J. Mol. Biol. (1996) **256**, 623–644

JMB

Residue–Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading

Sanzo Miyazawa[1] and Robert L. Jernigan[2]*

[1]*Faculty of Technology Gunma University, Kiryu Gunma 376, Japan*

[2]*Room B-116, Building 12B Laboratory of Mathematical Biology, DBS, National Cancer Institute, National Institutes of Health, Bethesda MD 20892-5677, USA*

**Corresponding author*

Attractive inter-residue contact energies for proteins have been re-evaluated with the same assumptions and approximations used originally by us in 1985, but with a significantly larger set of protein crystal structures. An additional repulsive packing energy term, operative at higher densities to prevent overpacking, has also been estimated for all 20 amino acids as a function of the number of contacting residues, based on their observed distributions. The two terms of opposite sign are intended to be used together to provide an estimate of the overall energies of inter-residue interactions in simplified proteins without atomic details. To overcome the problem of how to utilize the many homologous proteins in the Protein Data Bank, a new scheme has been devised to assign different weights to each protein, based on similarities among amino acid sequences. A total of 1168 protein structures containing 1661 subunit sequences are actually used here. After the sequence weights have been applied, these correspond to an effective number of residue–residue contacts of 113,914, or about six times more than were used in the old analysis. Remarkably, the new attractive contact energies are nearly identical to the old ones, except for those with Leu and the rarer amino acids Trp and Met. The largest change found for Leu is surprising. The estimates of hydrophobicity from the contact energies for non-polar side-chains agree well with the experimental values. In an application of these contact energies, the sequences of 88 structurally distinct proteins in the Protein Data Bank are threaded at all possible positions without gaps into 189 different folds of proteins whose sequences differ from each other by at least 35% sequence identity. The native structures for 73 of 88 proteins, excluding 15 exceptional proteins such as membrane proteins, are all demonstrated to have the lowest alignment energies.

Keywords: hydrophobicity; contact energy; residue packing energy; sequence threading; sequence sampling weights

Introduction

The understanding of protein folding is a long-standing goal in structural biology. Although a large number of native structures of proteins are already known and more are being elucidated rapidly, usually only relatively small fluctuations near native structures have been examined in detail with molecular dynamics simulations that employ potentials with full atomic representations of proteins. Much less is known about the denatured state and the full breadth of the protein folding process. Simulating the entire protein folding process, which occurs on a time scale ranging from milliseconds to seconds, would require enormous computational power because of the high dimensional space of protein conformations and the complexity of the energy surface. The present-day capabilities of computers usually limit the time scale of molecular dynamics simulations to nanoseconds. Consequently, simplified models are required for the study of the protein folding process. Simplifications can be made to both geometry and potential functions. Here, we address the design of simplified but realistic free energy potentials that include solvent effects.

Previously we evaluated empirically a set of such

Abbreviations used: PDB, Protein Data Bank; s.d., standard deviation.

0022–2836/96/080623–22 $12.00/0

effective inter-residue contact energies for all pairs of the 20 amino acids, under the basic assumption that the average characteristics of residue–residue contacts observed in a large number of crystal structures of globular proteins represent the actual intrinsic inter-residue interactions. For this purpose, the Bethe approximation (quasi-chemical approximation) with an approximate treatment of the effects of chain connectivity was employed with the number of residue–residue close contacts observed in protein crystal structures (Miyazawa & Jernigan, 1985). This empirical energy function included solvent effects, and provided an estimate of the long-range component of conformational energies without atomic details. A comparison of these contact energies with the Nozaki–Tanford transfer energies (Nozaki & Tanford, 1971) showed a high correlation, although on average the contact energies yielded about twice the energy gain indicated by the Nozaki–Tanford transfer energies (Miyazawa & Jernigan, 1985). However, when these contact energies are applied together with a simple assumption about the compactness of the denatured state to estimate the unfolding Gibbs free energy changes for single amino acid mutants of the tryptophan synthase α subunit (Yutani *et al*., 1987) and bacteriophage T4 lysozyme (Matsumura *et al*., 1988), they yield estimates that exhibit a strong correlation with the observed values, especially for hydrophobic amino acids. Also, the calculated energy values had the same magnitudes as the observed values for both proteins. This method can also explain the wide range of unfolding Gibbs free energy changes for single amino acid replacements at various residue positions of staphylococcal nuclease (Miyazawa & Jernigan, 1994). These facts all indicate that the inter-residue contact energies properly reflect actual inter-residue interactions, including hydrophobic effects originating in solvent effects.

It is generally thought that the native conformations of proteins correspond to the structures of lowest free energy. Thus, successful potential functions, such as those based on native structures, ought to yield the lowest free energy for the native conformations. It was demonstrated that classical semi-empirical potentials such as CHARMM (Brooks *et al*., 1983) cannot always identify non-native folds of proteins (Novotny *et al*., 1984). Our contact energies were demonstrated to discriminate successfully between native-like and incorrectly folded conformations in a lattice study of five small proteins (Covell & Jernigan, 1990). Sippl (1990) evaluated the potentials of mean force as a function of distance for two-body interactions between amino acids in protein structures from the radial distribution of amino acids in known protein native structures. The potentials of mean force for the interactions between C^{β} atoms of all amino acid pairs were used to calculate the conformational energies of amino acid sequences in a number of different folds, and it was found that the conformational energy of the native state is the lowest among the alternatives (Hendlich *et al*., 1990; Sippl & Weitckus, 1992; Jones *et al*., 1992). Pairwise contact potentials depending on inter-residue distance were also estimated by Bryant & Lawrence (1993).

Table 1. Summary of protein structures used in the present analysis

Number of protein structures[a]	1168
Number of protein subunit structures	1661
Number of protein families[b]	424
Effective number of proteins ($\Sigma_i w_i$)	251

[a] Structures whose resolutions were higher than 2.5 Å and which were determined by X-ray analyses and are larger than 50 residues. See text for details.

[b] A set of proteins with less than 95% sequence identity between any pair.

It should be noted here that a two-body residue–residue potential of mean force based on the radial distribution of residues will manifest peaks and valleys as a function of distance, even for hard spheres, which are effects of close residue packing. However, these would not be present in actual interaction potentials. That is, such a potential of mean force reflects not only the actual inter-residue interactions, but also includes the average effects of other residues upon the target residue pair, including those interposed between the target pair and especially the significant effects of residue packing in protein structures. There will be an over-counting if the sum of the potential is taken over all residue pairs. Thus, if the residue–residue potential in a protein is approximated by such a potential of mean force, the sum of the potential over all residue pairs is unlikely to yield the correct value for the total residue–residue interaction energy. In addition, even though these effective potentials have the important characteristics of low values for the native folds of proteins, they are unlikely to succeed in representing the actual potential surface far from the native conformation. Therefore, such potentials of mean force may not be appropriate for application in a study of a wide range of conformations, from the denatured state to the native conformation.

On the other hand, a direct account of the requirement that the native state be lowest in energy was taken by Crippen (1991), and Maiorov & Crippen (1992), who tried to fit empirically a feasible set of parameters, which corresponded to contact energies between amino acid groups separated at certain ranges of distance, in such a way that the total contact energies of native conformations were lower than other alternatives. As with all of these potentials, it is unknown how well this potential represents the actual potential surface far from the native conformation.

Lüthy *et al*. (1992) developed an empirical method to evaluate the correctness of protein models. Pseudo-potentials have been devised to find out which amino acid sequences fold appropriately into a known three-dimensional structure (Bowie *et al*.,

Table 2. Number of contacts: upper triangle[a] for random mixing and lower triangle[b] for actual counts in the protein sample

	N_i	SLV[c]	Cys	Met	Phe	Ile	Leu	Val	Trp	Tyr	Ala	Gly	Thr	Ser	Asn	Gln	Asp	Glu	His	Arg	Lys	Pro	
		314,460	1089	528	1723	3429	7813	5970	299	1780	10,219	9160	4894	6798	2888	1760	4854	4100	730	2492	5123	2273	C'_{ii}
		24,518	21,124	41,637	57,536	89,945	76,880	15,672	40,772	99,089	95,305	70,556	80,459	52,629	41,652	70,989	64,324	24,806	47,888	72,052	47,860	SLV[c]
SLV[c]	12,785,07	4,010,169	2128	1178	2496	3624	5388	4521	1033	2670	4410	4395	3032	3469	2184	1630	2476	2060	1191	1965	2160	2065	Cys
Cys	10,318	12,373	5003	946	2961	4161	6581	5041	964	2416	5023	4015	3020	3046	1974	1600	2638	2410	1221	2089	2195	1940	Met
Met	10,354	7172	1247	1125	3566	8302	13,335	10,531	2170	5294	10,178	8516	6590	6625	4277	3405	5653	4844	2593	4140	4508	4109	Phe
Phe	21,163	6577	2573	4546	4778	6620	18,622	14,980	2820	7004	14,508	12,158	9153	9075	5808	4644	7478	6746	3424	5795	6186	5678	Ile
Ile	29,069	12,219	3830	5188	10,183	8432	16,070	23,502	4540	10,854	23,146	18,339	14,102	14,095	8989	7655	11,584	10,788	5514	9456	9895	8882	Leu
Leu	44,483	13,346	5395	8218	17,228	26,396	21,432	10,497	3612	8844	19,229	15,857	11,786	11,964	7356	5923	9312	8345	4422	7147	7867	7256	Val
Val	38,207	26,678	4768	5343	12,746	19,830	32,623	14,091	525	1998	3503	3064	2393	2452	1541	1236	1880	1640	940	1524	1499	1545	Trp
Trp	8160	5866	1168	1393	2498	3101	4810	3723	519	2752	8598	7554	5875	5917	3874	3001	4811	4030	2221	3662	3759	3654	Tyr
Tyr	20,365	21,398	2045	2765	5549	6770	10,806	7777	1975	2491	10,471	15,644	11,298	11,446	7212	5794	9222	8103	4289	7077	7612	6923	Ala
Ala	47,643	97,613	4306	5014	10,516	16,465	24,837	21,733	3453	7942	11,789	7237	9785	9962	6133	4860	7736	6570	3640	5864	6138	5981	Gly
Gly	46,381	120,773	4168	3376	6523	9071	13,533	13,584	2946	7235	17,079	10,881	4041	7551	4679	3757	5899	5078	2734	4509	4724	4511	Thr
Thr	32,587	82,795	2475	2595	5281	7958	11,387	9906	1708	5102	11,323	11,312	4387	4266	4743	3747	5854	4957	2767	4479	4692	4639	Ser
Ser	36,710	111,609	3012	2205	5567	6443	10,084	9059	1889	5537	11,447	12,019	9465	5458	1766	2396	3791	3232	1784	2853	3126	2873	Asn
Asn	24,236	76,966	1549	1410	2940	3331	5751	4792	1381	3792	6465	7319	5790	5920	2605	1148	2989	2619	1375	2404	2420	2306	Gln
Gln	19,199	58,151	1375	1490	3056	3772	6091	4541	1057	3350	5037	4918	4412	3957	3368	1253	2721	4301	2255	3683	3992	3648	Asp
Asp	32,749	107,790	1584	1389	3217	4318	5729	4606	1444	5105	7836	8632	7362	8327	6075	3550	2951	2154	2004	3459	3678	3229	Glu
Glu	30,640	98,646	1189	1827	3114	4498	6706	5241	1521	4601	5933	5248	6172	6343	4550	3096	4001	1884	762	1820	1868	1817	His
His	11,907	24,908	1209	1273	2581	2525	4133	3215	1091	2441	3600	3425	3071	3073	2053	1332	3603	2800	1224	1766	2946	2889	Arg
Arg	22,741	62,394	1120	1463	2956	4065	6653	4867	1588	4433	5205	5656	4720	4790	3325	3062	8981	8570	1883	1558	1972	2913	Lys
Lys	32,058	124,460	1200	1456	3098	4199	6531	5437	1367	4678	6019	6054	4940	5117	4260	3304	9501	10,234	1554	1803	1354	1624	Pro
Pro	24,594	63,958	1983	1916	3930	4600	7632	6518	2253	4653	6368	6827	4806	4843	3077	2779	3483	2989	2097	3036	2665	1824	
	N_i	SLV[c]	Cys	Met	Phe	Ile	Leu	Val	Trp	Tyr	Ala	Gly	Thr	Ser	Asn	Gln	Asp	Glu	His	Arg	Lys	Pro	

Totals are: N_r = 54,356.2, N_{rr} = 113,913.5, $2N_{r0}$ = 113,569.1.

[a] Scaling factors are $C'_{ii} \times 10$, $C'_{i0} \times 20$, $C_{ii} \times 10$, and $C_{ij} \times 20$.

[b] Scaling factors are $N_i \times 10$, $N_{ii} \times 10$, $N_{ij} \times 20$.

[c] SLV, effective solvent molecules.

Table 3. Contact energies in RT units; e_{ij} for upper half and diagonal and e'_{ij} for lower half

		Cys	Met	Phe	Ile	Leu	Val	Trp	Tyr	Ala	Gly	Thr	Ser	Asn	Gln	Asp	Glu	His	Arg	Lys	Pro	
	Cys	–5.44	–4.99	–5.80	–5.50	–5.83	–4.96	–4.95	–4.16	–3.57	–3.16	–3.11	–2.86	–2.59	–2.85	–2.41	–2.27	–3.60	–2.57	–1.95	–3.07	Cys
	Met	0.46	–5.46	–6.56	–6.02	–6.41	–5.32	–5.55	–4.91	–3.94	–3.39	–3.51	–3.03	–2.95	–3.30	–2.57	–2.89	–3.98	–3.12	–2.48	–3.45	Met
	Phe	0.54	–0.20	–7.26	–6.84	–7.28	–6.29	–6.16	–5.66	–4.81	–4.13	–4.28	–4.02	–3.75	–4.10	–3.48	–3.56	–4.77	–3.98	–3.36	–4.25	Phe
	Ile	0.49	–0.01	0.06	–6.54	–7.04	–6.05	–5.78	–5.25	–4.58	–3.78	–4.03	–3.52	–3.24	–3.67	–3.17	–3.27	–4.14	–3.63	–3.01	–3.76	Ile
	Leu	0.57	0.01	0.03	–0.08	–7.37	–6.48	–6.14	–5.67	–4.91	–4.16	–4.34	–3.92	–3.74	–4.04	–3.40	–3.59	–4.54	–4.03	–3.37	–4.20	Leu
	Val	0.52	0.18	0.10	–0.01	–0.04	–5.52	–5.18	–4.62	–4.04	–3.38	–3.46	–3.05	–2.83	–3.07	–2.48	–2.67	–3.58	–3.07	–2.49	–3.32	Val
	Trp	0.30	–0.29	0.00	0.02	0.08	0.11	–5.06	–4.66	–3.82	–3.42	–3.22	–2.99	–3.07	–3.11	–2.84	–2.99	–3.98	–3.41	–2.69	–3.73	Trp
	Tyr	0.64	–0.10	0.05	0.11	0.10	0.23	–0.04	–4.17	–3.36	–3.01	–3.01	–2.78	–2.76	–2.97	–2.76	–2.79	–3.52	–3.16	–2.60	–3.19	Tyr
	Ala	0.51	0.15	0.17	0.05	0.13	0.08	0.07	0.09	–2.72	–2.31	–2.32	–2.01	–1.84	–1.89	–1.70	–1.51	–2.41	–1.83	–1.31	–2.03	Ala
	Gly	0.68	0.46	0.62	0.62	0.65	0.51	0.24	0.20	0.18	–2.24	–2.08	–1.82	–1.74	–1.66	–1.59	–1.22	–2.15	–1.72	–1.15	–1.87	Gly
	Thr	0.67	0.28	0.41	0.30	0.40	0.36	0.37	0.13	0.10	0.10	–2.12	–1.96	–1.88	–1.90	–1.80	–1.74	–2.42	–1.90	–1.31	–1.90	Thr
	Ser	0.69	0.53	0.44	0.59	0.60	0.55	0.38	0.14	0.18	0.14	–0.06	–1.67	–1.58	–1.49	–1.63	–1.48	–2.11	–1.62	–1.05	–1.57	Ser
	Asn	0.97	0.62	0.72	0.87	0.79	0.77	0.30	0.17	0.36	0.22	0.02	0.10	–1.68	–1.71	–1.68	–1.51	–2.08	–1.64	–1.21	–1.53	Asn
	Gln	0.64	0.20	0.30	0.37	0.42	0.46	0.19	–0.12	0.24	0.24	–0.08	0.11	–0.10	–1.54	–1.46	–1.42	–1.98	–1.80	–1.29	–1.73	Gln
	Asp	0.91	0.77	0.75	0.71	0.89	0.89	0.30	–0.07	0.26	0.13	–0.14	–0.19	–0.24	–0.09	–1.21	–1.02	–2.32	–2.29	–1.68	–1.33	Asp
	Glu	0.91	0.30	0.52	0.46	0.55	0.55	0.00	–0.25	0.30	0.36	–0.22	–0.19	–0.21	–0.19	0.05	–0.91	–2.15	–2.27	–1.80	–1.26	Glu
	His	0.65	0.28	0.39	0.66	0.67	0.70	0.08	0.09	0.47	0.50	0.16	0.26	0.29	0.31	–0.19	–0.16	–3.05	–2.16	–1.35	–2.25	His
	Arg	0.93	0.38	0.42	0.41	0.43	0.47	–0.11	–0.30	0.30	0.18	–0.07	–0.01	–0.02	–0.26	–0.91	–1.04	0.14	–1.55	–0.59	–1.70	Arg
	Lys	0.83	0.31	0.33	0.32	0.37	0.33	–0.10	–0.46	0.11	0.03	–0.19	–0.15	–0.30	–0.46	–1.01	–1.28	0.23	0.24	–0.12	–0.97	Lys
	Pro	0.53	0.16	0.25	0.39	0.35	0.31	–0.33	–0.23	0.20	0.13	0.04	0.14	0.18	–0.08	0.14	0.07	0.15	–0.05	–0.04	–1.75	Pro
$e_{rr} - 2.55$	e_{ir}	–3.57	–3.92	–4.76	–4.42	–4.81	–3.89	–3.81	–3.41	–2.57	–2.19	–2.29	–1.98	–1.92	–2.00	–1.84	–1.79	–2.56	–2.11	–1.52	–2.09	
$e_r - 3.60$	e_i	–4.29	–4.73	–5.57	–5.29	–5.71	–4.72	–4.41	–3.87	–3.17	–2.53	–2.63	–2.27	–2.14	–2.35	–2.02	–2.07	–2.94	–2.43	–1.82	–2.53	
$f_r - 3.60$	f_i	–5.58	–6.14	–7.39	–7.09	–7.88	–6.15	–5.34	–4.60	–3.24	–2.22	–2.48	–1.92	–1.74	–1.93	–1.54	–1.49	–2.91	–2.07	–1.17	–1.97	
N_{ir} / N_i	2.096	2.723	2.722	2.780	2.811	2.893	2.728	2.537	2.493	2.143	1.840	1.973	1.771	1.699	1.720	1.598	1.508	2.075	1.787	1.343	1.629	
q_i 7.162	6.281	6.646	6.137	5.870	6.042	6.087	6.155	5.793	6.037	6.334	6.284	6.486	6.582	6.574	6.469	6.487	6.235	6.241	6.318	6.569	5.858	

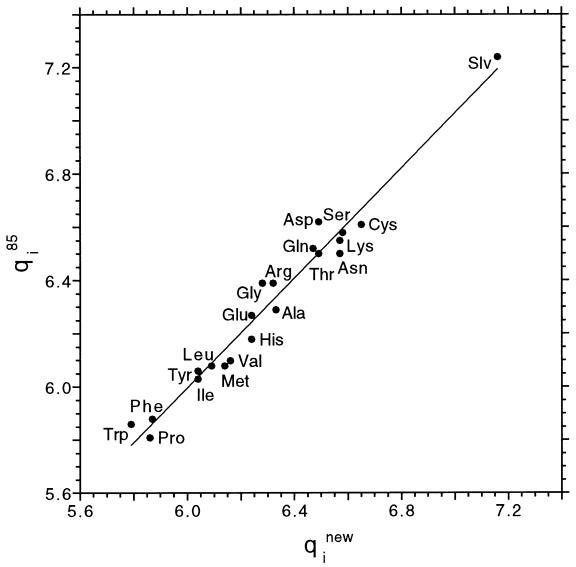


Figure 1. A comparison of the coordination number, q_i, between the previous analysis (1985) and this work for each type of amino acid. Slv denotes solvent. The ordinate indicates the values previously reported by Miyazawa & Jernigan (1985) and the abscissa the values obtained in this work. A continuous line shows the regression line that is $y = -0.19 + 1.03x$; the correlation coefficient is 0.98.

1991; Nishikawa & Matsuo, 1993). The pseudo-potential devised by Nishikawa & Matsuo (1993) was composed of four terms, side-chain packing, hydration, hydrogen bonding and local conformational potentials, and was empirically derived from the statistical features observed in 101 known protein structures. A slightly modified form of the Sippl potential was used to take account of the effect of side-chain packing in proteins. All other terms were also evaluated as potentials of mean force. This function was also demonstrated to be an appropriate measure of the compatibility between sequences and structures of proteins. They share a number of common characteristics with empirical energy potentials, but they are not designed to be used explicitly as an empirical energy function. In the case of Nishikawa & Matsuo (1993), each of the four terms are summed with weights in the total energy.

On the other hand, the present two-body contact energies are estimated with the Bethe approximation with the basic assumption of regarding residue–residue contacts in protein structures to be the same as those in mixtures of unconnected amino acids and solvent molecules. The Bethe approximation is a well-known second-order approximation to the mean field approximation used to describe a system consisting of a mixture of multiple molecular species interacting with each other (Hill, 1960). Both approximations are usually used to calculate a partition function for such a system from a given set of interaction energies between molecules. In the mean field approximation, contacts between species would be approximated to be random, and a partition function of the system would be developed. In the Bethe approximation, the effects of interactions are taken into account to estimate the average numbers of contacts. Thus, the Bethe approximation is the lowest order approximation to be able to provide an estimate of a set of contact energies between species from a given set of the average numbers of contacts between them.

Therefore, if residue–residue contacts in protein structures can be reliably represented to be the same as those in mixtures of unconnected amino acids and solvent molecules, the Bethe approximation will give us a reasonable estimate of actual interaction energies between amino acids. Of course, it must be examined whether or not this basic approximation is appropriate to describe the contacts in real protein structures. Also, a limitation, both of this method and of methods using a potential of mean force to evaluate inter-residue interactions, lies in the fact that the effects of specific amino acid sequences on the statistical distribution of inter-residue distances are completely neglected, although the characteristics of the protein being a chain are included as a mean field.

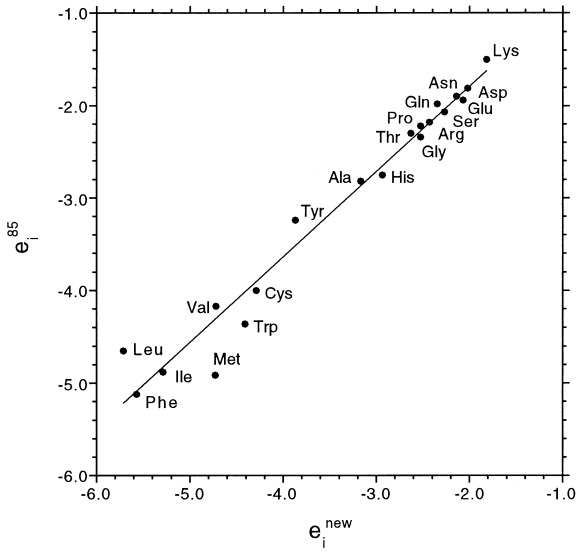
Here, the effective inter-residue contact energies for all amino acid pairs are re-evaluated using the same method as before, but with significantly more protein structures than before. In the original work, only 42 globular proteins were used to calculate contact frequencies between amino acids. Since 1985, many additional protein structures have been reported, and now more than 1000 protein structures are available. However, one complication arises from the fact that there are many homologous proteins in the Protein Data Bank (PDB; Bernstein *et al*., 1977). For example, the structures of many single amino acid mutants of T4 lysozyme are included in the PDB. Therefore, to use all of this structural data, an unbiased sampling of protein structures from the PDB is required in the calculation of the contact frequencies. Here, a sampling weight for each protein has been devised based on a sequence homology matrix giving the extent of sequence identity of all pairs of sequences.

In the estimation of the attractive contact energies, the Bethe lattice is used for conformational space. However, an additional repulsive potential based on many-body residue packing is needed to properly estimate the long range energies of protein conformations. Here, the repulsive energies that result from tight packing of residues are evaluated as a function of the numbers of contacting residues based on their distributions in known protein structures.

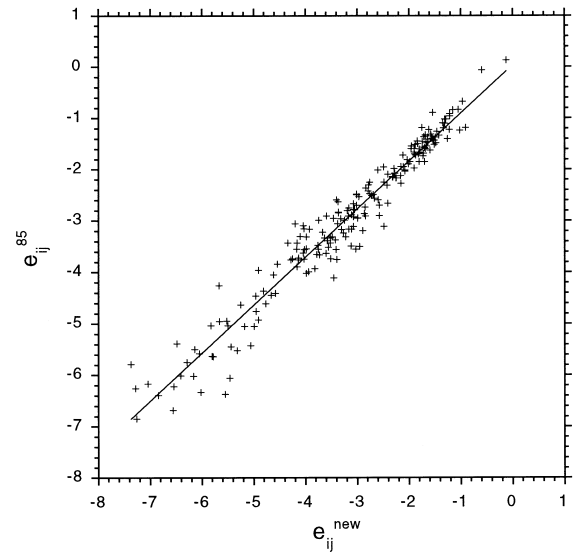
Results

Sample weighting

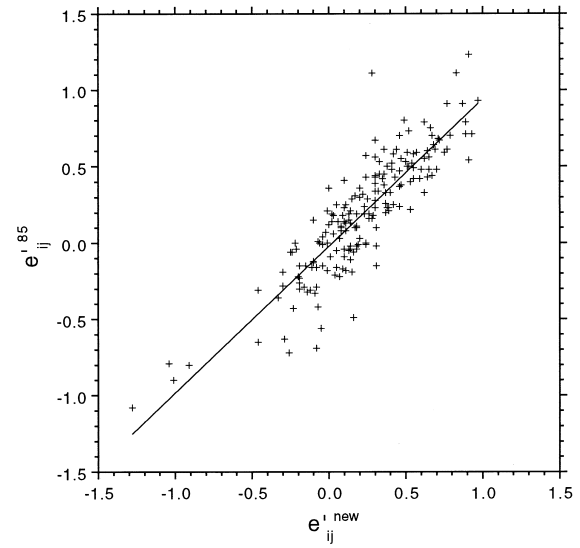
According to the procedure described in Methods, 1168 protein structures, whose structures have been analyzed by X-ray and whose resolution has been reported to be better than 2.5 Å, are chosen, and then each of the 1661 sequences in



A



B



C

Figure 2(A–C)

those structures is sampled with a weight determined as described below on the basis of the sequence identity matrix. Sequences are aligned in a conventional pairwise manner with a global alignment method (Needleman & Wunsch, 1970), using the log odds matrix for 250 PAM (Dayhoff *et al*., 1978) as a scoring matrix for amino acid similarity. Penalty for a gap (of k residues) is taken to be $12 + 4(k - 1)$ with the cut-off value of 48, but no penalty is used for terminal gaps. Sequence identities are then calculated for the aligned pairs of sequences. In the original work, the numbers of contacts in protein structures were counted in complete assembly. In the present calculation, only the coordinates of subunits explicitly given in the PDB files are used. As listed in Table 1, the effective number of proteins is 251. The effective number of residues is 54,356, and the total effective number of contacts is 113,914 in the present analysis; by contrast, these values were 9040 and 18,192 in the original analysis, respectively. The total effective number of contacts is 6.3 times the previous data. Thus, on average, the standard deviations in the numbers of contacts could be expected to be reduced by a factor of 1/2.5.

Estimates of contact energies

The effective numbers of contacts observed in the protein structures are listed in the lower triangle of Table 2. The entries in Table 2 have been multiplied by ten for diagonal elements and by 20 for off-diagonal elements. The numbers in the upper triangle of Table 2 correspond to the correction factors, C and C', for the estimation of contact energies; see equations (29), (30), (36) and (37) in this paper, and also equations (10) to (15) of Miyazawa & Jernigan (1985). The coordination number for each residue type has been estimated in the same way with equation (33) of Miyazawa & Jernigan (1985) from the volume of each type of residue at the center and the average volume of its surrounding residues. These newly estimated coordination numbers are listed in the last row of Table 3 and are very similar to the previous

Figure 2. A, A comparison of the average contact energy, e_i, in RT units for each type of amino acid between the previous analysis (1985) and this work. The ordinate values are those reported by Miyazawa & Jernigan (1985) and the abscissa the values obtained in the present analysis. The continuous line is the regression line, $y = 0.06 + 0.92x$; the correlation coefficient is 0.98. B, A comparison of the contact energy, e_{ij}, in RT units of each type of contact between the previous analysis (1985) and this work. A continuous line shows the regression line that is $y = 0.03 + 0.93x$; the correlation coefficient is 0.97. C, A comparison of the energy, e'_{ij}, in RT units, which is the energy difference accompanying the formation of a contact pair i–j from contact pairs i–i and j–j, between the previous analysis (1985) and this work. The regression line is $y = -0.02 + 0.96x$, and the correlation coefficient is 0.88.

estimates (1985) as shown in Figure 1; the regression line is $y = -0.19 + 1.03x$ and the correlation coefficient is 0.98.

Effective contact energies are estimated from these numbers and are listed in dimensionless units, units of RT, in Table 3. Figure 2 shows a comparison between the new contact energies and those from the previous analysis. The average contact energies, e_i, for each type of residue are compared in Figure 2A, and each of the contact energies, e_{ij}, is compared in Figure 2B. The ordinate values are those reported by Miyazawa & Jernigan (1985) and the abscissa values those obtained in the present work. Correlations between both estimates are high; the correlation coefficient is 0.98 for e_i and 0.97 for e_{ij}. The difference between these two estimates tends to be larger for hydrophobic amino acids than for hydrophilic ones. On average, the present estimates of contact energies are slightly more negative than the previous estimates, perhaps reflecting large structures present in the new data; the regression line is $y = 0.06 + 0.92x$ for e_i, and $y = 0.03 + 0.93x$ for e_{ij}. The present estimate of the mean contact energy, e_r, is equal to –3.6, which is more negative than the old estimate, –3.2.

Also, the differences between individual new and old contact energies tend to be large for infrequent amino acids such as Trp and Met. This might be expected. However, there is one large difference for a common amino acid, Leu. The contact energy of any pair with Leu is significantly more negative in the present analysis. All of the top ten contact pairs with large differences involve pairs with Leu. The difference in the contact energy for the Leu–Leu pair is the largest among them, and the present estimate, –7.37, is more negative than the previous one, –5.79. If the coordination number for Leu were estimated to be smaller in the present work than previously, then the contact energy for any pair with Leu would be estimated to be more negative. However, the two estimates of the coordination number of Leu are almost identical (see Table 2 and Figure 1). The comparison of the distribution of the number of contacts for Leu clearly shows that the present distribution is shifted toward more contacts than the previous one. Therefore, there are actually more contacts with Leu in the present data than before, but the reason is unclear.

As shown in equation (27), the estimation of contact energies, e_{ij}, requires the estimation of n_{00}, which is not so accurate. On the other hand, relative differences between contact energies do not depend on such a parameter; see equation (28). As already stated, the absolute values of the contact energies might be expected to be less reliable than their relative differences. The required scaling factor for absolute energy specification could be adjusted by making e_{rr} consistent with experimental estimates of the average attractive energy between residues.

Figure 2C shows the comparison of the values of e'_{ij} between the two estimates. The estimation of e'_{ij}, which is the energy difference accompanying the formation of the two contact pairs i–j from the contact pairs i–i and j–j, does not require estimation of n_{00}. Therefore, their estimates might be expected to be more accurate than the absolute values of the contact energies, e_{ij}. The regression line in Figure 2C almost passes through the origin with unit slope; it is $y = -0.02 + 0.96x$. Even though the correlation between the two estimates is clearly not as good as for the contact energies e_{ij}, its value is still quite good at 0.88.

The important qualitative characteristics of the contact energies observed in the previous analysis hold in the present results; the values of e'_{ij} clearly show: (1) a relatively large energy loss accompanying the formation of Cys–X contacts from Cys–Cys and X–X contacts, probably reflecting the loss of disulfide bonds; (2) the large favorable electrostatic interactions between positively charged (Lys, Arg) and negatively charged (Glu, Asp) amino acids; and (3) the segregation of hydrophobic and hydrophilic residues.

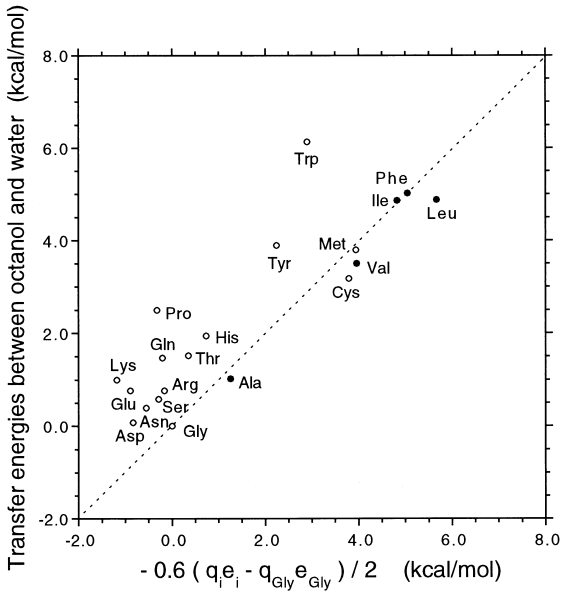


Figure 3. Transfer free energies of amino acids relative to glycine between octanol and water corrected by Sharp *et al*. (1991) and the corresponding values of $-0.6q_ie_i/2$ in this work. $-RTq_ie_i/2$ corresponds to the average contact energy gain of an i type residue completely surrounded by other residues in the protein crystal structures. $RT = 0.6$ kcal/mol has been employed to transform the dimensionless contact energies into kcal/mol units. The modified transfer free energies of 20 *N*-acetyl amino acid amides between octanol and water are taken from Table III of Sharp *et al*. (1991); these values include volume corrections for solute–solvent size differences. The line of unit slope is shown by a dotted line. Filled circles show the values for non-polar side-chains (Ala, Val, Ile, Leu, Phe) and open circles for other amino acids; Gly is located at the origin. The regression line, not shown here, is $y = -0.14 + 0.96x$, and the correlation coefficient is 0.98 for non-polar side-chains. Although polar amino acids are given in this Figure, the comparison is strictly meaningful only for non-polar residues.

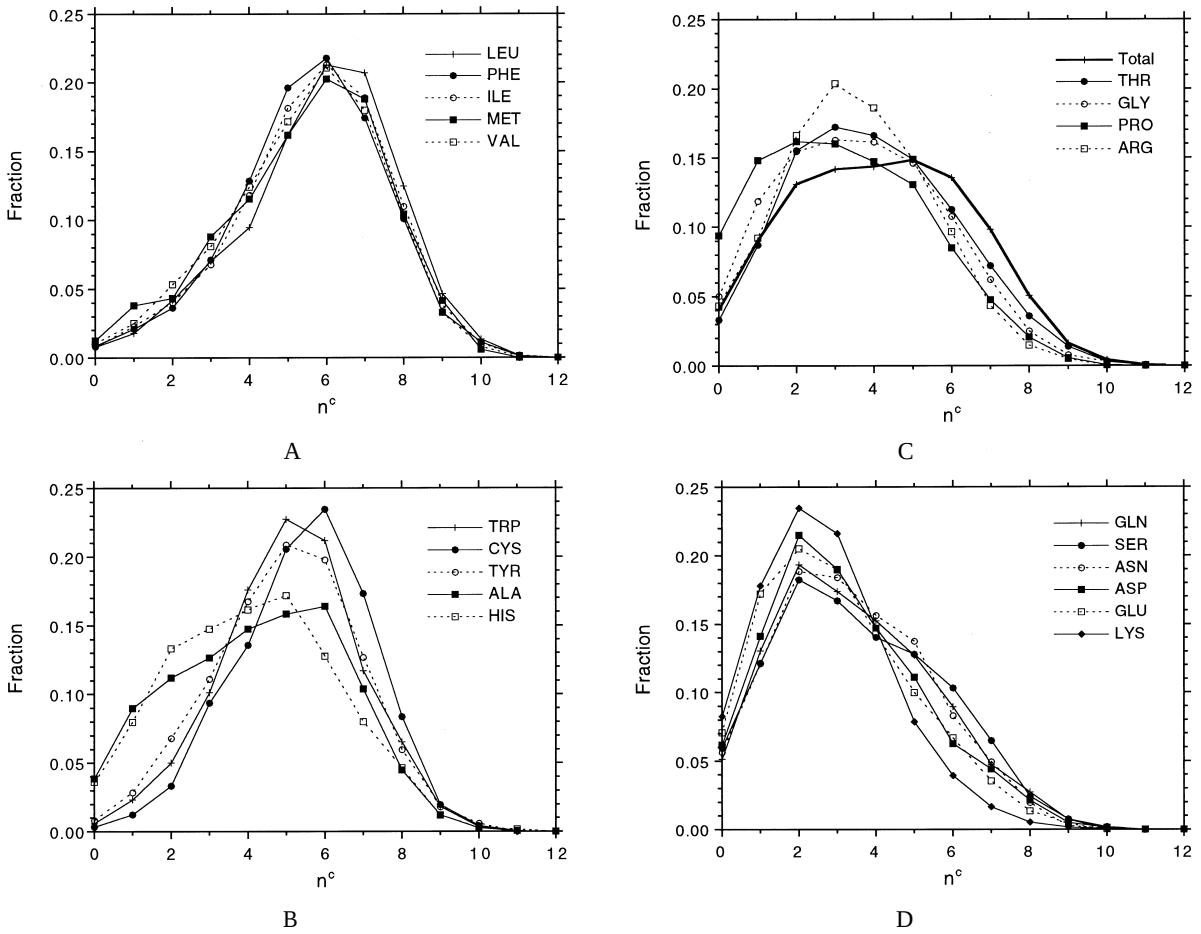


Figure 4. The frequency distribution of the number of contacts for each type of amino acid. The distribution for each type of amino acid is shown by order of increasing values of e_{ir}, which measures the hydrophobicity of the amino acid, in A to D. Total in C shows the frequency distribution of the number of contacts, irrespective of amino acid type.

Comparison with experimental transfer energies

Figure 3 shows the comparison of the contact energies with the transfer free energies, taken from Table III of Sharp *et al*. (1991), for amino acids relative to glycine between octanol and water. The abscissae show values of $-0.6q_ie_i/2$ that correspond to the average contact energy gain of an *i* type of residue completely surrounded by other residues in protein crystal structures. For comparison, a value of $RT = 0.6$ kcal/mol has been employed in this Figure to express the dimensionless contact energies in kcal/mol units. Sharp *et al*. (1991) re-evaluated the transfer free energies of 20 amino acids between octanol and water taken from Fauchère & Pliška (1983) by including volume corrections for solute–solvent size differences. The filled circles show the values for non-polar side-chains, Ala, Val, Ile, Leu and Phe, and the open circles for other amino acids; Gly is located at the origin. The plots for the most hydrophobic side-chains are located nearly directly on the dotted line of the unit slope, i.e. the two values are nearly identical. Their coincidence strongly supports the present estimates of long-range interactions between side-chains. Although polar amino acids are also plotted in this Figure, the comparison is really meaningful only for non-polar residues, because organic solvents cannot represent circumstances surrounding polar residues in native protein structures; e_i for polar residues includes not only hydrophobic energies but also the average of other interaction energies with surrounding residues and water in proteins, such as hydrogen bonds and electrostatic energies. Thus, the extent of agreement is, overall, better than might be expected.

Distributions of the number of residue contacts

The distribution of the number of contacts for each of the 20 types of amino acids in protein crystals is shown in order of increasing values of e_{ir}, which measure the hydrophobicity of amino acids, in Figure 4A to 4D. Amino acids whose values of e_{ir} are similar will show similar frequency distributions, although the coordination numbers must be taken into account. For example, the distribution for Cys is shifted somewhat toward more contacts than that for Trp, even though $e_{\mathrm{cys},r}$ for Cys is more positive than that for Trp, because the coordination number for Cys is significantly larger than that for

Trp. The distributions for non-polar amino acids have single peaks around $n^c = 6$ and those for the most polar amino acids near $n^c = 2$. These numbers of contacts are typical for residues buried completely inside of proteins or for residues exposed on the surfaces of proteins. This indicates that the system consists nearly of two phases, buried residues and surface residues. The distribution for Ser clearly shows the presence of a shoulder near $n^c = 6$, as well as a single peak at $n^c = 2$. For such residue types that are ambivalent in character, the distributions can readily be decomposed into two peaks with maxima near 2 and 6.

Table 4 shows only the high density portion of the distribution of the number of contacts for each of the 20 types of amino acid in protein crystals. Each number in the table is an effective number because it is the sum of the numbers of contacts weighted by the sampling weight of equation (50) to remove any sampling bias from homologous protein sequences included in the protein sample. Repulsive packing energies have been estimated directly from the values in this table by using equation (43).

Average residue–residue energies

Replacing $e_{i_p j}$ in equation (19) by the average contact energy e_{i_p} of the i_p type of amino acid, the average contact energy for the pth residue is represented by:

$$\langle E_p^c \rangle \equiv \frac{1}{2} e_{i_p} n_p^c \tag{1}$$

and depends linearly on the number of contacts with the pth residue, $n_{p.}^c$ Figure 5 shows the dependence of the long range interaction energy on the number of contacts for each type of amino acid. The ordinate corresponds to the sum of the average attractive contact energy (equation (1)) and the repulsive packing energy (equation (43)). In the repulsive region shown in this Figure, the repulsive packing energy, the second term in equation (43), does not depend so much on the type of amino acid. This is expected, because repulsive packing energies should reflect only packing density and not depend strongly on the type of amino acid at the center; likewise, the coordination number of the amino acid does not depend much on the type of amino acid.

Total energies of individual proteins

As described in Methods, the long-range energy defined by equation (13) has been calculated for a set of 189 representatives of protein structures in the PDB, which differ from each other by at least 35% sequence identity, but do not include too many unknown atomic coordinates, which were selected by Orengo *et al*. (1993).

The numbers of residues in contact with each residue along a chain are calculated by counting residues within 6.5 Å of each residue according to equation (14): the contact energies between these residue pairs are then evaluated from Table 3 and summed up according to equation (19). Also repulsive energies are estimated for the residue according to equation (40). The hard core repulsion defined by equation (41) is not included here with $e^{hc} = 0$, because known protein structures were usually refined to remove such close contacts, and such close contacts are easily removed by structure refinement. The repulsive packing energy is estimated according to equation (43), if the number of contacts at the pth residue is above the threshold value q_{i_p}. Then, those contact energies and repulsive energies are summed over all residues in a protein to estimate the total long-range energy.

As discussed in the previous work, the total contact energies are expected to consist of two terms, a term that is proportional to the chain length and another term that is proportional to the surface area of the proteins, that is, the two-thirds power of the chain length for proteins whose shapes are similar to each other:

$$\sum_p E_p^c = \sum_{i(\neq 0)} \sum_{j(\neq 0)} e_{ij} n_{ij} \tag{2}$$

$$= e_r n_{rr} \tag{3}$$

$$\simeq e_v \sum_{i(\neq 0)} \frac{1}{2} q_i n_i - e_s n_{r0} \tag{4}$$

where

$$e_v = \left(\sum_{i(\neq 0)} e_i q_i N_i \right) \Big/ \left(\sum_{i(\neq 0)} q_i N_i \right) \tag{5}$$

$$= -3.26$$

$$e_s = \left(\sum_{i(\neq 0)} e_i N_{i0} \right) / N_{r0} \tag{6}$$

$$= -2.57$$

In Figure 6A, the long-range energies per residue calculated with equation (13) are plotted in dimensionless units against the inverse of the one-third power of their chain lengths. Monomeric proteins, whose shapes could be similar to each other, are shown by filled circles, and the regression line for them is shown by a continuous line that is given by $E^{long}/n_r = -8.5 + 10.1 n_r^{-1/3}$. The correlation coefficient is 0.67. The value of the intersect of the regression line is more positive than expected from equation (4), probably because of repulsive packing energies included in the total long-range energies of the proteins.

Alignment energy for residues within a protein structure

In the present contact energies, any contact has a favorable contribution to protein stability, even if it is between polar residues, because contact energies between residues are all negative. Therefore,

Table 4. The high density portion in the distribution of the number of contacts for each amino acid

n^c	Total	Cys	Met	Phe	Ile	Leu	Val	Trp	Tyr	Ala	Gly	Thr	Ser	Asn	Gln	Asp	Glu	His	Arg	Lys	Pro
5	8057.8	212.1	167.2	415.4	527.5	720.4	654.2	185.6	425.0	754.2	677.2	485.1	470.1	333.5	243.9	364.2	306.2	204.6	338.7	251.8	320.8
6	7372.4	242.1	209.7	461.4	619.4	947.3	804.8	172.7	402.6	780.6	499.2	366.8	378.5	201.1	172.1	204.5	204.6	152.0	219.3	125.4	208.4
7	5334.9	178.7	194.4	368.8	549.7	921.2	686.4	95.5	258.0	494.4	288.0	235.1	237.9	119.9	90.8	143.9	108.4	95.4	98.5	53.1	116.8
8	2747.7	86.4	107.9	213.0	318.6	553.6	392.8	53.3	121.2	213.0	115.8	116.1	88.5	47.8	52.0	70.0	40.7	55.4	33.8	17.3	50.4
9	883.3	20.3	42.7	69.2	110.0	205.8	124.7	14.9	36.4	57.8	35.9	45.2	27.3	6.5	13.7	16.2	12.4	14.2	11.4	4.9	13.8
10	226.9	4.5	6.1	23.7	28.8	59.9	30.0	2.9	11.8	11.6	16.0	9.5	7.1	1.9	1.1	4.9	0.7	2.9	0.6	1.7	1.1
11	28.4	0.5	0.0	2.5	4.9	5.6	4.7	0.0	0.8	2.8	1.9	1.8	0.1	0.0	0.0	0.1	0.0	2.0	0.0	0.0	0.8
12	0.8	0.0	0.0	0.2	0.0	0.0	0.2	0.0	0.0	0.2	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
13	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
q_i	6.28	6.65	6.14	5.87	6.04	6.09	6.16	5.79	6.04	6.33	6.28	6.49	6.58	6.57	6.47	6.49	6.24	6.24	6.32	6.57	5.86

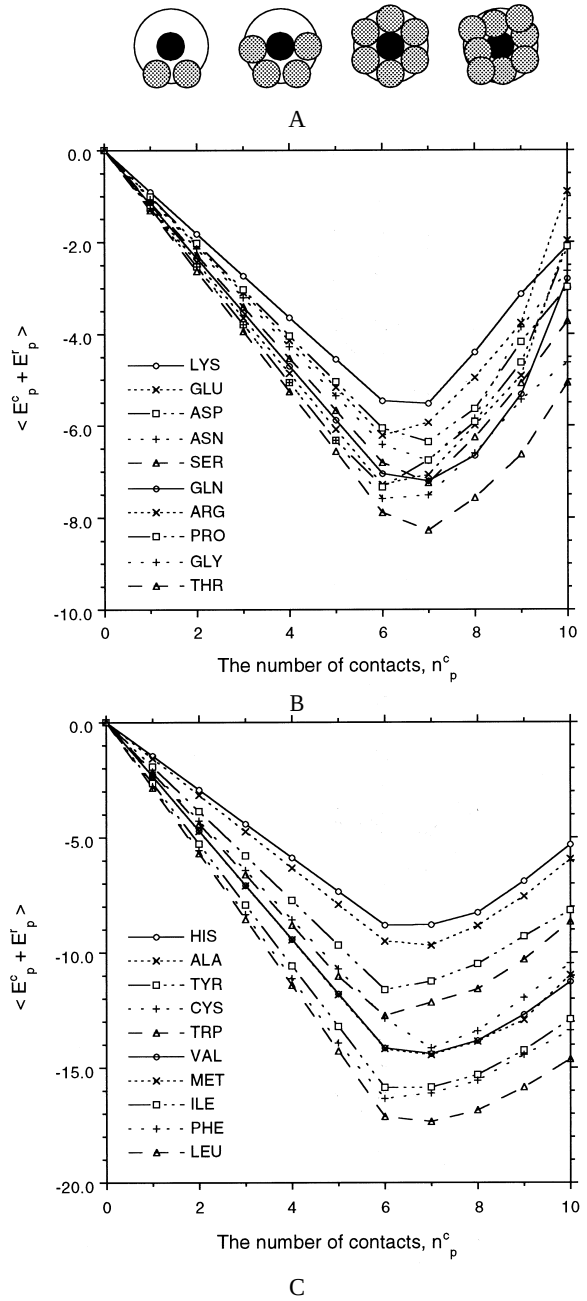


Figure 5. The dependence of the average long-range energy on the number of contacts for each residue type. A, Ranges of packing density as reflected in n_p^c values. B, Values for the more polar residue types. C, The less polar residue types in order of decreasing values of e_{ir}, which measures the hydrophobicity of the amino acid. The ordinates show the sum of the average contact energy and the repulsive packing energy, in dimensionless *RT* units; see equation (1) for the definition of the average contact energy and equation (40) for the repulsive packing energy. The hard core repulsion energy is not included here; $e^{hc} = 0$ in equation (41).

conformations with the most contacts tend to be the most stable in the assessment of the total contact energy. However, a rigorous treatment of binding requires an estimate of the entropy loss as well as the binding energy. On the other hand, non-representative proteins whose shapes differ significantly from a globular form, or small proteins like inhibitors, are often stabilized by disulfide bonds whose effects on protein stability are not counted here. Other effects not accounted for here are effects of the backbone and details of denatured state entropies. Also, as already noted, the estimate of e_{rr}, which is defined by equation (34) and reflects the overall compactness of proteins, is less reliable. Consequently, an assessment of protein stability based on the total number of contacts in a protein could lead to an incorrect result. Thus, in order to measure the stability of any protein structure for a given sequence, it might be better to use the energy that does not include the homogeneous energy for protein collapse but consists only of the remaining energy for aligning residues with the contacts assumed to be present within the protein structure. The alignment energy of residues within a protein structure is therefore defined here to be the total contact energy minus the average collapse energy:

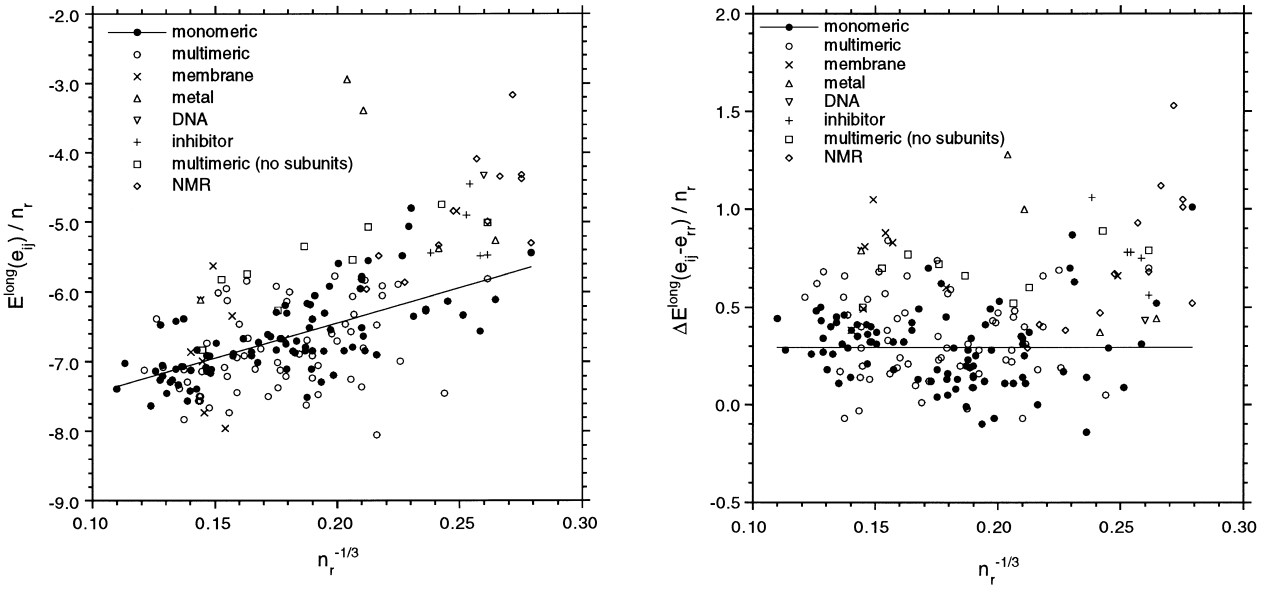
$$E_p^c(e_{ij} - e_{rr}) = E_p^c(e_{ij}) - E_p^c(e_{rr}) \qquad (7)$$

where $E_p^c(e_{ij} - e_{rr})$ is a function of $(e_{ij} - e_{rr})$ that is calculated by replacing e_{ij} with $(e_{ij} - e_{rr})$ in equation (18). $e_{ij} - e_{rr}$ here is named the alignment energy for contacts between residues of type *i* and *j*, and represents the relative preference of the *i*–*j* contacts compared to the average attraction.

It should also be noted that $E^c(e_{ij} - e_{rr})$ does not depend on n_{r0} but has only a linear dependence on the chain length; because e_s is almost equal to e_{rr}, the second term in equation (4) is negligible. The value of e_{rr} is estimated to be –2.55 in *RT* units as given in Table 3.

It is clear from equation (7) that $E_p^c(e_{ij} - e_{rr})$ will be positive unless hydrophobic residues are buried inside proteins and hydrophilic ones are exposed to solvent on the surfaces of proteins. In other words, the alignment energy, $E_p^c(e_{ij} - e_{rr})$, is consistent with the polar-out and non-polar-in rule observed in protein structures. Here it should be noted that equation (7) can be applied to proteins in water, but should not be used in a hydrophobic environment.

The total alignment energy in the form of equation (7) is appropriate for the consideration of the stabilities of protein conformations for a given sequence, but it is still inappropriate for the stabilities of a given fold for different protein sequences. In the latter case, a simple comparison of the energies of a given fold among protein sequences would be meaningless, because the ensemble of protein conformations could depend on the protein sequence. The stability of a specific conformation for a protein is determined in relation to the whole ensemble of protein conformations. Therefore, unless the whole ensemble of protein conformations is known, reference energies, each of which reflects to some extent the partition function for a



A B

Figure 6. A, The dependence of the long-range energy per residue on the inverse of the one-third power of chain length. B, The long-range energy per residue with a collapse energy subtracted to remove the protein size dependence. See equation (13) for the definition of the long-range energy, and equation (12) for $\Delta E^{long}(e_{ij} - e_{rr})$. The long-range energies include the contact energies and the repulsive packing energies, but not the hard core repulsion energies, i.e. $e^{hc} = 0$ in equation (41). All energies here are given in *RT* units. The representative protein structures used here are 189 protein structures that differ from each other by at least 35% sequence identity and are those selected by Orengo *et al*. (1993) (see their Table 1). Proteins with many unknown atomic coordinates are not included. The filled circles show the values for monomeric proteins determined by X-ray, not including membrane proteins, metal binding proteins, DNA binding proteins and inhibitors, and multimeric proteins not given in their complete assembly in the coordinate files. The open circles are proteins whose structures are given in an at least partial, if not full, assembly of subunits. A continuous line shows the regression line for the monomeric proteins. In A, the regression line is $E^{long}(e_{ij}) / n_r = -8.5 + 10.1 n_r^{-1/3}$, and the correlation coefficient is 0.67. In B, a collapse energy has been removed, so that the regression line is almost flat, $\Delta E^{long}(e_{ij} - e_{rr}) / n_r = 0.30 - 0.01 n_r^{-1/3}$, and the correlation coefficient is 0.001. The entry names and sequence identifiers of the PDB files used in preparing this Figure are:

Membrane proteins:
1PRC-L 1PRC-M 1PRC-C 2P0R 1SN3 1VSG-A 1HGE-A 1HGE-B 1PRC-H
Metal binding proteins:
1CY3 1PRC-C 5RXN 2HIP-A 2CDV
DNA binding proteins:
1HDD-C
Inhibitors:
1H0E 1PI2 3EBX 20V0 5PTI
Multimeric proteins without subunit interactions:
2WRP-R 1UTG 1R0P-A 2TMV-P 2RHE 2STV 3PGM 61DH 1PYP
Structures determined by NMR:
1C5A 1HCC 1ATX 1SH1 2SH1 1EPG 4TGF 3TRX 1EG0 1APS 1IL8-A 2GB1
Other monomeric proteins:
1MBC 1MBA 1ECD 2LH3 2LHB 1R69 4ICB 4CPV 1LE2 1YCC 1CC5 451C 1IFC 1RBP 1SGT 4PTP 2SGA 2ALP 2SNV 1CD8 1CD4 1ACX 1PAZ 1PCY 1GCR 2CNA 3PSG 1F3G 8I1B 1ALD 1PII 6XIA 2TAA-A 4ENL 5P21 4FXN 2FCR 2FX2 3CHY 5CPA 8DFR 3DFR 3ADK 1GKY 1RHD 4PFK 3PGK 2GBP 8ABP 2LIV 1TRB 1IPD 4ICD 1PGD 8ADH 2TS1 1PHH 31ZM 1LZ1 1RNH 7RSA 1CRN 1CTF 1FXD 2FXB 4FD1 1FDX 4CLA 9RNT 1RNB-A 1FKF 1SNC 1UBQ 3B5C 9PAP 3BLM 2CPP 1CSC 1ACE 1C0X 1GLY 1LAP 1LFI 2CYP 8ACN 2CA2
Other multimeric proteins:
1HBB-A 2SDH-A 1ITH-A 1C0L-A 1LMB-A 3SDP-A 2SCP-A 2HMZ-A 256B-A 2CCY-A 1GMF-A 1BBP-A 2FB4-H 3HLA-B 1C0B-A 2AZA-A 2PAB-A 1BMV-1 1BMV-2 2PLV-1 1TNF-A 2MEV-1 2MEV-2 2MEV-3 2PLV-2 2PLV-3 2LTN-A 2RSP-A 2ER7-E 5HVP-A 1NSB-A 5TIM-A 2TRX-A 1CSE-E 1GP1-A 4DFR-A 8CAT-A 4MDH-A 1GD1-0 7AAT-A 1HRH-A 1RVE-A 2SIC-I 8ATC-B 2TSC-A 2SAR-A 1MSB-A 1B0V-A 1FXI-A 1TGS-I 1TPK-A 9WGA-A 3HLA-A 8ATC-A 2CPK-E IGST-A IOVA-A 7API-A IWSY-B 2GLS-A 2PMG-A 6TMN-E 3GAP-A

sequence, are needed in order to measure the stabilities of a given fold for protein sequences. Here, we choose the total energy expected for a typical protein with the given amino acid composition and chain length to be the reference energy of the native structure for the given protein sequence, in order to compare the alignment energies of the native folds among a wide range of

protein sequences; see Miyazawa & Jernigan (1994) for a reference state that is appropriate to measure stability changes due to limited amino acid changes in a protein. That is, the following difference in energy is considered:

$$\Delta E^{\text{long}} \equiv E^{\text{long}} - (E^{\text{long}} \text{ of a typical native structure for a given sequence}) \quad (8)$$

$$\sim E^{\text{long}} - \sum_p f_{i_p} \cdot (\text{the average number of contacts per residue in a typical native structure}) \quad (9)$$

where the latter sum is the sum of the average contact energies per residue of residue type i_p over all positions in the protein. f_i as a function of e_{ij} is estimated by averaging over all proteins; that is:

$$f_i(e_{ij}) \equiv e_i \frac{N_{ir}}{N_{rr}} \frac{N_r}{N_i} \quad (10)$$

The values of $f_i(e_{ij})$ are given in Table 3. The average number of contacts per residue in a typical native structure for a given sequence is estimated as follows, ignoring the chain length dependence.

(the average number of contacts per residue in a typical native structure):

$$\sim N_{rr} / N_r = 2.096 \quad (11)$$

Because the second term of equation (9) does not depend on protein conformation but only on the amino acid composition and the length of the protein, it is a scaling factor and does not have any effect in a comparison of different conformations for the same protein. However, to a certain extent, its use will allow us to discuss how compatible a given protein sequence is with a certain structure, as well as how stable a particular conformation is for a certain sequence.

Lastly, a linear dependence on chain length is removed by comparison on a ''per residue'' basis, and the following quantity, which is appropriate for assessing the stability of one protein structure among other folds, is obtained:

$$\Delta E^{\text{long}}(e_{ij} - e_{rr}) / n_r \sim [E^{\text{long}}(e_{ij} - e_{rr}) - \sum_p f_{i_p}(e_{ij} - e_{rr}) \cdot (N_{rr} / N_r)] / n_r \quad (12)$$

where $\Delta E^{\text{long}}(e_{ij} - e_{rr})$ and $f_{i_p}(e_{ij} - e_{rr})$ are calculated by substituting e_{ij} with $(e_{ij} - e_{rr})$ in equations (9) and (10). ΔE^{long} includes both attractive and repulsive terms; the latter will be important principally for poor quality structures.

In Figure 6B, the estimates of $\Delta E^{\text{long}}(e_{ij} - e_{rr}) / n_r$ for the protein representatives are plotted against $n_r^{-1/3}$. As expected, overall there is no correlation between the two quantities for monomeric proteins; the mean for monomeric proteins is slightly more positive than zero, because the repulsive packing energy is included in $\Delta E^{\text{long}}(e_{ij} - e_{rr}) / n_r$. Therefore, this alignment energy has removed the dependence on the size of the protein. Membrane proteins, which are shown as crosses, tend to have much higher values of $\Delta E^{\text{long}}(e_{ij} - e_{rr}) / n_r$ than the mean for monomeric proteins. This is to be expected, because membrane proteins are not stable in water; in membrane proteins, portions exposed and embedded in the membrane are highly hydrophobic, and the surface is hydrophobic rather than hydrophilic, resulting in relatively high values of $\Delta E^{\text{long}}(e_{ij} - e_{rr})$. The same type of exception can be found for the metal binding proteins and the DNA binding proteins, in which metals and DNA have been treated here only as holes filled with water in the calculation of the total contact energies. Also, the multimeric cases, shown as open circles, tend to be located above the continuous line, probably because some coordinates of the inter-molecular neighbors are incompletely given in the partially assembled structures. On the other hand, the high values of energies for proteins whose structures were determined by NMR may indicate the relatively poorer resolution of these structures.

Discrimination for the native structures among other folds

The threading of sequences into other folds (Hendlich *et al*., 1990; Jones *et al*., 1992) or more generally the inverse folding (Bowie *et al*., 1991) is a good method to evaluate how well a given energy scale can discriminate the native structures as the lowest energy conformations among other folds. A more extensive study of inverse folding, in which deletions and additions are included, will be reported in a subsequent paper. Here, a result for simple threading of protein sequences without gaps into other folds is reported as a demonstration of the discrimination power of our alignment energy.

A total of 88 proteins determined to a resolution better than 2.5 Å by X-ray analyses, which are structurally dissimilar to each other with values smaller than 80 on the scale of Orengo *et al*. (1993) for structure similarity, were threaded into each of the 189 representatives of protein structures, which differ from each other by at least 35% sequence identity and were selected by Orengo *et al*. (1993) (see their Table 1). The 88 proteins are a subset of the 189 protein representatives whose entry names are listed in the caption of Figure 6. Coordinate files with too many unknown atomic coordinates are excluded from these data sets. Proteins classified within the multidomain group by Orengo *et al*. (1993) are also excluded from the set of sequences to be threaded.

$\Delta E^{\text{long}}(e_{ij} - e_{rr})$ is calculated for protein sequences threaded at all possible positions in all other protein structures, and their means and standard deviations are also calculated; no gaps in either the sequences or the structures are allowed. The positions of the native energies in the distributions of all threadings are then measured in units of standard deviation

(s.d.) where negative values indicate the native energies are below the mean. Table 5 lists values per residue, $\Delta E^{\text{long}}(e_{ij} - e_{rr})/n_r$, for the protein sequences threaded in their own native structures, as well as the ranks and the positions of the native energies in units of s.d. in the distributions of all threadings; proteins are sorted by the increasing order of the values, in units of standard deviation, of $\Delta E^{\text{long}}(e_{ij} - e_{rr})/n_r$. Favorable cases for proteins with more negative values than –5 s.d. are listed in Table 5A and the unfavorable higher ones in Table 5B.

For most proteins, the native structures have s.d. values of $\Delta E^{\text{long}}(e_{ij} - e_{rr})/n_r$ significantly large in magnitude and are ranked at the lowest energy. However, there are some proteins for which the native structures are not best or significantly better than all others. Proteins with values worse (higher) than –5 s.d., listed in Table 5B, are always membrane proteins or proteins, whose coordinates are given in isolated forms without their counterparts, such as small inhibitors, multimeric proteins and proteins binding metallic ions or other molecules. The relative proportions of binding regions on their surfaces may be significantly large for these small proteins.

Figure 7 shows the correlation between the values of $\Delta E^{\text{long}}(e_{ij} - e_{rr})/n_r$ in *RT* units and in s.d. units. It should be noted that their values in s.d. units do not depend on the second term of equation (12), but their absolute values in *RT* units do. Since the native energies in s.d. units depend not only on the native energies but also on the means and standard deviations of the energy distributions of threadings, a good correlation is not expected, especially in the low energy region. However, a fact that insignificant native folds have relatively high energies indicates that the energy function, $\Delta E^{\text{long}}(e_{ij} - e_{rr})/n_r$, may not only be used as an energy function to evaluate the stabilities of protein structures for a given sequence, but also as a scoring function to assess the compatibilities of protein sequences to a given structure. A more detailed discussion is planned for a subsequent paper.

Discussion

A basic assumption underlying the present estimation of contact energies is that, for a large enough sample, the effects of specific amino acid sequences will average out, and then the numbers of residue–residue contacts observed in a large number of protein crystals will represent the actual intrinsic interresidue interactions. As already noted in the original paper, this assumption is compatible with the ''principle of structural consistency'' originally proposed by Go (1983) and also called the ''principle of minimal frustration'' in the energy landscape view of proteins by Bryngelson & Wolynes (1987), because the present assumption is equivalent to the assumption that on average the intrinsic contact interactions are those consistent with the high stability of native structures. This

Table 5. Positions of native folds in the energy distributions of threadings

A. *Proteins with favorable native threadings*

PDB name	Length	Rank	Threadings	$\Delta E^{\text{long}}(e_{ij} - e_{rr})/n_r$[a]	In units of s.d[b]
7AAT-A	401	1	1681	0.17	–11.1
1PGD	469	1	765	0.27	–10.9
1PII	452	1	953	0.18	–10.1
2LIV	344	1	3084	0.35	–9.9
2GBP	309	1	4402	0.32	–9.9
8ADH	374	1	2254	0.29	–9.9
1PHH	394	1	1808	0.31	–9.9
2ER7-E	330	1	3548	0.20	–9.8
4ENL	436	1	1146	0.40	–9.7
2TS1	317	1	3972	0.21	–9.6
1ALD	363	1	2538	0.38	–9.5
1GD1-O	334	1	3406	0.40	–9.5
4FXN	138	1	16,873	–0.10	–9.3
3ADK	194	1	11,413	0.12	–9.1
3PGK	415	1	1426	0.45	–9.1
5TIM-A	249	1	7560	0.18	–8.8
1GKY	186	1	12,058	0.18	–8.8
1RVE-A	244	1	7876	0.24	–8.7
4PFK	319	1	3972	0.41	–8.7
1RHD	293	1	5173	0.31	–8.6
8CAT-A	498	1	548	0.62	–8.6
1IPD	345	1	3052	0.42	–8.6
6XIA	387	1	1953	0.46	–8.5
1MBC	153	1	15,196	–0.01	–8.5
3LZM	164	1	14,113	0.08	–8.4
2FCR	173	1	13,264	0.16	–8.3
1NSB-A	390	1	1889	0.55	–8.3
2TSC-A	264	1	6713	0.16	–8.3
1COL-A	197	1	10,951	0.12	–8.2
8ATC-B	146	1	15,196	0.15	–8.2
3CHY	128	1	18,067	–0.07	–8.2
4PTP	223	1	9338	0.38	–8.1
1PAZ	120	1	19,074	0.11	–7.9
1GCR	174	1	13,171	0.13	–7.8
2TRX-A	108	1	20,669	–0.07	–7.8
4DFR-A	159	1	14,598	0.20	–7.7
2CNA	237	1	8389	0.33	–7.7
4CPV	108	1	20,530	0.14	–7.6
1F3G	151	1	15,407	0.23	–7.6
1COB-A	151	1	15,407	0.31	–7.4
1MSB-A	115	1	19,724	0.22	–7.4
1LZl	130	1	17,821	0.28	–7.4
4CLA	213	1	10,054	0.12	–7.4
2LTN-A	181	1	12,563	0.24	–7.4
5CPA	307	1	4493	0.40	–7.3
4ICB	76	1	25,579	–0.14	–7.3
6LDH	329	1	3548	0.50	–7.2
5P21	166	1	13,922	0.29	–7.2
1RNH	148	1	15,300	0.34	–7.2
1CSE-E	274	1	6151	0.57	–7.0
1BOV-A	69	1	26,735	0.05	–6.9
1FKF	107	1	20,812	0.25	–6.9
1UBQ	76	1	25,579	0.14	–6.8
2AZA-A	129	1	17,943	0.43	–6.8
1YCC	108	1	20,669	0.31	–6.8
1RBP	175	1	13,081	0.45	–6.8
256B-A	106	1	20,957	0.31	–6.7
1ACX	108	1	20,669	0.34	–6.7
1FXD	58	1	28,625	0.31	–6.6
3B5C	86	1	23,674	0.17	–6.5
1LMB-A	87	1	23,056	0.19	–6.5
1GMF-A	119	1	19,203	0.23	–6.4
9RNT	104	1	21,250	0.37	–6.4
2RSP-A	115	1	18,565	0.28	–6.4
9WGA-A	170	1	13,451	0.59	–6.1
7RSA	124	1	18,565	0.53	–6.1
2SIC-I	107	1	20,812	0.40	–6.1
2SAR-A	96	1	22,443	0.40	–5.8
1PRC-C[c]	333	1	3441	0.79	–5.7
2PAB-A	114	1	18,692	0.45	–5.7
2HIP-A[d]	71	1	26,400	0.37	–5.5
2RHE[e]	114	1	19,857	0.52	–5.4
5RXN[f]	54	1	29,347	0.44	–5.0

[a] Alignment energies per residue in *RT* units; see equation (12) for definition.

[b] The position of the native energy in the distribution of all threadings in units of s.d., where negative values are for native energies below the mean.

[c] Photosynthetic reaction center; four protoporphyrin IX are bound.

[d] High potential iron sulfur protein; four Fe and four S are bound.

[e] Bence-Jones protein.

[f] Rubredoxin; small Fe binding protein.

Table 5—*continued*

B. *Proteins with less favorable native threadings*

PDB name	Length	Rank	Threadings	ΔE^{long} $(e_{ij} - e_{rr})$ $/ n_r$[a]	In units of s.d.[b]	Comment
2OVO	56	1	28,983	0.56	–4.9	Ovomucoid third domain; 3 S–S bonds
2STV	184	1	11,413	0.72	–4.5	Coat protein of S.T. virus; multimeric
2WRP-R	104	1	20,812	0.60	–4.5	Trp repressor; DNA binding
1SN3	65	5	27,410	0.66	–4.1	Scorpion neurotoxin; membrane binding protein
1TPK-A	88	11	23,674	0.69	–4.1	Tissue plasminogen activator, Kringle-2 domain
3EBX	62	15	27,925	0.78	–3.8	Erabutoxin B; inhibitor to acetylcholine receptor
1UTG	70	15	26,567	0.89	–3.7	Uteroglobin; progesterone binding protein
1PI2	61	26	27,752	0.78	–3.7	Bowman–Birk proteinase inhibitor; no enzyme
5PTI	58	44	28,625	0.75	–3.6	Trypsin inhibitor; no enzyme
2POR	301	3	4778	1.05	–3.3	Porin; integral membrane protein
2CDV	107	50	20,812	1.01	–2.8	Cytochrome c_3; small protein with 4 hemes
1CRN	46	>100	30,832	1.01	–2.6	Crambin
1HOE	74	>100	25,905	1.06	–2.2	α-Amylase inhibitor without enzyme
1PRC-L	273	>100	6205	0.88	–2.0	Photosynthetic reaction center; membrane protein
1CY3	118	>100	19,333	1.28	–1.4	Cytochrome c_3; small protein with 4 hemes

[a] Alignment energies per residue in *RT* units; see equation (12) for definition.
[b] The position of the native energy in the distribution of all threadings in units of s.d., where negative values are for native energies below the mean.

assumption is also equivalent to the assumption that the distribution of the numbers of contacts in protein structures is a ''self-averaging property'' in the terms of Bryngelson *et al*. (1995), which means it is, in the present case, the property of heteropolymers rather than of the detailed amino acid order of protein sequences. Gutin *et al*. (1992) indicated that the Boltzmann-like statistics observed in protein structures are a general property of the stable structures of heteropolymer chains, and that the ''temperature'' in these statistics is not the usual temperature of the medium but a ''selective temperature'', at which the native structure is ''frozen out'' from an exponentially large set of other structures; see also Săli *et al*. (1994) for the estimation of the selective temperature or critical temperature.

We have not stated explicitly what temperature should be taken for the conversion of the contact energies from *RT* units to kcal/mol in our original paper, but a melting temperature is implied, which is just low enough for the native structure to be marginally stable and high enough for the energy landscape of protein to show minimum frustration and for the ''self-averaging property'' of contacts to be satisfied. In the analyses given in our papers (Miyazawa & Jernigan, 1994, 1985), however, room temperature was used to translate the *RT* units into kcal/mol. The reason is that the contact energies estimated here are free energies which depend on temperature. In general, the conformational energy of a protein is implicitly not an energy but rather a free energy, because it is usually a potential of mean force that is obtained by integrating over solvent coordinates and therefore includes energetic and entropic solvent effects. Therefore, as long as the melting temperature is not too far away from room temperature, it is preferable to use the experimental room temperature for unit conversions. The use of average melting temperature or room temperature for the conversion differs significantly from the claim of Gutin *et al*. (1992) that a critical temperature estimated to be a factor of 1.5 higher than room temperature should be used for the conversion, but does not contradict the claim of Săli *et al*. (1994) that the critical temperature could also be less than room temperature, because rapid folding into the stable native structure occurred slightly above the critical temperature and the folding temperature could be equated with room temperature.

The hydrophobic effect is a dominant force in stabilizing the native structure of a globular protein. However, there is a lack of agreement as to its precise magnitude. The hydrophobicity of a small molecule is usually measured by the transfer energies, for example, of amino acids from a non-polar solvent to water. However, it is not immediately evident that the free energy change accompanying a process in which residues are buried in the folding process of protein is the same as that accompanying the transfer process of amino acids from water to a nonaqueous solvent, because both processes are obviously different (Lee, 1993). In our previous work on contact energies, we pointed out that the estimates of the contact energies of amino acid side-chains were almost twice as large in magnitude as the usual estimates of hydrophobicities from the transfer energies of Nozaki & Tanford (1971). In order to measure directly the contribution of the hydrophobic effect to the stability of proteins, many experiments measuring the stability change caused by amino acid replacements have been performed (Yutani *et al*., 1984; Matsumura *et al*., 1988; Shortle *et al*., 1990). These experiments showed that the unfolding free energy changes upon single amino acid replacements could be significantly larger in magnitude than those expected from the transfer free energies of amino acids (Yutani *et al*., 1984; Shortle *et al*., 1990), but were also highly variable (Shortle *et al*., 1990).

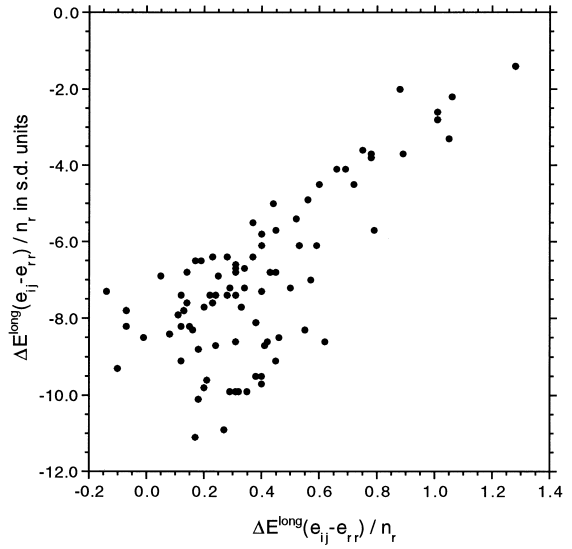


Figure 7. The correlation between the values of $\Delta E^{\text{long}}(e_{ij} - e_{rr}) / n_r$ of the native structures of 88 proteins in RT units and those in standard deviation units from its distribution of threaded structures. A total of 88 proteins determined to a resolution better than 2.5 Å by X-ray analyses, which are structurally dissimilar to each other with values smaller than 80 on the scale of Orengo *et al*. (1993) for structure similarity, are threaded into each of the 189 representatives of protein structures, which differ from each other by at least 35% sequence identity and were those selected by Orengo *et al*. (1993) (see their Table 1). $\Delta E^{\text{long}}(e_{ij} - e_{rr}) / n_r$ is then calculated according to equation (12) for all threadings, with no gaps allowed, and its mean and standard deviation are calculated. The 189 protein representatives are the same ones used for Figure 6. A total of 88 proteins are a subset of these 189 protein representatives and are listed in Table 5. Coordinate files with too many unknown atomic coordinates are excluded from these data sets. The correlation coefficient is 0.75.

The effect of a cavity created by such amino acid replacements was considered to be one of the factors causing this discrepancy (Shortle *et al*., 1990; Nicholls, 1991). Eriksson *et al*. (1992) reported that the unfolding free energy changes showed correlations with the increases in the sizes of cavities observed in the protein structures of mutant T4 lysozymes, and that the value of the unfolding free energy change extrapolated to zero cavity size coincided with the value expected from the transfer free energy of an amino acid. Lee (1993) estimated the free energy changes for cavities formed by replacing a bulky side-chain with a smaller side-chain, assuming the protein structure to remain completely rigid. He pointed out that most experimental values of unfolding free energy changes for such replacements fell into a range between the maximum expected for the full cavity size and the minimum expected for no cavity, because rearrangement of the protein occurs to fill the cavity.

On the other hand, Sharp *et al*. (1991) argued that the previous estimates of the hydrophobic effect derived from analyses of solute partition data did not fully account for changes in volume entropy, and they provided new estimates by including a volume correction term. Their overall estimates were almost twice as large as the previous ones and similar in magnitude to the estimates in this work of the contact energies for the transfer energy of side-chains from the outside to the interior of a protein. Pace (1992) also reported that the estimate of hydrophobicity with the volume correction could explain the values of unfolding free energy changes observed for 72 aliphatic side-chain mutants from four proteins in which a larger side-chain was replaced by a smaller side-chain.

As shown in Figure 3, the estimates by Sharp *et al*. (1991) of the transfer free energies between octanol and water for non-polar side-chains are almost identical with the present estimates of equivalent quantities for hydrophobic side-chains. Here, it should be noted that the contributions of amino acid side-chains to the contact energy are calculated to be equal to the difference of the contact energies between Gly and other amino acids and these estimates are expected to be more reliable than the absolute values of the contact energies of amino acids.

Miyazawa & Jernigan (1994) showed, based on the previous estimates of contact energies, that the contact energy changes of protein structures due to single amino acid replacements were large enough to account for the observed values of unfolding free energy changes. Also, the large variation of unfolding free energy changes among residue positions for single amino acid replacements, observed by Shortle *et al*. (1990), was accounted for if the free energy changes in the denatured state were taken into account. The free energy change in the denatured state of a protein due to single amino acid replacements had not been considered and had usually been ignored. The large variation of unfolding free energy changes may be due to the different environments surrounding each residue in both the native structure and the denatured state. Miyazawa & Jernigan (1994) indicated that staphylococcal nuclease, under the experimental conditions of Shortle *et al*. (1990), was not fully expanded in the denatured state and also that the compact denatured state might have a substantially native-like topology.

In the original work to evaluate the contact energies, the number of protein structures used was only 42, including 30 monomeric proteins (Miyazawa & Jernigan, 1985). These proteins were chosen with the criteria that their chain lengths were longer than 100 residues and that few atomic positions were missing. Also, in cases where coordinates are available for several closely homologous proteins, only one representative was used; the minimum difference between amino acid sequences for homologous proteins was 50%. The numbers of contacts were calculated for the complete assembly. Now, more than 1000 protein structures are available. Here, we use only the coordinates of subunits as given in the PDB files, without

calculating any additional subunit intersections. That is, if a PDB file is given in a monomeric state, the numbers of contacts have been counted in the monomeric state, even though matrices to generate multiple symmetric subunits are given in the PDB file. The ratio of molecular surface contacts to the total number of contacts

$$\left\{\frac{N_{ro}}{(N_{rr} + N_{ro})}\right. \text{ is about } 0.33.$$

Generally, subunit–subunit interfaces and enzyme–substrate interfaces are not the major portion of a molecular surface. Therefore, ignoring the molecular binding in some proteins should not affect the estimated values of contact energies so much.

There are many homologous proteins in the PDB. Any statistical analysis of sequence–structure relations of proteins requires an unbiased sampling of proteins. One way is to choose representative sets of proteins that are sufficiently dissimilar to each other. Usually, selections of protein representatives are based on sequence dissimilarity (Hobohm *et al*., 1992) or structure dissimilarity (Orengo *et al*., 1993; Fischer *et al*., 1996). The alternative method used here is to assign a different sampling weight to each protein according to the extent of redundancy (sequence identity), to remove sampling bias. This has not been done previously, but we show that such a sampling weight for each protein can be based on a similarity matrix between proteins in the data set. Here, each element of the similarity matrix has been evaluated as the ratio of identical residues between aligned protein sequences, and then the sequence identity matrix has been diagonalized. The present data set contains 1661 protein sequences. To reduce computational time for the calculation of the 1661 by 1661 sequence identities among sequences and for diagonalization of the resulting sequence identity matrix, the matrix size was reduced by regarding highly similar sequences as identical, i.e. if they are more than 95% identical. If sequence identities lower than a certain value ought to be ignored, then matrix elements of sequence identity, equation (44), whose values fall below a threshold value, could be replaced by zero; 30% identity, where sequence pairs may be evolutionally unrelated, may be a good choice for such a threshold. For statistical analyses similar to the present one, this method will certainly be better if homologous proteins are equally good structures, because all available data are used and then the statistical error becomes lower for a larger number of samples.

Remarkably, all characteristics of the contact energies found in the original analysis hold for the present results, for which the total effective number of contacts is 6.3 times more than in the original data. The new values differ from the original estimates of the contact energies to a significant extent only for the infrequent amino acids Trp and Met and the one other exception, Leu. The reason why the contacts with Leu are more frequently observed is not clear.

The contact energies are attractive energies only, and volume exclusion between residues needs to be taken into account, unless a lattice model with a coordination number small enough to prevent overpacking is used. In a lattice model, volume exclusion between residues can be satisfied readily by assuming that a lattice site cannot be occupied by multiple residues. When a lattice model is not used, a repulsive potential is needed for evaluation of the conformational packing energy of a protein. This has been one motivation for the present work. The van der Waals' surfaces of side-chains are highly anisotropic, making spherically symmetric repulsive potentials inappropriate to represent effective two-body repulsive interactions between two residues. The present many-body potential has been formulated to depend instead on packing density to represent the repulsive interactions around a central residue in a protein and to treat packing density effects more realistically than would be possible with a spherically symmetric two-body potential. However, such a repulsive potential is not in effect for small numbers of neighboring residues. In addition, there is a hard core, which cannot be penetrated, for any side-chain with any conformation. However, a preliminary investigation showed that there is relatively little residue type dependence for these hard cores, and that the minimum size of such a hard core may correspond closely to the van der Waals' size of a methyl group. Therefore, the repulsive potential between residues here is represented as the sum of two kinds of potentials, a hard core potential that is a two-body potential and a repulsive packing potential that is a many-body potential depending on the number of residues immediately surrounding a residue. The repulsive packing potentials for 20 types of amino acids have been estimated as a function of the number of contacts from the high density portion of the distributions observed in known protein structures. This tail portion of the distribution, which is defined as the region of high coordination numbers of amino acids, should reflect the effective repulsive packing energy between residues.

We have shown how this new potential for interactions among residues can be used to calculate the total energies of a set of proteins and how these energies behave roughly as expected. Also, calculations of threading sequences into other folds have demonstrated that the native structures have the lowest alignment energies of residues among all other folds for most of protein representatives in the PDB, the exceptions being proteins such as membrane proteins and others having bound ligands. The application to threading here may not be the most demanding test of the potentials, because it only requires discrimination among relatively few conformations. However, the results are excellent and give, at least, a taste of the level

of success achievable with it. This new potential is appropriate for application to a broad range of protein conformation simulations and inverse folding calculations.

Methods

Long-range inter-residue interaction energy

Long-range interaction energies are simplified for inter-residue interactions at the residue level without any atomic details of side-chains. They are approximated to consist of two terms, a short-range attractive term that becomes effective only when two residues are in close proximity, and a repulsive term that results from the overlap of residues at high packing densities. In the case of a lattice model, volume exclusion between residues is implicitly included in the model, because a site cannot be occupied by more than one residue. However, if conformational space is defined continuously rather than discretely, a repulsive energy potential is essential to account for volume exclusion among residues. In the following, the long range inter-residue energy of a protein is represented as a sum of these two terms over all residues in a protein:

$$E^{\text{long}} = \sum_p (E_p^{\text{c}} + E_p^{\text{r}}) \tag{13}$$

where p is a residue sequence index in a protein.

Contact energy

Residues are represented by single points at the centers of their side-chain atom positions; the positions of C^{α} atoms are used for glycine residues. Residues whose centers are closer than R^{c} are defined to be in contact. The limiting value $R^{\text{c}} = 6.5$ Å for contacts was chosen on the basis of the occurrence of the first peak in the radial distribution of residues in the interior of proteins (Miyazawa & Jernigan, 1985).

Intra-residue interactions and nearest-neighbor interactions also lead to contact formation among nearest neighbor residues, and therefore in the present evaluation of the long range interactions, these nearest-neighbor contacts are explicitly excluded.

Thus, a contact between the pth and qth residues is defined using:

$$\Delta_{pq}^{\text{c}} \equiv \begin{cases} 0 & \text{if } |p-q| \leqslant 1 \\ \mathbf{H}(R^{\text{c}} - d_{pq}) & \text{if } |p-q| > 1 \end{cases} \tag{14}$$

$$R^{\text{c}} \equiv 6.5 \text{ Å} \tag{15}$$

where $\mathbf{H}$ is the Heaviside function defined as:

$$\mathbf{H}(x) \equiv \begin{cases} 1 & \text{if } x \geqslant 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{16}$$

and d_{pq} is the distance between the pth and qth residues. Maiorov & Crippen (1992) pointed out that, in the use of discrete functions for the definitions of contacts, slight changes in interatomic distances can produce significantly different lists of contacts. Since residue–residue contacts are defined here by using the distance between their side-chain centers, they are relatively insensitive to small variations in interatomic distances. However, a sigmoidal function with a transition width of about 1 Å might have been used to advantage, instead of the Heaviside function, in equation (16) to account for uncertainties in the boundary region. The number of residues of type j in contact with the pth residue whose amino acid is i_p type is:

$$n_{i_p j}^{\text{c}} = \sum_{q(\neq p)} \delta_{j_q j} \Delta_{pq}^{\text{c}} \tag{17}$$

where δ is the Kronecker δ function.

The contact energy for the pth residue, E_p^{c}, in which the energy of a conformation with no residue–residue contacts is defined to be zero, is represented as:

$$E_p^{\text{c}}(e_{ij}) = \frac{1}{2} \sum_{q(\neq p)} e_{i_p j_q} \Delta_{pq}^{\text{c}} \tag{18}$$

$$= \frac{1}{2} \sum_{j(\neq 0)} e_{i_p j} n_{i_p j}^{\text{c}} \tag{19}$$

where e_{ij} is the energy difference accompanying the formation of contacts between i and j types of amino acids from those amino acids exposed to solvent, and is defined as follows (Miyazawa & Jernigan, 1985):

$$e_{ij} \equiv E_{ij} + E_{00} - E_{i0} - E_{j0} \tag{20}$$

where 0 means effective solvent molecules, and E_{00} and E_{i0} are the absolute interaction energies between a pair of solvents and between an i type of residue and an effective solvent, respectively. Here, it should be noted that a lattice model is used to take account of interactions among solvent molecules and amino acids in a protein. Residues in a protein are assumed to occupy lattice sites or cells as a linear chain. Each vacant cell is regarded as being occupied by an effective solvent molecule.

$n_{i_p j}^{\text{c}}$ can be summed to produce the following equation:

$$n_{ij} = \frac{1}{2} \sum_p \sum_{q(\neq p)} \Delta_{i_p i}^{\text{c}} \Delta_{j_q j}^{\text{c}} = \frac{1}{2} \sum_p n_{i_p j}^{\text{c}} \delta_{i_p i} \tag{21}$$

$$\frac{1}{2} q_i n_i = \sum_{j=0}^{\text{c}} n_{ij} \tag{22}$$

$$n_{ij} = n_{ji} \tag{23}$$

where n_{ii} and $2n_{ij}$ are the total numbers of contacts between two residues of the same type, i, and between i and j types of amino acids; n_i is the total number of i type amino acids in a protein and q_i is the coordination number for the i type of amino acid, that is, the average number of residues that completely surround the i type of amino acid. Let us define here the following quantities for the typical residue r:

$$n_{ir} = n_{ri} \equiv \sum_{j(\neq 0)} n_{ij} \tag{24}$$

$$n_{rr} \equiv \sum_{i(\neq 0)} n_{ir} \tag{25}$$

$$n_r \equiv \sum_{i(\neq 0)} n_i \tag{26}$$

The summations above are taken over all 20 amino acid types.

How to estimate contact energies

The contact energies have been re-evaluated by using a newer, much larger protein data set with the same procedure described by Miyazawa & Jernigan (1985), in which the following assumptions and approximations were employed; the original notation is used here.

(1) For a large enough sample, the effects of specific amino acid sequences average out and the numbers of non-bonded residue–residue contacts observed in a large number of protein crystal structures will then represent the actual intrinsic differences of interactions among residues in proteins. Of course, nearest-neighbor contacts along chains are significantly affected by the amino acid sequences of proteins. Therefore, the contacts between nearest-neighbor residues are explicitly excluded in the counting of contacts. The effect of a protein being a chain may remain and might affect somewhat the observed number of residue–residue contacts. Therefore, the contact energies estimated here should be regarded as effective inter-residue energies.

(2) By taking account of the effects of the chain connectivity as imposing a limit to the size of the system, i.e. the total number of lattice sites, the system is then regarded as an equilibrium mixture of unconnected residues and effective solvent molecules.

(3) The Bethe approximation (quasi-chemical approximation), which gives an exact solution for the Bethe lattice, and in which contact pair formation can be regarded as a chemical reaction, is used to estimate the contact energies from the numbers of contacts observed in known protein structures. In the Bethe approximation, e_{ij} satisfies the following equilibrium equation:

$$\exp(-e_{ij}) = \frac{\bar{n}_{ij}\bar{n}_{00}}{\bar{n}_{i0}\bar{n}_{j0}} \tag{27}$$

where $\bar{n}_{ij}$ represents the statistical average of n_{ij}. Note that all energies here are taken to be dimensionless, i.e. in units of RT. Usually this equation is used to evaluate $\bar{n}_{ij}$ from known e_{ij}, but it is used inversely here to estimate the contact energies from the numbers of contacts observed in known protein structures. However, the equation above includes n_{00}, the number of contacts between effective solvent molecules, which is rather difficult to estimate accurately. However, the differences between contact energies, such as e'_{ij} defined in the next paragraph, do not depend on n_{00} and so the estimates of such relative quantities ought to be more reliable than the absolute values of contact energies; e.g.:

$$\exp(e_{ij} - e_{kl}) = \frac{\bar{n}_{ij}\bar{n}_{k0}\bar{n}_{l0}}{\bar{n}_{i0}\bar{n}_{j0}\bar{n}_{kl}} \tag{28}$$

does not include n_{00}.

(4) n_{00} is evaluated through an estimation of the number of effective solvent molecules. The number of effective solvent molecules for each protein is chosen to yield the total number of residue–residue contacts equal to its expected value for the hypothetical case of hard sphere interactions with $e_{ij} = 0$ among residues and effective solvent molecules, representing the effects of chain connectivity. An effective solvent molecule is taken to have the same volume (of water molecules) as an average residue.

(5) The numbers of contacts, n_{ij}, are counted for each protein structure and the sums of n_{ij} over all protein samples, N_{ij}, are calculated. The numbers of contacts with effective solvent molecules, n_{i0}, are estimated with equation (22); the coordination number for each amino acid type is estimated from the volume of each type of amino acid at the center and the average volume of its surrounding residues. That is, incomplete coordination spheres are completed with solvent molecules. Then, the contact energies are estimated from N_{ij} with a composition correction, so that if all residues and effective solvent molecules were randomly mixed, the estimated values of the contact energies would be zero:

$$e'_{ij} = -\frac{1}{2}\ln\left(\frac{N_{ij}^2 C_{ii} C_{jj}}{N_{jj} N_{jj} C_{ij}^2}\right) \quad \text{for } i, j \neq 0 \tag{29}$$

$$e'_{i0} = -\frac{1}{2}\ln\left(\frac{N_{i0}^2}{N_{ii} N_{00}} \frac{C'_{ii} C'_{00}}{C'^2_{i0}}\right) \tag{30}$$

where:

$$e'_{ij} \equiv e_{ij} - (e_{ii} + e_{jj})/2 \tag{31}$$

$$e_{ij} = e'_{ij} + e'_{00} - e'_{i0} - e'_{j0} \tag{32}$$

C_{ij} and C'_{ii} are correction factors and are equal to the number of contacts expected between residues of i and j types and the number of contacts expected between residues of the same type i, respectively, when residues and effective solvent molecules are randomly mixed; see equations (10) to (15) of Miyazawa & Jernigan (1985) for their detailed definitions. e'_{ij} is the energy difference detailing the residue interaction specificity accompanying the formation of a contact pair i–j from contact pairs i–i and j–j.

According to the procedure described by Miyazawa & Jernigan (1985) and briefly summarized above, the contact energies, e_{ij}, between i and j types of amino acids are re-evaluated with equations (29) and (30) above, or equations (10) to (15) of Miyazawa & Jernigan (1985). However, the further useful quantities for interactions with the average residue r are defined as:

$$\exp(-e_{ir}) \equiv \frac{\bar{n}_{ir}\bar{n}_{00}}{\bar{n}_{i0}\bar{n}_{r0}} \tag{33}$$

$$\exp(-e_{rr}) \equiv \frac{\bar{n}_{rr}\bar{n}_{00}}{\bar{n}_{r0}\bar{n}_{r0}} \tag{34}$$

These are not estimated by the procedure described in the earlier paper, but directly from the data by the following equations:

$$e_{ir} = e'_{ir} - e'_{i0} - e'_{r0} \tag{35}$$

$$e'_{ir} = -\tfrac{1}{2}\ln(C_{ii}/N_{ii}) \tag{36}$$

$$e_{rr} = -2e'_{r0} = \ln(C'_{00}/N_{00}) \tag{37}$$

As stated by Miyazawa & Jernigan (1985), if e_{ir} is more negative than e_{rr}, then amino acids of the i type will tend to be buried inside a protein; otherwise, they will tend to be exposed to solvent on the surface of the protein.

Also, the average contact energy of each type of amino acid is estimated by:

$$e_i = \sum_{j(\neq 0)} e_{ij} N_{ij} / N_{ir} \tag{38}$$

$$e_r = \sum_{i(\neq 0)} e_i N_{ir} / N_{rr} \tag{39}$$

e_i may be used to compare with other estimates of hydrophobic energies. However, e_{ir} will be more appropriate as a one-dimensional measure of hydrophobicity. e_i and e_{ir} correspond to the mean energy and the

effective mean energy, respectively; refer to equation (35) of Miyazawa & Jernigan (1985).

Repulsive energy

Here, each residue has been represented by the point at the center of its side-chain heavy atom positions. Generally, the van der Waals' surface of a side-chain is highly anisotropic, and its hard core would not be well represented by a sphere. However, if a repulsive potential between residues is averaged over all possible side-chain conformations, then the anisotropic character in the average repulsive potential will be weakened by the flexibility manifested in the side-chain conformations. Therefore, a symmetric repulsive potential is appropriate at the present level of simplification. However, a repulsive force resulting from van der Waals' overlaps is a very short range force, and therefore even a soft core symmetric repulsive potential may be inappropriate. Here, the repulsive force of van der Waals' overlaps is approximated as the sum of two terms, a hard core repulsion between residues and a repulsive packing potential, depending on the packing density of residues:

$$E_p^{\mathrm{r}} = e_p^{\mathrm{hc}} + e_p^{\mathrm{r}} \tag{40}$$

The hard core potential is taken, in turn, as a two-body interaction potential given by:

$$e_p^{\mathrm{hc}} \equiv \frac{1}{2} \sum_{q(\neq p)} e^{\mathrm{hc}} \mathbf{H}(r^{\mathrm{c}} - d_{pq}) \tag{41}$$

e^{hc} is the parameter for the positive energy of hard core repulsion and will have a large positive value. Depending on side-chain conformation, the side-chain centers of two residues can sometimes come as close as the van der Waals' separation between two methyl groups. Therefore, the hard core radius, $r^{\mathrm{c}}/2$, is independent of residue type and here is simply the van der Waals' core of a methyl group, with a value of 1.9 Å.

The repulsive packing potential is a many-body interaction potential dependent on n_p^{c}, the number of residues in contact with the pth residue:

$$n_p^{\mathrm{c}} \equiv \sum_{j(\neq 0)} n_{i_p j}^{\mathrm{c}} \tag{42}$$

This packing density energy in RT units is estimated as follows:

$$e_p^{\mathrm{r}} \equiv \mathbf{H}(n_p^{\mathrm{c}} - q_{i_p}) \left[\left(\frac{q_{i_p}}{n_p^{\mathrm{c}}} - 1 \right) E_p^{\mathrm{c}} - \ln \left(\frac{N(i_p, n_p^{\mathrm{c}}) + \epsilon}{N(i_p, q_{i_p}) + \epsilon} \right) \right] \tag{43}$$

$N(i, n^{\mathrm{c}})$ is the total number of residues of type i that are surrounded by n^{c} residues in the set of protein structures. The value of $N(i, q_i)$ is obtained by interpolation. To avoid the divergence of the logarithm function in equation (43), a small positive number, ϵ, has been added to both the numerator and the denominator, so that the sum of the contact energy and repulsive energy takes on a positive value even at $N(i, n^{\mathrm{c}}) = 0$ for any amino acid; here, a value of $\epsilon = 10^{-6}$ is employed.

The first term of the repulsive packing potential in equation (43) represents the effect of exceeding the limiting number of contacts, which is equal to the coordination number q_{i_p}, to offset the extra contact energy; the actual value of E_p^{c} from equation (19) is used. The second term is estimated from the ratio of the observed frequency of packing density to that for the limiting value for packing density. The repulsive energy is applied through the Heaviside function only if the number of surrounding residues exceeds a threshold value, q_{i_p}. The distribution of the number of contacts for each amino acid is assumed in the Bethe approximation to be determined by the contact energies, except for the high density region, which should reflect the repulsive energies among residues. It can be seen in Table 3 that the estimated coordination numbers for the 20 types of amino acids here range only from 5.79 for Trp to 6.65 for Cys.

Protein structures used in the present statistical analysis

The proteins used here are all those in the Protein Data Bank (PDB) (Bernstein *et al*., 1977) that satisfy the following conditions.

(1) Proteins whose structures are determined by X-ray analysis and whose resolution is equal to, or better than, 2.5 Å. All protein structures determined by NMR are excluded.
(2) Protein subunits composed of 50 or more residues.
(3) Membrane proteins are excluded, because, inappropriate to them, it is assumed that incomplete coordination spheres are completed with water molecules. Thus, the number of inter-residue contacts compiled here are those observed only in soluble proteins in order to derive contact energies appropriate for soluble proteins. However, some characteristics of neighbor pairs might be preserved, except for the overall inversion of structure compared to soluble proteins.

The number of protein subunit structures satisfying these criteria is 1661 (see Table 1).

In this data set, there are many proteins whose sequences are similar to one another. To obtain statistically unbiased results, unbiased sampling is required. There are two ways to do this: either (i) use protein representatives that are sufficiently dissimilar to each other in their sequences; or (ii) use a different statistical weight for each protein related to its extent of similarity to other sequences. So far, most statistical analyses have used a representative set of proteins. Usually, protein representatives are chosen by specifying an upper limit for sequence identity (Hobohm *et al*., 1992) or structural similarity (Orengo *et al*., 1993; Fischer, *et al*., 1996, among them). However, it is not clear what value is best as an upper limit of similarity in protein representatives. Also, in such a method, many good structures may be discarded. In this work, the second approach has been taken.

Sampling weight

A sequence identity matrix is defined here as:

$$\begin{aligned} I_{\mu\nu}^{\mathrm{s}} &\equiv \text{sequence identity between sequence } \mu \text{ and } \nu \\ &\equiv \frac{2 \times (\text{number of identical residues in the alignment})}{(\text{length of sequence } \mu) + (\text{length of sequence } \nu)} \end{aligned} \tag{44}$$

The sequence identity is taken as the fraction of identical residues in the alignment of two sequences. The sequence identity matrix is a real symmetric, non-negative matrix.

Each element of $I^s_{\mu\nu}$ is between 0 and 1, and the diagonal elements are equal to one:

$$0 \leqslant I^s_{\mu\nu} = I^s_{\nu\mu} \leqslant 1$$
$$I^s_{\mu\mu} = 1 \tag{45}$$

The sequence identity matrix is similar to a correlation matrix in the sense that the value of $I^s_{\mu\nu}$ represents the correlation between the amino acid sequences μ and ν. Let us define λ_μ as the μth eigenvalue and V_μ the μth column eigenvector that is orthogonal to all others and normalized to be equal to one. That is:

$$I^s V_\mu = \lambda_\mu V_\mu \tag{46}$$

$$V^T_\mu V_\nu = \delta_{\mu\nu} \tag{47}$$

V^T_μ means the transpose of V_μ. The total sum of the eigenvalues is equal to the trace of the sequence identity matrix, and so it is equal to the number of sequences used:

$$\sum_\mu \lambda_\mu = \mathrm{Tr}\ I^s = \sum_\mu 1 = N_{\mathrm{prot}} \tag{48}$$

where N_{prot} is the number of sequences used. From equation (45), it follows that a sequence identity matrix must be positive semidefinite:

$$0 \leqslant \lambda_i \leqslant N_{\mathrm{prot}} \tag{49}$$

Let us consider some special cases. If a whole set of sequences can be divided into groups where individual sequences from different groups are completely dissimilar, each group can be handled independently. In case there is a sequence which is completely dissimilar to any other sequence, at least one eigenvalue will be equal to one. If all sequences are the same, all other eigenvalues, except one, must be equal to zero; if $I^s_{\mu\nu} = 1$ for any μ and ν, rank $I^s = 1$, and therefore the number of non-zero eigenvalues must be equal to one.

On the basis of these characteristics, the following procedure may be used to remove redundant information that comes from similarities among sequences. The sampling weight, w_ν for the νth sequence is taken to be:

$$w_\nu \equiv \left(\sum_\mu (\lambda_\mu - (\lambda_\mu - 1) \cdot \mathbf{H}(\lambda_\mu - 1)) \cdot V_\mu V^T_\mu \right)_{\nu\nu} \tag{50}$$

$$0 < w_\nu \leqslant 1 \tag{51}$$

This definition of sampling weight satisfies all requirements. If, and only if, a sequence has zero sequence identity with any other sequence, then the sampling weight for that sequence will be equal to one. If N_{prot} sequences in the data set are all identical, then the statistical weights for those sequences will be equal to $1/N_{\mathrm{prot}}$. Generally, sampling weights take a value between one and $1/N_{\mathrm{prot}}$, and are approximately equal to the inverse of the number of similar sequences. The effective number of proteins can be defined as the total sum of the sampling weights:

$$N^{\mathrm{effective}}_{\mathrm{prot}} \equiv \sum_\mu w_\mu \leqslant N_{\mathrm{prot}} \tag{52}$$

In a similar way, the effective number of residues is defined.

Please note the use of sequence identity here. Although this approach for weighting could be used equally well with any sequence similarity measure, the present problem dictates that only identities be considered; otherwise part of the data for deriving contact pairs would have been discarded.

The number of protein subunits chosen according to the criteria described in the previous section is more than 1600. To reduce computational time expediently for the calculation of the sequence identity matrix and its diagonalization, proteins that have more than 0.95 sequence identity are regarded as the same sequence and are removed from the sequence identity matrix. Then a statistical weight for each of the individual proteins is taken to be w_μ / m, where w_μ is the weight for that protein family μ, and m is the number of members in the family. Protein representatives for this purpose are chosen in the following procedures. (1) If a protein is less similar than 0.95 sequence identity to any protein representative already chosen, this protein is regarded as a new protein. (2) The above procedure is iterated until all proteins are examined. There are 424 protein representatives remaining with less than 0.95 sequence identity to any other in the PDB, and the total effective number of proteins is found with equation (52) to be 251.

References

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Bowie, J. U., Lüthy, 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., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Computer Chem.* **4**, 187–217.

Bryant, S. H. & Lawrence, C. E. (1993). An empirical energy function for threading protein sequence through the folding motif. *Proteins: Struct. Funct. Genet.* **16**, 92–112.

Bryngelson, J. D. & Wolynes, P. G. (1987). Spin glasses and the statistical mechanics of protein folding. *Proc. Natl Acad. Sci. USA*, **84**, 7524–7528.

Bryngelson, J. D., Onuchic, J. N., Socci, N. D. & Wolynes, P. G. (1995). Funnels, pathways and the energy landscape of protein folding: a synthesis. *Proteins: Struct. Funct. Genet.* **21**, 167–195.

Covell, D. G. & Jernigan, R. L. (1990). Conformations of folded proteins in restricted spaces. *Biochemistry*, **29**, 3287–3294.

Crippen, G. M. (1991). Prediction of protein folding from amino acid sequence over discrete conformation spaces. *Biochemistry*, **30**, 4232–4237.

Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1978). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure 1978* (Dayhoff, M. O., ed.), vol. 3. suppl. 5, pp. 345–352, National Biomedical Research Foundation, Washington, DC.

Eriksson, A. E., Baase, W. A., Zhang, X.-J., Heinz, D. W., Blaber, M., Baldwin, E. P. & Matthews, B. W. (1992). Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science*, **255**, 178–183.

Fauchère, V. L. & Pliška, V. (1983). Hydrophobic parameters π of amino acid side-chains from the partitioning of *N*-acetyl-amino-acid amides. *Eur. J. Med. Chem.* **18**, 369–375.

Fischer, D., Tsai, C. J., Nussinov, R. & Wolfson, H. J. (1996). A 3-D sequence independent representation of the protein databank. *Protein Eng*., in the press.

Go, N. (1983). Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.* **12**, 183–210.

Gutin, A. M., Badretdinov, A. Y. & Finkelstein A. V. (1992). Why are the statistics of globular protein structures Boltzmann-like? *Mol. Biol. (USSR)*, **26**, 94–102.

Hendlich, M., Lackner, P., Weitckus, S., Floechner, H., Froschauer, R., Gottsbachner, K., Casari, G. & Sippl, M. J. (1990). Identification of native protein folds amongst a large number of incorrect models; the calculation of low energy conformations from potentials of mean force. *J. Mol. Biol.* **216**, 167–180.

Hill, T. L. (1960). *Statistical Mechanics*. Addison-Wesley, Reading, MA.

Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992). Selection of representative protein data sets. *Protein Sci.* **1**, 409–417.

Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature*, **358**, 86–89.

Lee, B. (1993). Estimation of the maximum change in stability of globular proteins upon mutation of a hydrophobic residue to another of smaller size. *Protein Sci.* **2**, 733–738.

Lüthy, R., Bowie, J. U. & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature*, **356**, 83–85.

Maiorov, V. N. & Crippen, G. M. (1992). Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol.* **227**, 876–888.

Matsumura, M., Becktel, W. J. & Matthews, B. W. (1988). Hydrophobic stabilization in T4 lysozyme determined directly by multiple substitutions of Ile3. *Nature*, **334**, 406–410.

Miyazawa, S. & Jernigan, R. L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, **18**, 534–552.

Miyazawa, S. & Jernigan, R. L. (1994). Protein stability for single substitution mutants and the extent of local compactness in the denatured state. *Protein Eng.* **7**, 1209–1220.

Needleman, S. B. & Wunsch, C. B. (1970). A general method applicable to the search for similarities in the amino acid sequences of two proteins. *J. Mol. Biol.* **48**, 443–453.

Nicholls, A., Sharp, K. A. & Honig, B. (1991). Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins: Struct. Funct. Genet.* **11**, 281–296.

Nishikawa, K. & Matsuo, Y. (1993). Development of pseudoenergy potentials for assessing protein 3-D–1-D compatibility and detecting weak homologies. *Protein Eng.* **6**, 811–820.

Novotny, J., Bruccoleri, R. E. & Karplus, M. (1984). An analysis of incorrectly folded protein models; implications for structure predictions. *J. Mol. Biol.* **177**, 787–818.

Nozaki, Y. & Tanford, C. (1971). The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxine solutions; establishment of a hydrophobicity scale. *J. Biol. Chem.* **246**, 2211–2217.

Orengo, C. A., Flores, T. P., Taylor, W. R., & Thornton, J. M. (1993). Identification and classification of protein fold families. *Protein Eng.* **6**, 485–500.

Pace, C. N. (1992). Contribution of the hydrophobic effect to globular protein stability. *J. Mol. Biol.* **226**, 29–35.

Săli, A., Shakhnovich, E. & Karplus, M. (1994). Kinetics of protein folding: a lattice model study of the requirements for folding to the native state. *J. Mol. Biol.* **235**, 1614–1636.

Sharp, K. A., Nicholls, A., Friedman, R. & Honig, B. (1991). Extracting hydrophobic free energies from experimental data: relationship to protein folding and theoretical models. *Biochemistry*, **30**, 9686–9697.

Shortle, D., Sites, W. E. & Meeker, A. K. (1990). Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease. *Biochemistry*, **29**, 8033–8041.

Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. *J. Mol. Biol.* **213**, 859–883.

Sippl, M. J. & Weitckus, S. (1992). Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. *Proteins: Struct. Funct. Genet.* **13**, 258–271.

Yutani, K., Ogasawara, K., Tsujita, T. & Sugino, Y. (1987). Dependence of conformational stability on hydrophobicity of the amino acid residue in a series of variant proteins substituted at a unique position of tryptophan synthase α subunit. *Proc. Natl Acad. Sci. USA*, **84**, 4441–4444.

Edited by B. Honig

(Received 17 April 1995; accepted in revised form 17 November 1995)