while the electrostatic term strongly favors the native structure. Similarly, the electrostatic term for the pairs 1fdx/5rxn and 1ppt/1cbh favors the decoy but the hydrophobic term more than compensates for this effect leading to the correct identification of the native structure.

The fact that our scoring function performs almost as well for the native and pseudo-native set (Tables 1, 3) indicates that its success does not depend on the algorithm used to add side chains. To show that the scoring function is also fairly insensitive to the protocol used for the minimization, we applied the scoring function to the *unminimized* structures in all three sets. In this case it identified 23 out of 23 of the correct structures in the EMBL deliberately misfolded set, 22 out of 23 native structures in the pseudo-native set, and 26 out of 30 in the CASP1 set, with three of the four errors in the CASP1 set occurring for the models built on a template with very high sequence identity to the target. Similar results were obtained even when comparing unminimized native structures with optimized decoys, indicating that the scoring function is identifying important features of the native fold that are independent of the detailed structure.

Because of the large number of decoys in the Park and Levitt set (~3,900 structures), it was not possible to apply our full free energy function to the entire set. However, the simplified PEEF was efficient enough to apply to all the structures in this set (requiring <1 s of computer time per structure on an IRIS R10000) and yielded excellent results as shown in Table 4. Identification of the X-ray structure in this set, as in the CASP1 set, is challenging for such simplified functions because native secondary structure is preserved, and many of the decoys have RMSDs less than 3 Å from native. However, we achieved good discrimination despite the fact that no processing of either the native or decoy structures in this set was performed (i.e., no minimization), except as described below. These results further confirm that the scoring function we have developed appears to measure characteristics that

Table 4. Rank of the X-ray structures in the Park and Levitt set based on the hydrophobic component (E_{hp}), the electrostatic component (E_{coul}), and the total score (E_{total}) calculated with the simplified PEEF

PDB code	# Decoys	Rank		
		E_{hp}	E_{coul}	E_{total}
1ctf	631	1	1	1
1r69	676	1	5	1
4pti	688	16	3	1
1sn3	661	132	1	3
2cro	675	2	115	4
4rxn	678	1	116	10
3icb	654	3	103	42
1clb	654	46	2	1

are optimized in native structures. For three of the seven proteins in the Park and Levitt set, the X-ray structure has the minimum energy as measured by our simplified PEEF. For two cases, 1sn3 and 2cro in Table 4, the X-ray structure is ranked third and fourth, respectively. Of the five decoys that have a lower energy than the native structure in these cases, only one has an RMSD greater than 3 Å. Moreover, application of the complete free energy function to these five decoys unambiguously identifies the native structure.

The remaining two cases are complicated by the presence of charged ligands/cofactors in the native structures, which were used to construct the decoys. Rubredoxin (4rxn) contains an iron-sulfur cluster and bovine calcium binding protein (3icb) contains two calciums. The structure used to represent the native state for these proteins is simply the X-ray structure solved in the presence of the ligands, but with ligands removed from the PDB file. It is unlikely that this conformation of the protein would be identical to its actual conformation in the absence of ligands (the apo form). Our failure to rank the native as the most stable based on the simplified PEEF is, therefore, not unexpected. When the apo structure of the calcium binding protein (1clb) is used to represent the native state, it has the lowest energy (see the results for 1clb in Table 4). For rubredoxin, no structure without the iron-sulfur cluster exists. However, of the 10 decoys that are identified as more stable based on the simplified PEEF, seven have RMSDs less than 2 Å.

Discussion

In agreement with previous work (Janardhan & Vajda, 1998; Lazaridis & Karplus, 1998; Vorobjev et al., 1998), we have shown that PEEFs can be quite effective in discriminating native from incorrectly folded structures. The PEEF used in this work is based on a gas phase evaluation of the Coulomb, van der Waals, and covalent energy and a gas to water solvation free energy obtained from the FDPB method (Equation 1). An important new result of this paper is that the gas phase Coulomb energy is by itself capable of identifying the native structure. On this basis, a simplified PEEF is developed that combines a gas phase electrostatic term and a hydrophobic contact term. Because this function can be applied directly to protein structures without requiring time-consuming minimization protocols, it overcomes a major hurdle in the application of PEEFs to real structure prediction problems.

The fact that the total electrostatic energy favors the native state of a protein relative to other possible compact, folded conformations has been observed previously in different contexts. In an early study of 14 crystal structures, Wada and Nakamura (1981) showed that there is a favorable distribution of charge-charge and charge-backbone interactions in native proteins. More recently, Spassov and coworkers (Spassov et al., 1994, 1997) have argued convincingly that the total electrostatic energy in a large set of nonhomologous proteins is more favorable than what would be expected for a random distribution of charge. That the coulombic component of the total electrostatic energy seems to be optimized independently has also been suggested by Oliva and Moulton (1999), who showed that local coulomb interactions (<6 Å) are more frequently favorable than unfavorable in native structures.

Our work builds on these studies, but extends them in a number of ways. First, we have shown that the Coulomb energy is an effective means of discriminating native folds from misfolded decoys, suggesting that the scoring function given in Equation 2 can be used in structure prediction problems. Moreover, comparisons of the native structure to decoys whose structures have been at

least partially optimized is a more stringent test of electrostatic optimization than previous comparisons to randomly generated charge distributions (Spassov et al., 1994, 1997).

More importantly, the argument illustrated in Figure 1 shows how this optimization can occur despite a generally unfavorable solvation free energy, and further emphasizes the importance of charge-charge and charge-dipole interactions as a determinant of a protein's fold. That is, the general picture that emerges is that Coulomb forces are at least partially optimized in folded proteins despite the cost in solvation free energy. Why the two effects generally work in opposite directions is evident from the discussion of Figure 1, but can be also understood by considering two cases in which the positive and negative charges are either dispersed evenly over the surface of a protein, or polarized in such a way that all positive charges are clustered together at one pole of a protein with negative charges clustered at the other. The Coulomb energy will clearly be lower in the first case in which the protein will have no dipole moment, implying a less favorable solvation energy. Given the large dipole moment that is formed in the second case, the solvation term will be favored. Thus, our results taken together with the argument illustrated in Figure 1 imply that, given two compact conformations, the one with the more favorable Coulomb term is likely to have the more favorable total electrostatic energy as well. This is consistent with the recent observation by Grimsley et al. (1999) that the effect of mutations on the stability of proteins when the mutations did not significantly alter the total buried polar area (i.e., reversing the charge of exposed ionizable residues in the experiments they describe) correlated well with a simple calculation of the change in Coulomb energy upon mutation.

Table 5. Individual charge-charge (C/C) and charge-dipole (C/D) interaction energies for each charged residue in scorpion neurotoxin (1sn3) in its native conformation and in its decoy conformation (the conformation of chymotrypsin inhibitor)^a

Residue	Native			Decoy		
	C/C	C/D	Total	C/C	C/D	Total
N-term	-48	-3	-50	-16	-3	-18
C-term	-40	15	-26	13	5	18
Lys1	-18	-15	-33	0	11	11
Glu2	-2	-19	-22	-10	-9	-19
Lys7	-45	19	-26	-21	3	-18
Lys8	-5	-4	-9	-44	1	-43
Asp10	-59	-24	-83	-30	12	-19
Lys13	-17	-6	-23	-10	4	-6
Lys18	-29	3	-26	-5	3	-2
Glu21	8	-9	0	-17	-2	-20
Glu23	2	-25	-23	-20	11	-9
Asp26	-17	-10	-27	4	14	18
Glu28	-18	-30	-47	19	23	42
Lys30	-68	-7	-75	-41	-12	-53
Lys32	-16	-38	-54	2	13	15
Glu49	8	-17	-10	-1	13	12
Glu53	-55	5	-50	-24	-17	-41
Lys63	-33	-37	-70	-25	-5	-31

^aEnergies are reported in kcal/mol and are calculated with the formula used in our simplified PEEF (see Materials and methods).

It is important to emphasize that our results say nothing about whether electrostatic forces drive protein folding, in a thermodynamic sense, relative to the unfolded state. Indeed, previous work from our lab has argued that electrostatic forces are net destabilizing (Yang & Honig, 1995a, 1995b; Yang et al., 1996). In this paper we are comparing two compact conformations and arguing that electrostatic forces are optimized in the native structure relative to other alignments or topologies rather than to a hypothetical unfolded state. In this sense, electrostatics may be viewed as playing a role in the specificity of folded conformations much as it is believed to do in the binding of two proteins to form a complex.

We now consider the structural origins of the electrostatic optimization we have discussed. In general, both charge–charge and charge–dipole interactions favor the native structure. Typical interaction energies for each charged residue in a native and decoy structure are shown in Table 5 for the pair scorpion neurotoxin/chymotrypsin inhibitor (1sn3/2ci2) in the EMBL deliberately misfolded set. The source of the difference in the charge–charge term is a generally more favorable distribution of ionizable amino acids in the native protein, and is consistent with the observations of Wada and Nakamura (1981) and Spassov et al. (1994). As shown in Figure 2A, in the native structure there are, in general, more short-range attractive charge–charge interactions than repulsive interactions, while for the decoys, the distribution of distances at which favorable and unfavorable interactions occur is essentially

identical, indicating a random charge distribution. Nevertheless, in 13 out of 23 native/decoy pairs in the EMBL set, the sum of favorable interactions in the decoy is greater than that in the native structure. The more favorable electrostatic energy in the native structure in these cases arises because the sum of unfavorable interactions is much smaller in the native structure than in the decoy structure. As shown in Figure 2, many of these interactions occur at long range, and it is thus important to include them in a scoring function for the evaluation of models.

The source of the difference in the charge–dipole term can be traced to more specific structural features, as is illustrated by the specific example of scorpion neurotoxin/chymotrypsin inhibitor shown in Table 5. It can be seen that the interaction energy of the charged residues with the dipoles in the native structure is, in general, favorable, while in the decoy, the charge–dipole interaction energy is generally close to zero or unfavorable. Perhaps surprisingly, in both the native and decoy structures, most of the dipoles that interact significantly with a given charge are randomly oriented with respect to that charge so that the sum of the interaction energies is, therefore, approximately zero. In the native structures, however, certain local regions of the sequence have all or most of their backbone amide and carbonyl groups oriented so as to favorably to interact with a given charge.

For scorpion neurotoxin, these regions are shown in Figure 3 along with the ionizable groups with which they interact. Fig-

Fig. 2. **A:** The sum of all unfavorable (top) or favorable (bottom) charge–charge interactions in all the native structures in the EMBL deliberately misfolded set occurring at the distance specified on the x-axis and calculated by the charge–charge term in our the simplified PEEF. **B:** The same as A but for decoy structures.

Fig. 3. Charged residues in the sequence of scorpion neurotoxin that have a large net favorable charge–dipole interaction. (A) K1 with loop residues 52–55; (B) D10 and K63 with loop residues 6–12; (C) E28 and K32 with loop residues 31–37; (D) D21, D23, K30, K32 interacting with a helix dipole (helix 22–30).

ure 4A identifies these regions within the context of the entire protein. They correspond to turns that serve to define the topology of the structure as well as to a helical region against which all other

secondary structure elements in this protein pack. In the decoy, in which the sequence of scorpion neurotoxin is modeled onto the backbone of chymotrypsin inhibitor, the backbone dipoles in turns

Fig. 4. The colored regions indicate where in the native structures backbone dipoles interact favorably with charged residues in the sequence. (A) Scorpion neurotoxin (1sn3); (B) chymotrypsin inhibitor (2ci2).

and secondary structure elements are randomly oriented with respect to the charges in the scorpion neurotoxin sequence. Consequently, the total charge–dipole interaction energy for each charged residue in this structure is small and frequently unfavorable. In contrast, in the native structure of the chymotrypsin inhibitor, specific turns and secondary structure elements highlighted in Figure 4B interact extremely favorably with certain charges in the chymotrypsin inhibitor sequence. This pattern of favorable interactions between ionizable groups and the peptide backbone in the native structure but not the decoy is characteristic of many of the protein pairs in the EMBL deliberately misfolded set.

We have purposely excluded hydrogen bonds from our simplified PEEF. The inclusion of a dipole–dipole term for hydrogen bonds would introduce too much noise into the energy function because its proper calculation would require the computationally expensive and potentially inaccurate placement of hydrogens on hydroxyl groups. Moreover, calculation of such a term for main chain–main chain interactions would be dominated by the secondary structure composition of the folds being compared, which could potentially involve other energetic factors that are not accounted for in our simplified PEEF. It is then important to realize that the Coulomb interactions we calculate with Equation 3 arise entirely from the distribution of ionizable residues in the protein.

Our results raise a number of questions that deserve further study. How general is our finding that the total electrostatic energy of native structures is optimized relative to other possible compact, folded conformations? Our results using the Park and Levitt set and preliminary results on a set of decoys generated by an ab initio folding algorithm (D. Petrey, unpubl. obs.) support the observation that the vacuum electrostatic term is frequently individually optimized. Are the structural features described above that correlate with the favorable difference in Coulomb energy generally found in native structures? If so, then it should be possible to incorporate this information in protein structure prediction algorithms. Perhaps the most interesting question is whether coulombic interactions place restrictions on the possible stable tertiary topologies of a given peptide chain. If so, this might suggest that there is a common charge distribution in a subset of proteins with a similar fold. Work is currently underway to address these issues.

Materials and methods

Conformational free energies

Differences in the free energy between minimized conformations of the native and decoy structures were calculated using the following formula (see Fig. 5):

$$\Delta G_{\text{decoy} \rightarrow \text{native}} = -G_{\text{solvation}}^D + \Delta G_{\text{vacuum}} + G_{\text{solvation}}^N \quad (1)$$

For each native/decoy pair, the total accessible area of both the native and decoy were roughly equivalent. Given the small free energy/surface area coefficient for the transfer of molecules from the gas to the aqueous phase (Sitkoff et al., 1994), surface area-dependent terms were ignored in the energy evaluation. Similarly, the contribution of side-chain configurational entropy, which we calculated by the method of Pickett and Sternberg (1993), were found to be small and were also ignored. Consequently, the solvation energies of the native and decoy structures, $G_{\text{solvation}}^D$ and

Fig. 5. Thermodynamic cycle used to evaluate the energy difference between the decoy and native structure.

$G_{\text{solvation}}^N$, only include the electrostatic contribution to this term. ΔG_{vacuum} is calculated as

$$\Delta G_{\text{vacuum}} = \Delta G_{vdW} + \Delta G_{coul} + \Delta G_{cov}$$

where G_{vdW} is the van der Waals energy, G_{coul} is the coulombic energy, and G_{cov} is the sum of the bond, bond angle, and torsional angle energies. These terms were obtained from minimized values reported in the output of the program X-PLOR (Brünger, 1992). All differences in energy are calculated with the decoy structure as the reference state.

All molecular mechanics components of this study were carried out using the PARAM22 parameter set (MacKerrel et al., 1998) from CHARMM (Brooks et al., 1983). Hydrogen addition and minimization were performed using the program X-PLOR (Brünger, 1992). Hydrogen addition was carried out in two stages. A first round of hydrogen addition was performed using a dihedral angle search step size of 10°. The resulting structure was subjected to another round of hydrogen addition using a dihedral angle search step size of 3°.

Energy minimization was performed on each structure in two rounds as well with constraints placed on the heavy atom dihedral angles using a force constant of 20 kcal/mol·radian². A first round of 100 steps of Powell minimization was performed including only the bond angle, bond length, van der Waals, and dihedral constraint energy terms. A second round of minimization was performed with the addition of the electrostatic and improper energy terms and carried out to a tolerance gradient of 0.05 or a maximum of 2,000 steps. All ionizable residues as well as the chain termini were modeled as charged except for histidine, which was modeled as neutral. A dielectric constant of 10 and a cutoff of 12 Å for the nonbonded energy terms was used during the minimization procedure. A dielectric constant of 10 has been found previously to produce Ramachandran plots for glycine and alanine dipeptides that are good approximations to the same Ramachandran plots calculated with the finite difference Poisson–Boltzmann FDPB method (Yang et al., 1996) and was used here as a simple approx-

imization to the effect of the solvent during the minimization. However, as is consistent with a thermodynamic cycle in which Coulomb interactions are calculated in the gas phase, a dielectric constant of 1 was used in the free energy evaluation.

Solvation energies were calculated using the FDPB method implemented in the program DelPhi (Nicholls & Honig, 1991). As in previous work using CHARMM parameters (Yang & Honig, 1995a), the protein was assigned an internal dielectric constant of 1. The solvent was assigned a dielectric constant of 80, a lattice spacing of 2.25 grids/Å was used, and a lattice dimension that ensured that a cube the size of the largest x , y , or z dimension filled 80% of the lattice.

Solvent-accessible areas were calculated using an algorithm developed by Sridharan (Sridharan et al., 1992) using atomic radii from CHARMM22 (MacKerrel et al., 1998). Area calculations were performed on models that included heavy atoms only. Polar atoms are defined to be all backbone and side-chain nitrogens and oxygens, backbone and side-chain carbonyl, and carboxyl carbons and the sulfur of cysteine. All other heavy atoms are classified as nonpolar.

Simplified PEEF

The simplified PEEF is defined as

$$\Delta E_{\text{decoy} \rightarrow \text{native}} = \Delta E_{\text{hp}} + \Delta E_{\text{coul}} \quad (2)$$

where E_{hp} is a hydrophobic term and E_{coul} is a Coulomb term. E_{hp} is obtained by counting the number of "hydrophobic contacts." Hydrophobic contacts occur when two atoms on the side chains of hydrophobic residues are within a water-molecule diameter of each other. Each such contact is arbitrarily assigned a stabilizing contribution of 1 kcal/mol. This value is chosen to make the scale of hydrophobic term of the same order as the electrostatic term. The hydrophobic residues are defined to be alanine, valine, leucine, isoleucine, phenylalanine, cysteine, methionine, proline, and tryptophan.

E_{coul} is calculated by assuming that the charge distribution can be described by a system of point charges and point dipoles. The positions of the point charges are obtained from the positions of the side chain of the ionizable residues (aspartate, glutamate, lysine, and arginine), slightly different from the procedure used by Wada and Nakamura (1981). For lysine, charges are placed at the NZ position, for arginines they are placed at the CZ position, and for aspartate and glutamate they are placed at the average position of the side-chain carboxylate atoms (CG, OD1, and OD2 for aspartate and CD, OE1, and OE2 for glutamate). Point dipoles are placed at the center of amide or carbonyl groups. Hydroxyl donor and acceptor pairs (i.e., serine, threonine, and tyrosine side chains) are excluded because their positions are not easily recovered from the positions of the heavy atoms and may contribute too much noise to the scoring function. Dipole directions for the remaining donor dipoles can be placed unambiguously. The electrostatic energy of such a system assuming a unit charge on the ionizable residues and modeling the dipoles as two opposite charges of magnitude 0.5 atomic units separated by 1 Å is given by

$$E_{\text{coul}} = 332 \sum_{i < j} \frac{q_i q_j}{r_{ij}} + 166 \sum_{i, j} \frac{q_i \mu_j \cdot \mathbf{r}_{ij}}{r_{ij}^3} \quad (3)$$

where the μ_i s are unit vectors in the direction of the dipole and all constants have been incorporated into the leading terms. If any charges interact at a distance of <3.5 Å, the distance between them was set to 3.5. The program used to calculate the simplified scoring function was created with the Troll software library of macromolecular analysis tools (D. Petrey, in prep.).

Native and decoy conformations

All structures in the EMBL Deliberately Misfolded set and CASP1 set were obtained from the Prostar web site (prostar.carb.nist.gov/Prostar). The Park and Levitt set was obtained from the "Decoys 'R Us" Web site (dd.stanford.edu). To ensure that the same protocol was used on all structures, wherever hydrogens were present they were removed and re-added with X-PLOR (Brünger, 1992). A complication arising in the use of a physical-chemical potentials is that results of the energy calculations can be difficult to interpret if the two systems being compared are not chemically identical. This situation occurred for some of the native/decoy pairs. When it was a simple matter to make the native and decoy structures chemically identical, this was done. Terminal residues that were known to be in the sequence but did not produce any density in the X-ray structure were deleted from any models that contained them. Where the difference was only in the presence of a few atoms in a residue, these atoms were rebuilt in extended conformation. In cases where extensive model building was required, such as when the model contained an internal insertion or deletion compared to the native structure, the relevant pairs of proteins were excluded from further study. These included the pair 5pad/1rei in the EMBL deliberately misfolded set crabpi_vriend, edn_biosym, edn_weber, mchpr_vihinen, ndk_abagyan, ndk_vihinen, p450_abagyan, and p450_weber from the CASP1 set. For protein NDK in the CASP1 set, the representative of the native state was taken from the PDB (PDB code 1nue) rather than from the Prostar Web site because the PDB structure appeared to be better resolved, including a C-terminal glutamate that the structure from Prostar did not.

Acknowledgments

This work was supported by grants 5-T32-GM08281, 2ROI-GM41371, and 5R37-GM30518-18 from the NIH and grant DBI-9601463 from the NSF.

References

- Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. 1983. CHARMM: A program for macromolecular energy, minimization and dynamic calculations. *J Comp Chem* 4:187–217.
- Brünger AT. 1992. *X-PLOR manual, version 3.1. A system for X-ray crystallography and NMR*. New Haven, Connecticut: Yale University Press.
- Eisenberg D, McLachlan AD. 1986. Solvation energy in protein folding and binding. *Nature* 319:199–203.
- Grimsley G, Shaw K, Fee L, Alston R, Huyghues-Despointes B, Thurlkill R, Scholtz J, Pace C. 1999. Increasing protein stability by altering long-range coulombic interactions. *Protein Sci* 8:1843–1849.
- Holm L, Sander C. 1992. Evaluation of protein models by atomic solvation preference. *J Mol Biol* 225:93–105.
- Huang ES, Subbiah S, Levitt M. 1995. Recognizing native folds by the arrangement of hydrophobic and polar residues. *J Mol Biol* 252:709–720.
- Janardhan A, Vajda S. 1998. Selecting near-native conformations in homology modeling: The role of molecular mechanics and solvation terms. *Protein Sci* 7:1772–1780.
- Lazaridis T, Karplus M. 1998. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J Mol Biol* 288:477–487.

- Lazaridis T, Karplus M. 2000. Effective energy functions for protein structure prediction. *Curr Opin Struct Biol* 10:139–145.
- MacKerel AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, 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.
- Mosimann S, Meleshko R, James MNG. 1995. A critical assessment of comparative molecular modeling of tertiary structures of proteins. *Proteins* 23:301–317.
- Nicholls A, Honig B. 1991. A rapid finite difference algorithm utilizing successive over-relaxation to solve the Poisson-Boltzmann equation. *J Comp Chem* 12:435–445.
- Novotny J, Brucoleri R, Karplus M. 1984. An analysis of incorrectly folded protein models. Implications for structure predictions. *J Mol Biol* 177:787.
- Oliva M, Moulton J. 1999. Local electrostatic optimization in proteins. *Protein Eng* 12:727–735.
- Park B, Levitt M. 1996. Energy functions that discriminate X-ray and near-native folds from well constructed decoys. *J Mol Biol* 258:367–392.
- Pickett SD, Sternberg MJE. 1993. Empirical scale of side-chain conformational entropy in protein folding. *J Mol Biol* 231:825–839.
- Sali A, Blundell TL. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 234:779–815.
- Sitkoff D, Sharp KA, Honig B. 1994. Accurate calculation of hydration free energies using macroscopic solvent models. *J Phys Chem* 98:1978–1988.
- Spassov V, Ladenstein R, Karshikoff A. 1997. Optimization of the electrostatic interactions between ionized groups and peptide dipoles in proteins. *Protein Sci* 6:1190–1196.
- Spassov VZ, Karshikoff AD, Ladenstein R. 1994. Optimization of the electrostatic interactions in proteins of different functional and folding type. *Protein Sci* 3:1556–1569.
- Sridharan S, Nicholls A, Honig B. 1992. A new vertex algorithm to calculate solvent accessible surface areas. *Biophys J* 61:A174.
- Vorobjev YN, Almagro JC, 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.
- Wada A, Nakamura H. 1981. Nature of the charge distribution in proteins. *Nature* 293:757–758.
- Yang A-S, Hitz B, Honig B. 1996. Free energy determinants of secondary structure formation: III. β -turns. *J Mol Biol* 259:873–882.
- Yang A-S, Honig B. 1995a. Free energy determinants of secondary structure formation: I. α -Helices. *J Mol Biol* 252:351–365.
- Yang A-S, Honig B. 1995b. Free energy determinants of secondary structure formation: II. Antiparallel β -sheets. *J Mol Biol* 252:366–376.