Number of residues in a sphere around a certain residue can be used as a hydrophobic penalty function of proteins

Kazunori Toma

Computer Science Department, Asahi Chemical Industry Co. Ltd., Samejima, Fuji, Shizuoka, Japan

A novel hydrophobic penalty function of proteins is proposed and assessed with several test cases. The number of residues in a defined sphere around a certain residue is averaged over the data set proteins. Differences between the standard values thus obtained and calculated values are summed up, residue by residue, with the weight of standard deviations to give the penalty value. This penalty function is applied to the structures of randomly shuffled sequences, incorrectly folded structures and partially denatured structures displayed on a graphics terminal, and is shown to discriminate the native structure from others fairly well, although the present parameter set is tuned for proteins of about 100–150 residues. From the results of present study and the known correlation with other hydrophobic parameters of amino acids, the penalty function can be considered as a practical amino acid residue-level hydrophobic penalty function.

Keywords: protein structure, α-carbon model, hydrophobic penalty function, residues in a sphere

INTRODUCTION

The prediction of a protein's structure from its amino acid sequence is one of the unsolved fundamental problems in molecular biology. While the demand for such a method is increasing, as more and more amino acid sequences are deduced from nucleic acid sequences, the problem has remained unanswered for about three decades. Although there have been many approaches, one of the crucial obstacles at present is the lack of proper potential functions to evaluate the protein folding.[1]

Detailed, empirical potential energy functions have been developed for proteins and peptides. Currently, these are used mainly for structure, dynamical and free energy calculations around the native structure of proteins.[2] Although studies on folding have been performed with these potential energy functions, their success in predicting protein structures has been rather limited because of the so-called multiple-minima problem and because of the enormous computational time required. These problems seem to be very hard to overcome at present in spite of efforts by Scheraga and his coworkers.[3,4]

While the potential functions mentioned above are powerful tools for studying the detailed structure of proteins, other classes of potential functions or model systems have also been proposed to surmount the problems mentioned above. These approaches involve some simplification of protein structures in their formulations, which has the advantage of reducing both the dimensionalities of problems and the required computational time. Among these approaches, lattice models are rather successful in describing fundamental aspects of protein folding;[5–7] however, they are so oversimplified that it is often too difficult to reconstruct realistic models of proteins.

Because it is very natural to represent each amino acid residue as a single unit, another class of simplified formulations uses an empirical potential function based on such models as an α-carbon model. This type of study was first reported more than a decade ago,[8–10] as computers become more powerful, it is appropriate to reconsider such older approaches.[11,12] Again, the most fundamental problem of this approach is that no proper potential function is known. The problem is so deep that the best result obtained by this type of approach comes no closer to the native structure than a 4.0-Å root-mean-square displacement.[11] This level of accomplishment simply means that the resultant structure is somewhat globular.[13]

One possibly misleading formulation comes from the idea of a pairwise potential between amino acid residues. Although it is tempting to postulate a pairwise potential between amino acid units (like those used to model atomic potentials), *if we employ an α-carbon system it is difficult* to clarify physical origin of such a pairwise potential. There-

Address reprint requests to Kazunori Toma at the Computer Science Department, Asahi Chemical Industry Co., Ltd., 2–1, Samejima, Fuji, Shizuoka 416, Japan.

Received 26 June 1990; accepted 10 July 1990

fore, in this paper a type of penalty function, rather than a potential function, is proposed using the number of amino acid residues in a defined sphere around a certain residue; however, this penalty function is also empirical and lacks physical motivation. While Nishikawa and Ooi used this idea with considerable success to predict radial distributions of amino acids from sequence information,[14,15] a direct application in the three-dimensional (3D) context is examined in this paper. In the following formulation, the fundamental nature of this penalty function and possible directions of future developments are discussed.

METHODS

The residues in a sphere (RIS) of a given size are calculated as the number of residues in a sphere of defined radius around a given residue. This kind of value is traditionally called a contact number, because it approximates the number of residues in contact if an appropriate radius is chosen to define the sphere.[16] However, as this paper is not concerned with the problem of residues being in contact, this quantity will be referred to as RIS for the sake of clarity. The position of each residue was represented by the α-carbon coordinate. All residues other than the central residue were counted. Radii from 6 Å to 14 Å, with 1 Å increments, were used to generate standard RIS values. The standard RIS value for each amino acid is defined as an average of real values of all the proteins in the data set. Standard deviations (SD) for each standard value were also calculated. In the following discussion, RIS06 denotes the RIS for a 6-Å radius; RIS14 denotes the RIS for a 14-Å radius; and so on. The penalty value for each radius is defined as the sum of the absolute values of the difference between the real RIS and the averaged RIS for each amino acid, divided by the SD:

$$\text{Penalty value} = \Sigma\ (|\text{RIS(real)} - \text{RIS(standard)}|/\text{SD})$$

In the following discussion, this value is further divided by the residue number to eliminate protein size dependence, and the resultant value is called the "RIS penalty value."

The following 27 protein structures having better than 1.7-Å resolution from the Protein Data Bank (PDB)[17] were used to generate average and standard deviation values for each RIS; actinidin (PDB entry name 2ACT, 218 residues), α-lytic protease (2ALP, 198), acid proteinase (3APR, 325), azurin (2AZA, 129), cytochrome C′ (2CCY, 127), cytochrome C3D (2CDV, 107), chymotrypsin (4CHA, 239), carboxypeptidase A (5CPA, 307), parvalbumin (1CPV, 108), cytochrome C peroxidase (2CYP, 293), cytochrome C (3CYT, 103), dihydrofolate reductase (3DFR, 162), dihydrofolate reductase (4DFR, 159), erythrocruorin (1ECD, 136), elastase (3EST, 240) ferredoxin (4FD1, 106), glutathione reductase (3GRS, 461), insulin (1INS, 51), lysozyme (2LZM, 164), myohemerythrin (2MHR, 118), ovomucoid 3rd domain (2OVO, 56), papain (9PAP, 212), plastocyanin (1PCY, 99), trypsin inhibitor (4PTI, 58), proteinase B (3SGB, 185), thermolysin (3TLN, 316) and cytochrome C551 (351C, 82). In the case of dimeric proteins, the first monomer was used, and the first conformation was used if there were alternate conformations. Groups other than amino acid residues (like hemes) were ignored.

The shuffled sequences of 3EST and 4PTI were made by randomly changing the order of amino acids while keeping the original composition; the same 3D structure was used for all sequences in RIS penalty value calculations. Incorrectly folded examples were taken from the literature,[18,19] although the original PDB α-carbon coordinates of hemerythrin (1HMQ) and immunoglobulin VL domain (1MCP, 1–113 residues of the first chain considered) were used without any structural change. Imaginary partially unfolded structures were constructed from the native structure of 3EST on an E&S PS340 graphics terminal using the graphics program ALPHA.[20] These structures correspond to unfolded substructures around one or two residue points. In the tables and the figures, the native structures are implied by PDB, and other structures are tentatively termed SH*i*, for the shuffled sequences, or DN*i*, for the handmade denatured structures.

RESULTS

Standard values

The average RIS values and SDs of amino acids with radii 6–14 Å, in 1-Å increments, are summarized in Table 1. For a defining radius of 5 Å all of the amino acids had standard RIS values of about 2.5. Spheres more than 14 Å in radius exceed the size of most of data set proteins, and had RIS values similar to those of RIS14. Therefore, radii in the range 6–14 Å are shown in the table. Because the SDs are rather large, decimal fractions of averaged values are meaningless in any physical sense, and these numbers are shown only to indicate that they are the actual numbers used to calculate the penalty values. The values listed in the table did not change very much, even if half of a data set was used to evaluate them.

It is clear that the hydrophobic residues have larger RIS values than the hydrophilic residues. However, the sphere-size dependence of the RIS values can also be seen. Among them, the behaviors of *Cys* and *Pro* are rather unique, seen most clearly in RIS06. *Cys* gives the largest value and *Pro* the smallest only at this point, indicating that the RIS of small spheres yields packing information rather than hydrophobic information. Thus each point has a different structural meaning, and it was concluded that all radial points should be included in the assessment of the penalty function.

Standard deviations in the table may give some idea about the location of each amino acid in proteins. Because *Gly* and *Ala* have almost average values and always give the largest SDs among amino acids, they must be located rather uniformly. By contrast, the smallest RIS values and SDs are for *Lys*, indicating that it is located strictly on the outside of proteins. This result suggests that the SD must also be incorporated in the penalty value, which can be done as mentioned in the method section. With this scaling, RIS values of different radial size can be compared directly.

Residues-in-sphere penalty values are calculated for data set proteins to illustrate the general behavior, and Table 2 summarizes the results in order of protein size. The amplitudes of RIS penalty values for spheres of different sizes are roughly comparable for a specific protein. Although the amino acid composition of a protein and its shape must have

Table 1. Standard RIS values of each amino acid at various sphere sizes. Average values are shown in the first row and standard deviations are in the second row for each amino acid. Numbers in the second column show the sample number in the data set

	NO.	RIS06	RIS07	RIS08	RIS09	RIS10	RIS11	RIS12	RIS13	RIS14
ALA	424	5.712	8.175	10.024	13.559	18.014	23.325	28.384	34.594	41.425
		1.836	2.469	3.077	4.473	6.119	8.167	10.136	12.652	15.139
ARG	157	5.389	7.758	9.541	12.752	17.274	22.382	27.726	33.707	40.561
		1.666	2.146	2.651	3.698	5.205	6.741	8.369	10.106	12.195
ASN	241	5.162	7.407	9.137	12.124	15.921	20.668	25.581	31.075	37.041
		1.560	2.099	2.767	3.879	5.058	6.812	8.420	9.920	11.659
ASP	275	5.211	7.236	8.702	11.724	15.385	20.244	24.978	30.556	36.607
		1.665	2.359	2.924	3.982	5.454	7.217	8.689	10.424	12.878
CYS	110	6.109	8.745	11.155	15.455	20.200	26.436	31.918	38.355	45.364
		1.610	2.143	2.442	3.363	5.058	6.649	8.295	10.495	12.950
GLN	167	5.186	7.527	9.126	12.048	16.102	21.293	26.192	32.437	39.036
		1.491	2.147	2.682	3.491	5.104	7.058	8.845	10.952	13.295
GLU	220	5.182	7.445	8.841	11.755	15.586	20.709	25.541	31.305	37.523
		1.582	2.147	2.529	3.498	4.998	6.711	8.618	10.499	12.660
GLY	473	5.609	8.015	10.068	13.552	18.042	23.135	28.573	34.928	42.101
		2.016	2.657	3.606	5.020	6.907	9.056	11.199	13.977	16.786
HIS	101	5.644	8.149	10.099	13.604	18.436	24.069	29.663	36.426	44.158
		1.533	2.026	2.431	3.532	5.374	7.365	9.104	10.905	13.432
ILE	239	5.569	8.983	11.439	15.218	21.071	28.335	35.297	43.071	51.213
		1.553	2.012	2.600	3.419	4.949	6.966	8.920	11.135	13.769
LEU	312	5.712	8.686	11.038	14.808	20.298	27.090	33.913	41.385	49.308
		1.619	2.205	2.620	3.528	5.110	7.016	9.063	11.313	13.977
LYS	292	5.065	7.267	8.733	11.500	15.325	19.795	24.479	29.839	35.548
		1.539	1.902	2.159	2.892	4.073	5.529	6.773	8.156	9.936
MET	78	5.487	8.590	10.256	14.051	19.487	26.128	32.231	39.923	47.590
		1.672	2.253	2.665	3.552	5.325	7.338	9.155	11.547	14.247
PHE	190	5.811	8.516	10.468	14.274	19.589	26.453	32.911	40.542	48.842
		1.578	1.890	2.269	3.380	4.940	6.788	8.784	10.881	13.440
PRO	181	4.376	6.956	9.022	12.282	16.309	20.945	25.967	31.768	38.155
		1.667	2.330	3.142	4.214	5.789	7.732	9.627	11.543	13.920
SER	332	5.223	7.654	9.503	12.774	16.997	22.370	27.352	33.012	40.178
		1.876	2.568	3.278	4.662	6.553	8.971	10.956	12.853	16.021
THR	333	5.177	7.997	10.027	13.153	17.802	23.441	28.760	34.949	41.784
		1.610	2.239	2.856	3.919	5.367	7.373	9.356	11.583	14.334
TRP	82	5.683	8.744	10.780	14.622	20.573	27.537	34.061	41.707	49.976
		1.498	2.071	2.514	3.609	5.116	6.965	9.336	11.452	14.129
TYR	199	5.704	8.332	10.558	14.206	19.095	25.508	31.005	37.633	44.668
		1.696	2.220	2.669	3.785	5.021	6.451	7.829	9.788	11.606
VAL	353	5.680	8.955	11.275	14.895	20.221	27.326	33.643	41.105	48.816
		1.425	2.162	2.622	3.660	5.186	7.465	9.655	12.118	14.817
ALL	4759	5.427	8.026	9.966	13.349	17.942	23.622	29.117	35.528	42.538
		1.706	2.336	2.964	4.116	5.795	7.900	9.887	12.146	14.730

some influence on the result, the present standard value seems to be tuned for about 100–150 residue proteins, as proteins of that size generally show smaller values compared with larger or smaller proteins.

Randomly shuffled sequences

Because 3EST and 4PTI are not optimally sized proteins for the present standard values (as discussed in the former section), they are selected from the sample set proteins as rather severe examples for the following test. Using the original structure, native and randomly shuffled amino acid sequences were evaluated with the RIS penalty function. The sequences used are shown in Figures 1 and 2 and the results are summarized in Tables 3 and 4 for 3EST and 4PTI, respectively.

For proteins the size of 3EST, RIS penalty values for all spheres clearly discriminate the native structure from randomly shuffled sequence structures. This relative success seems to be related to sequence homology, because the SH1

Table 2. **RIS penalty values for data set proteins in the order of residue number shown in the second column**

	NO.	RIS06	RIS07	RIS08	RIS09	RIS10	RIS11	RIS12	RIS13	RIS14
1INS	51	0.895	0.872	0.810	0.853	0.897	0.896	1.009	1.044	1.092
2OVO	56	0.986	1.047	0.962	0.894	0.884	0.888	0.903	0.971	0.987
4PTI	58	0.813	0.852	0.863	0.803	0.778	0.814	0.801	0.840	0.871
351C	82	0.783	0.764	0.753	0.771	0.726	0.681	0.710	0.736	0.726
1PCY	99	0.736	0.751	0.702	0.627	0.604	0.615	0.617	0.629	0.623
3CYT	103	0.889	0.787	0.762	0.715	0.597	0.589	0.578	0.591	0.601
4FD1	106	0.744	0.786	0.882	0.807	0.795	0.793	0.802	0.754	0.709
2CDV	107	0.907	0.895	0.918	0.941	0.843	0.888	0.883	0.897	0.951
1CPV	108	0.771	0.713	0.624	0.634	0.548	0.549	0.517	0.534	0.545
2MHR	118	0.677	0.662	0.705	0.692	0.715	0.680	0.681	0.667	0.676
2CCY	127	0.818	0.652	0.649	0.684	0.684	0.647	0.629	0.614	0.602
2AZA	129	0.821	0.793	0.759	0.775	0.770	0.681	0.675	0.654	0.636
1ECD	136	0.707	0.588	0.684	0.678	0.692	0.684	0.636	0.635	0.629
4DFR	159	0.793	0.847	0.816	0.733	0.736	0.749	0.726	0.712	0.701
3DFR	162	0.772	0.851	0.791	0.784	0.742	0.770	0.738	0.723	0.724
2LZM	164	0.665	0.654	0.653	0.630	0.638	0.631	0.612	0.617	0.618
3SGB	185	0.891	0.933	0.953	0.907	0.970	0.961	0.962	0.918	0.928
2ALP	198	0.853	0.948	0.967	0.956	1.007	0.986	0.969	0.936	0.942
9PAP	212	0.787	0.810	0.813	0.820	0.850	0.880	0.847	0.847	0.823
2ATC	218	0.832	0.824	0.852	0.840	0.877	0.879	0.842	0.846	0.841
4CHA	239	0.781	0.851	0.940	0.961	0.952	0.940	0.926	0.925	0.924
3EST	240	0.815	0.944	0.964	0.963	0.989	0.990	0.959	0.969	0.972
2CYP	293	0.771	0.684	0.690	0.748	0.760	0.791	0.812	0.813	0.796
5CPA	307	0.801	0.799	0.811	0.872	0.860	0.845	0.873	0.886	0.874
3TLN	316	0.906	0.864	0.834	0.890	0.890	0.858	0.889	0.893	0.878
3APR	325	0.788	0.791	0.828	0.731	0.745	0.716	0.752	0.770	0.767
3GRS	461	0.774	0.808	0.813	0.780	0.776	0.813	0.821	0.810	0.803

```
     1        10        20        30        40        50        60
PDB  VVGGTEAQRNSWPSQISLQYRSGSSWAHTCGGTLIRQNWVMTAAHCVDRELTFRVVVGEH
SH1  VVNRGEQYVLKINIGIGWTTYTGSEYGGDYSVITIQGPALTSNLTVTGLSLSVWANVFVG
SH2  SVLQSVGTQSVTDRHKVVANAANVSVNFGVYAIQQTRLADGIVRQNSSLTESTSQQAVLN
SH3  MASTPLPNIYTITWRAARQGRSLDTVVVNWCQQGQSQYVRNRRYQQIGEAEVTNATTTLD

              70        80        90       100       110       120
PDB  NLNQNNGTEQYVGVQKIVVHPYWNTDDVAAGYDIALLRLAQSVTLNSYVQLGVLPRAGTI
SH1  YSDTCCNGLIVLWWGVQVTRLAVTARTVSCHVTRHASTEHHSQVRQYGEVDTCDAVQNQL
SH2  HYTSNYDNRAWLYIPTQLTVYLRQSLYVSGRVVWGHEGETGNIWCTCNLVPDFRHWASGT
SH3  IAYSTVTSCQSWVTNSVDLQVIGGSLDLNTVVWAFGSSVQNYAIMAVPRVRGVVTGLAHQ

            130       140       150       160       170       180
PDB  LANNSPCYITGWGLTRTNGQLAQTLQQAYLPTVDYAICSSSSYWGSTVKNSMVCAGGDGV
SH1  YRTSNRFALQNSNILGSVAPAVGFLAQAHLQCSPNTKMYNSADSVRYSANLAAGQLSGPK
SH2  VTFAYGVRGQVNTILTLKNQCEACWLGVISYQRRNSAAAGGGGKIWDMYCALTIPSGNNN
SH3  VCCGHTSTWGKGFKSDANVYGLNSVRAELSTVWGCLYLLNVNSVLSGTGNCASPQEHQDC

            190       200       210       220       230       240
PDB  RSGCQGDSGGPLHCLVNGQYAVHGVTSFVSRLGCNVTRKPTVFTRVSAYISWINNVIASN
SH1  CSNQQIRPHSVGGNSYGGVCYSINRNVCIPGRGSVWTTNLQWVAWDGLVTVGAQDRPLMG
SH2  MVPVVPVVGPYWGTSNNGSVGPLQGGCSLALYGSRGLCCDGSSQARHAVIDVQHTTLITS
SH3  ACLVSLWYGANYGQVNTNPNNVGHLSFSTHSAGGTVLQYPRKRGALGHSIRIPGVGDIGY
```

Figure 1. 3EST native and shuffled sequences

```
     1        10        20        30        40        50
PDB  RPDFCLEPPYTGPCKARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCMRTCGGA
SH1  SCFLAGFTPYFCFGRNKRPIYCPERANCRYQAGDRRNAELGKYGTGMAVKDAKPCCTL
SH2  MCGGINCPCPLGTCKEGSTRYAFFINGYKRQLTFEKYNVPRGYCACARDAKARPDAFR
SH3  ADTNPYAPCFGCAAGFCMPSYITCRGCRGEQRFGVRYEDTKRPYLGAIFNARKCKNLK
```

Figure 2. 4PTI native and shuffled sequences

Table 3. RIS penalty values for 3EST structures of native and shuffled sequences. Differences from the native sequence structures are shown in the second row of each sphere size

Sequence	PDB	SH1	SH2	SH3
RIS06	0.815	0.870	0.853	0.869
		55	38	54
RIS07	0.944	0.969	0.970	0.998
		25	26	54
RIS08	0.964	1.026	1.040	1.064
		62	76	100
RIS09	0.963	1.039	1.056	1.077
		76	93	114
RIS10	0.989	1.052	1.083	1.111
		63	94	122
RIS11	0.990	1.073	1.088	1.125
		83	98	135
RIS12	0.959	1.042	1.053	1.089
		83	94	130
RIS13	0.969	1.050	1.053	1.105
		81	84	136
RIS14	0.972	1.047	1.043	1.096
		75	71	124

Table 4. RIS penalty values for 4PTI structures of native and shuffled sequences. Differences from the native sequence structures are shown in the second row of each sphere size

Sequence	PDB	SH1	SH2	SH3
RIS06	0.813	0.871	0.864	0.919
		58	51	106
RIS07	0.852	0.926	0.945	0.915
		74	93	63
RIS08	0.863	0.890	0.930	0.893
		27	67	30
RIS09	0.803	0.870	0.895	0.880
		67	92	77
RIS10	0.778	0.803	0.819	0.828
		25	41	50
RIS11	0.814	0.878	0.877	0.868
		64	63	54
RIS12	0.801	0.827	0.856	0.851
		26	55	50
RIS13	0.840	0.850	0.825	0.833
		10	−15	−7
RIS14	0.871	0.878	0.858	0.854
		7	−13	−17

Table 5. RIS penalty values for native and incorrectly folded structures of 1MCP and 1HMQ. Differences from the native sequence structures are shown in the second row of each sphere size

Sequence structure	1MCP 1MCP	1MCP 1HMQ	1HMQ 1HMQ	1HMQ 1MCP
RIS06	0.679	0.787	0.745	0.759
		108		14
RIS07	0.786	0.752	0.704	0.944
		−34		240
RIS08	0.711	0.828	0.745	0.934
		117		189
RIS09	0.649	0.826	0.770	0.846
		177		76
RIS10	0.680	0.816	0.735	0.907
		136		172
RIS11	0.717	0.797	0.709	0.940
		80		231
RIS12	0.672	0.793	0.736	0.916
		121		180
RIS13	0.720	0.764	0.718	0.911
		44		193
RIS14	0.714	0.730	0.691	0.898
		16		207

sequence that gives the smallest difference in RIS penalty values is the most similar to the native sequence. This fact indicates that RIS can be used as a guide function to evaluate the validity of a postulated 3D structure for a given sequence.

At RIS13 and RIS14, SH2 and SH3 sequences of 4PTI give better values than the native sequence. The observed difficulty in discrimination can be understood in terms of the protein-size dependence of RIS. Average RIS values of all residues at different radii of 4PTI and other small proteins deviate considerably from the average of all proteins at large sphere size. This fact can be explained more simply in terms of amino acid volume. Because an amino acid residue has an average volume of about 100 Å^3,[21] even a sphere radius of 12 Å exceeds the size of small proteins like 4PTI, and the evaluation with RIS13 and RIS14 must be nonsense.

Incorrectly folded structures

We chose hemerythrin, a four α-helical bundle protein, and the VL domain of immunoglobulin, a two four-strand β-sheet protein, as similarly sized structures having no sequence homology. These proteins were chosen as the basis of the incorrectly folded proteins, which were made by exchanging sequences between them. They were used to evaluate the empirical potential functions by Novotný *et al*,[18,19] who showed that the incorporation of the solvent effect was necessary to discriminate between the native structure and incorrectly folded models. Therefore, this example is deemed as a good test of how well the RIS penalty function represents the solvent effect.

As expected from the hydrophobic nature of the RIS penalty function, the resultant penalty values for almost all radii clearly distinguish native structures from incorrectly folded ones, as shown in Table 5. The distinction is clearer for large radii than for small ones. Because amino acid sequences themselves are native, the RIS values for small spheres may share some problems with the atomic level empirical potential energy function, in that they are geared for local conformation and they cannot distinguish these structures without the solvent effect. The success of large-sphere RIS values warrants that they can model the solvent effect of hydrophobic nature. If we have the proper folding pattern database, this method can be generalized to the assessment of new sequences with known structures.

Partially denatured structures

Arbitrarily constructed denatured structures of 3EST are shown in the Color Plates. Again 3EST was chosen for its poor performance in Table 2. Structures DN1–DN6 were made by extending two parts of a structure to the opposite side around a single residue point; DN7 and DN8 were made by further extending DN4 at other points. The RIS penalty values calculated for each structure are summarized in Table 6. This seems to be the toughest test applied to the RIS penalty function in this paper, because there are some values that are smaller than those of the native structure. The RIS values for spheres larger than 10 Å in radius consistently distinguish the native structures from others. Because almost all of the local structures are conserved in those denatured structures, it is not so surprising that small-sphere RIS values fail to give the smallest penalty for the native structure in several cases.

One remarkable exception is DN4, which almost always gives a smaller penalty value than the native structure. The DN4 structure was constructed by rotating two substructures separated almost at the center of sequence (118th residue);

Table 6. RIS penalty values for native and partially denatured structures of 3EST. Differences from the native sequence structures are shown in the second row of each sphere size

Structure	PDB	DN1	DN2	DN3	DN4	DN5	DN6	DN7	DN8
RIS06	0.815	0.808	0.858	0.847	0.883	0.898	0.895	0.908	0.988
		−7	43	32	68	83	80	93	173
RIS07	0.944	0.941	0.944	0.934	0.906	0.955	0.909	0.913	1.009
		−3	0	−10	−38	11	−35	−31	65
RIS08	0.964	0.956	0.970	0.958	0.914	0.964	0.937	0.941	1.024
		−8	6	−6	−50	0	−27	−23	60
RIS09	0.963	0.958	0.970	0.965	0.917	0.975	0.993	0.979	1.050
		−5	7	2	−46	12	30	16	87
RIS10	0.989	0.991	0.999	1.003	0.910	0.962	1.067	0.953	1.026
		2	10	14	−79	−27	78	−36	37
RIS11	0.990	1.003	1.005	1.018	0.912	1.000	1.066	1.018	1.039
		13	15	28	−78	10	76	28	49
RIS12	0.959	0.985	0.973	1.001	0.873	0.988	1.075	1.046	1.022
		26	14	42	−86	29	116	87	63
RIS13	0.969	1.000	0.995	1.027	0.906	1.024	1.088	1.104	1.051
		31	26	58	−63	55	119	135	82
RIS14	0.972	1.000	0.994	1.022	0.886	1.037	1.086	1.129	1.046
		28	22	50	−86	65	114	157	74

the two substructures actually correspond to the domains of elastase.[22] Because domain is known as a rather independent folding structural unit and the present standard value is geared for protein structures about this size (as shown in Table 2), this ill-fated result must be considered as a reasonable consequence of the multiple-layer architecture. Further unfolding of DN4 creates structures that are distinguishable with RIS penalty values, such as DN7 and DN8.

DISCUSSION

The RIS standard values exhibit hydrophobic tendencies as shown in Table 1. The RIS penalty function can easily discriminate the incorrectly folded proteins, indicating it can model the solvent effect. Earlier works have demonstrated the hydrophobic nature of this type of function. Jernigan and his coworkers claimed that even the contact number of amino acids used in a pairwise way was of hydrophobic nature.[13,16] Nishikawa and Ooi employed RIS08 or RIS14 to predict rather successfully radial distributions of amino acid residues from sequence.[14,15] Although the formulations and applications are different, local contacts and radial distributions have much in common with the hydrophobic nature of amino acid residues. Also, the RIS value has been shown to correlate well with other hydrophobic parameters,[23] such as those proposed by Rose and Roy[24] and by Kyte and Doolittle.[25] Thus, the RIS penalty function can be deemed a promising hydrophobic penalty function of the amino acid residue level.

This penalty function may be the first example of a nonpairwise formulation of hydrophobic character, and may overcome some problems that are seen in the usual pairwise treatment. At the least this treatment clarifies the meaning of the standard deviation. It may also apply to the general model of hydrophobic interactions in which hydrophobic groups expel water molecules and move closer together. Although there has been some discussion that the old idea of hydrophobic interaction is entirely wrong,[26] the fact still remains that the hydrophobic residues gather on the inside of proteins.[27]

The present results show that the RIS value can be used as the penalty function to evaluate a protein structure. If we have a new amino acid sequence, we can compare it with the known 3D structures kept as a folding-type database and determine the most probable structure for the given sequence. We can do this type of database search by RIS penalty values as seen in the evaluation of incorrectly folded proteins. However care must be taken to avoid considering a single RIS value as sufficient to evaluate all incorrect structures. The RIS values for larger spheres seem to be better suited to this purpose. There are other more general problems, of course, such as insertions and deletions, and there is little hope of solving a novel folding solely by this method.

There are several other important points to consider in using the RIS penalty function: protein sizes (residue number), their shapes and natures (e.g., dimeric, membrane or fibrous). Of course, these are the indigenous problems of protein structure predictions in general. Among these, however, the size dependence of the penalty value is clear from the present study, and must be a primary target of improvement in the future development of the RIS penalty function.

Although the RIS values of small spheres give some information on local packing, they are rather specific to *Cys* and *Pro* and the discriminatory power of small-size RIS values is rather weak, as shown in the tables. Therefore, local structural information like secondary structures can be incorporated without serious conflict with the present penalty function. While this paper does not discuss the process of coordinate generation and folding simulation, but rather is limited to the fundamental nature of the RIS penalty function, coordinates can be generated independently from the penalty function in several ways. Because a systematic search of candidate conformations may make the fundamental problems of the RIS penalty function clearer, a Monte Carlo-type search for possible low-penalty-value structures is underway using the penalty function as a guide with local coordinate generation techniques.

ACKNOWLEDGEMENT

The author thanks Drs. N. Gō and J.-F. Gibrat for their comments and discussion.

REFERENCES

1 Gō, N. Theoretical studies of protein folding. *Ann. Rev. Biophys. Bioeng.* 1983, **12**, 183–210

2 van Gunsteren, W.F. The role of computer simulation techniques in protein engineering. *Protein Eng.* 1988, **2**, 5–13

3 Li, Z. and Scheraga, H.A. Monte Carlo minimization approach to the multiple-minima problem in protein folding. *Proc. Natl. Acad. Sci. USA* 1987, **84**, 6611–6615

4 Ripoll, D.R. and Scheraga, H.A. The multiple-minima problem in the conformational analysis of polypeptides. III. An electrostatically driven Monte Carlo method: Test on enkephalin. *J. Protein Chem.* 1989, **8**, 263–287

5 Gō, N., Abe, H., Mizuno, H. and Taketomi, H. In *Protein Folding* (N. Jaenicke, Ed.) Elsevier, Amsterdam (1980) 167–181

6 Skolnick, J. and Kolinski, A. Computer simulations of globular protein folding and tertiary structure. *Ann. Rev. Phys. Chem.* 1989, **40**, 207–235

7 Lau, K.F. and Dill, K.A. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules* 1989, **22**, 3986–3997

8 Levitt, M. and Warshel, A. Computer simulation of protein folding. *Nature* 1975, **253**, 694–698

9 Levitt, M. A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* 1976, **104**, 59–107

10 Kuntz, I.D., Crippen, G.M., Kollman, P.A. and Kimelman, D. Calculation of protein tertiary structure. *J. Mol. Biol.* 1976, **106**, 983–994

11 Wilson, C. and Doniach, S. A computer model to dynamically simulate protein folding: Studies with crambin. *Proteins* 1989, **6**, 193–209

12 Covell, D.G. and Jernigan, R.L. Conformations of folded proteins in restricted space. *Biochem.* 1990, **29**, 3287–3294

13 Hagler, A.T. and Honig, B. On the formation of protein tertiary structure on a computer. *Proc. Natl. Acad. Sci. USA* 1978, **75**, 554–558

14 Nishikawa, K. and Ooi, T. Prediction of the surface-interior diagram of globular proteins by an empirical method. *Int. J. Peptide Protein Res.* 1980, **16**, 19–32

15 Nishikawa, K. and Ooi, T. Radial locations of amino acid residues in a globular protein: Correlation with the sequence. *J. Biochem.* 1986, **100**, 1043–1047

16 Miyazawa, S. and Jernigan, R.L. Estimation of effective inter-residue contact energies from protein crystal structures: Quasichemical approximation. *Macromolecules* 1985, **18**, 534–552

17 Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F. Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. The protein data bank: A computer based archival file for macromolecular structures. *J. Mol. Biol.* 1977, **112**, 535–542

18 Novotný, J., Bruccoleri, R. and Karplus, M. An analysis of incorrectly folded protein models. Implications for structure predictions. *J. Mol. Biol.* 1984, **177**, 787–818

19 Novotný, J., Rashin, A.A. and Bruccoleri, R.E. Criteria that discriminate between native proteins and incorrectly folded models. *Proteins* 1988, **4**, 19–30

20 Toma, K. Simple protein model building tool. *J. Mol. Graphics* 1987, **5**, 101–102

21 Goldsack, D.E. and Chalifoux, R.C. Contribution of the free energy of mixing of hydrophobic side chains to the stability of the tertiary structure of proteins. *J. Theor. Biol.* 1973, **39**, 645–651

22 Richardson, J.S. The anatomy and taxonomy of protein structure. *Adv. Protein Chem.* 1981, **34**, 167–339

23 Cornette, J.L., Cease, K.B., Margalit, H., Spouge, J.L., Berzofsky, J.A. and DeLisi, C. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.* 1987, **195**, 659–685

24 Rose, G.D. and Roy, S. in Hydrophobic basis of packing in globular proteins. *Proc. Natl. Acad. Sci. USA* 1980, **77**, 4643–4647

25 Kyte, J. and Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 1982, **157**, 105–132

26 Privalov, P.L. and Gill, S.J. The hydrophobic effect: A reappraisal. *Pure Appl. Chem.* 1989, **61**, 1097–1104

27 Lim, W.A. and Sauer, R.T. Alternative packing arrangements in the hydrophobic core of λ repressor. *Nature* 1989, **339**, 31–36