J. Mol. Biol. (1992) **227**, 876–888

Contact Potential that Recognizes the Correct Folding of Globular Proteins

Vladimir N. Maiorov and Gordon M. Crippen

College of Pharmacy, University of Michigan
Ann Arbor, MI 48109, U.S.A.

(Received 3 March 1992; accepted 26 May 1992)

We have devised a continuous function of interresidue contacts in globular proteins such that the X-ray crystal structure has a lower function value than that of thousands of protein-like alternative conformations. Although we fit the adjustable parameters of the potential using only 10,000 alternative structures for a selected training set of 37 proteins, a grand total of 530,000 constraints was satisfied, derived from 73 proteins and their numerous alternative conformations. In every case where the native conformation is adequately globular and compact, according to objective criteria we have developed, the potential function always favors the native over all alternatives by a substantial margin. This is true even for an additional three proteins never used in any way in the fitting procedure. Conformations differing only slightly from the native, such as those coming from crystal structures of the same protein complexed with different ligands or from crystal structures of point mutants, have function values very similar to the native's and always less than those of alternatives derived from substantially different crystal structures. This holds for all 95 structures that are homologous to one or another of various proteins we used. Realizing that this potential should be useful for modeling the conformation of new protein sequences from the body of protein crystal structures, we suggest a test for deciding whether a nearly correct approximation to the native conformation has been found.

Keywords: protein structure prediction; protein folding; amino acid residue contacts; conformational potential functions; globular proteins

1. Introduction

The classical protein folding problem is to predict the three-dimensional conformation of a protein given only its amino acid sequence. Here, we consider a restricted version that we might call the multiple choice "recognition problem": given the amino acid sequence of a protein *and* a large selection of globular conformations that includes the correct native fold, choose the one native conformation. Such a situation naturally arises in attempting to predict a protein's conformation by homology modeling, where there may be several different ways to arrange variable loops. Other applications are the assessment of alternative conformations of a protein derived from nuclear magnetic resonance (n.m.r.†) experiments, or choosing between different chain tracings through the electron density in the early stages of determining a protein's X-ray crystal structure.

A number of different researchers have suggested various criteria for the recognition problem, such as the number of hydrophobic contacts (Bryant & Amzel, 1987). Novotny and co-workers (Novotny *et al.*, 1984, 1988) analyzed the accessible surface area in terms of its polar/apolar ratio and the distribution of this ratio for different amino acid side-chains, as well as atomic packing and empirical energy and free energy functions, in order to differentiate between a few examples of correct *versus* intentionally misfolded structures. Chiche and co-workers related solvation free energy (Eisenberg & McLachlan, 1986) to the correctness of a protein fold using the observed approximately linear dependence of the solvation energy on the protein chain length (Chiche *et al.*, 1990). One of the latest and most successful examples of the three-dimensional profile approach (Lüthy *et al.*, 1992) discriminated between the correct and an incorrect fold for seven different proteins, judging from their relative scores and from the general relation between the scores of correct crystal structures and their chain lengths. Moreover, they were able to detect an incorrectly folded segment in an otherwise correct structure.

In the approaches cited so far, the goal has been

† Abbreviations used: n.m.r., nuclear magnetic resonance; PDB, Brookhaven Protein Data Bank; r.m.s.d., root-mean-square deviation; CTS, complete training set; RTS, reduced training set.

0022-2836/92/190876-13 $08.00/0

to recognize the correct fold as better in some sense than only one or two alternative folds. We believe it is much more difficult to favor the native fold over large numbers of alternatives. Sippl and co-workers (Sippl, 1990; Hendlich *et al.*, 1990) constructed a potential of mean force for the interactions among C^{β} atoms from a survey of protein crystal structures that tended to prefer the native conformation of several proteins over some thousands of alternatives, but not in all cases. In our initial look at the problem (Crippen, 1991), we concluded that a discrete function of interresidue contacts could be constructed for some simple model cases that preferred the native conformation over absolutely all possible alternatives. When it came to extending this to real protein conformations, we produced a discrete contact potential based on the native and alternative conformations of only eight proteins that correctly preferred the native over tens of thousands of alternative for another 37 proteins. However, the remaining 11 proteins in our study were incorrectly predicted. For this level of success, it was important to define a contact in the way reiterated below, and to use relatively few adjustable parameters. In agreement with Sippl, extremely small proteins or oligopeptides, such as avian pancreatic peptide, were consistently difficult to account for, but the remaining erroneous proteins could be treated by including them in the training set, thereby producing a similar number of other proteins that would not fit.

In this study, we have increased the total set of protein crystal structures from 56 to 109, thereby creating a much more difficult fitting problem because each protein is presented with many more alternatives to choose from. Nevertheless, we are able to account for *all* the proteins we examined by learning to identify the kinds of protein native conformations that can be treated this way and by correctly dealing with homologous proteins.

2. Methods

The approach is basically the same as before (Crippen, 1991). Given a protein crystal structure, we note which residues are in contact, according to a carefully chosen definition. The correct crystal structure of a protein is taken to be its native or reference conformation, and many alternative conformations are generated by taking the atomic co-ordinates of all possible contiguous segments of the correct length from all the larger proteins in the data set. In each of these alternatives, there are a different set of contacts, of course, but if the native sequence is imposed on each alternative, we seek some potential function of the contacts that has a lower value for the reference than for any alternative.

(a) *Protein structure data*

The total set of protein crystal structures we considered were the 109 polypeptide chains in the 15 October 1990 release of the Brookhaven Protein Data Bank (PDB) (Abola *et al.*, 1987) with co-ordinates of N, C^{α}, C', C^{β} and O atoms, and no obvious chain breaks in the middle, as in our previous study (Crippen, 1991). Disordered or unre-

Table 1

List of the reference proteins used in this work, sorted by PDB code

PDB code	Resol. (Å)†	No. residues	Chain ID‡	Title and source
155c	2·5	121		Cytochrome c550, *P. denitrificans*
1abp	2·4	306		L-Arabinose-binding protein, *E. coli*
1acx	2·0	108		Actinoxanthin, *A. globisporus*
1bds	—	43		Sea anemone anti-hypertensive anti-viral protein
1bp2	1·7	123		Bovine pancreatic phospholipase A2
1cc5	2·5	83		Cytochrome c5, *Azotobacter*
1ccr	1·5	111		Cytochrome c, rice
1crn	1·5	46		Crambin, Abyssinian cabbage
1cse	1·2	63	I	Eglin C (complexed with subtilisin Carlsberg)
		274	E	Subtilisin Carlsberg (complexed with eglin C)
1ctf	1·7	68		L7/L12 50 S ribosomal protein (C-terminal domain), *E. coli*
1cts	2·7	437		Pig citrate synthase
1cy3	2·5	118		Cytochrome c3, *D. desulfuricans*
1ecd	1·4	136		Hemoglobin (erythrocruorin, deoxy), *C. thummi thummi*
1est	2·5	240		Porcine tosyl-elastase
1fdx	2·0	54		Ferredoxin, *P. aerogenes*
1fx1	2·0	147		Flavodoxin, *D. vulgaris*
1gcn	3·0	29		Porcine glucagon
1gcr	1·6	174		Calf γ-II crystallin
1hip	2·0	85		High potential iron protein (oxidized), *C. vinosum*
1hmg	3·0	175	B	Haemagglutinin, influenza virus
		328	A	
1hmq	2·0	113		Hemerythrin (met), sipunculid worm
1hoe	2·0	74		α-Amylase inhibitor, *S. tendae*
1hvp	—	99		Retrovirus HIV-1 protease
1lh4	2·0	153		Leghemoglobin (deoxy), yellow lupin
1lyz	2·0	129		Hen egg-white lysozyme
1lz1	1·5	130		Human lysozyme
1mba	1·6	146		Sea hare myoglobin
1mbd	1·4	153		Sperm whale myoglobin
1paz	1·55	120		Pseudoazurin (oxidized, Cu^{2+}, *A. faecalis*)
1pcy	1·6	99		Plastocyanin (Cu^{2+}), poplar
1pfk	2·4	320		Phosphofructokinase, *E. coli*
1phh	2·3	394		*p*-Hydroxybenzoate hydroxylase, *P. fluorescens*
1pp2	2·5	122		Calcium-free phospholipase A-2, rattlesnake
1ppt	1·37	36		Avian pancreatic polypeptide
1pyp	3·0	280		Yeast pyrophosphatase
1rei	2·0	107		Human Bence–Jones immunoglobulin variable portion
1rhd	2·5	293		Bovine rhodanese
1rn3	1·45	124		Bovine ribonuclease A
1sn3	1·8	65		Scorpion neurotoxin, variant 3
1tim	2·5	247		Chicken triose phosphate isomerase
1wrp	2·2	102		Bacterial TRP repressor
2abx	2·5	74		α-Bungarotoxin, braided krait venom
2act	1·7	218		Actinidin, kiwi fruit

Table 1 *(continued)*

PDB code	Resol. (Å)†	No. residues	Chain ID‡	Title and source
2alp	1·7	198		α-Lytic protease, *L. enzymogenes*
2aza	1·8	129		Azurin (oxidized), *A. denitrificans*
2b5c	2·0	85		Bovine cytochrome *b5*
2c2c	2·0	112		Cytochrome c2 (oxidized), *R. rubrum*
2cab	2·0	256		Human carbonic anhydrase (form B)
2ccy	1·67	127		Cytochrome *c'*, *R. molischianum*
2cdv	1·8	107		Cytochrome c3, *D. vulgaris*
2cna	2·0	237		Concanavalin A, jack bean
2cyp	1·7	293		Yeast cytochrome *c* peroxidase
2fb4	1·9	216	L	Human immunoglobulin light chain
2gn5	2·3	87		Bacteriophage gene *5* DNA-binding protein
2hhb	1·74	141	A	Human hemoglobin (deoxy)
		146	B	
2lhb	2·0	149		Sea lamprey hemoglobin V (cyano, met)
2lzm	1·7	164		T4 phage lysozyme
2mlt	2·0	26		Bee melittin
2ovo	1·5	56		Ovomucoid (third domain), pheasant
2pab	1·8	114		Human prealbumin
2pka	2·05	80	A	Porcine kallikrein A
		152	B	
2rhe	1·6	114		Human λ immunoglubulin variable domain (Bence–Jones)
2sga	1·5	181		Proteinase A, *S. griseus*
2sns	1·5	141		Staphylococcal nuclease
2sod	2·0	151		Bovine Cu,Zn superoxide dismutase
2ssi	2·6	107		*Streptomyces* subtilisin inhibitor
2stv	2·50	184		Tobacco necrosis virus coat protein
2taa	3·0	478		Taka-amylase A, *A. oryzae*
351c	1·6	82		Cytochrome c551 (oxidized), *P. aeruginosa*
3adk	2·1	194		Porcine adenylate kinase
3ebx	1·4	62		Sea snake erabutoxin B
3fab	2·0	207	L	Human λ immunoglobulin FAB'
		219	H	
3fxc	2·5	98		Ferredoxin, *S. platensis*
3fxn	1·9	138		Flavodoxin (oxidized), *Clostridium*
3gap	2·5	208		Catabolite gene activator protein, *E. coli*
3gpd	3·5	334		Human D-glyceraldehyde-3-phosphate dehydrogenase
3grs	1·54	461		Human glutathione reductase
3icb	2·3	75		Bovine calcium-binding protein
3ins	1·5	21	A	Pig insulin
		30	B	
3pgk	2·5	415		Yeast phosphoglycerate kinase
3rp2	1·9	224		Rat mast cell protease
4ape	2·1	330		Endothiapepsin, fungal
4dfr	1·7	159		Dihydrofolate reductase, *E. coli*
4fd1	1·9	106		*Azotobacter* ferrodoxin
4mdh	2·5	333		Porcine cytoplasmic malate dehydrogenase

Table 1 *(continued)*

PDB code	Resol. (Å)†	No. residues	Chain ID‡	Title and source
4pti	1·5	58		Bovine pancreatic trypsin inhibitor
4rhv	3·0	40	4	Human rhinovirus 14 coat protein
		236	3	
		255	2	
		273	1	
4sbv	2·8	199		Southern bean mosaic virus coat protein
4tln	2·3	316		Bacterial thermolysin
5cpa	1·54	307		Bovine carboxypeptidase A α
5cpv	1·6	108		Carp calcium-binding parvalbumin B
5cyt	1·5	103		Cytochrome *c* (reduced), tuna
5rxn	1·20	54		Rubredoxin (oxidized, Fe^{3+}), *Clostridium*
6ldh	2·0	329		Dogfish lactate dehydrogenase
7api	3·0	36	B	Human modified α-1-antitrypsin
		339	A	
8adh	2·4	374		Horse apo-liver alcohol dehydrogenase
8cat	2·5	498		Bovine catalase
9pap	1·65	212		Papain (Cys25 oxidized), *Papaya*
9wga	1·8	170		Wheat-germ agglutinin

† A – sign denotes n.m.r. (1bds) and model (1hvp) protein structures for which the notion of the resolution is not applicable.

‡ In the case of more than 1 chain in a PDB file, the chain identifiers are given.

solved residues at the N or C termini are not included in the polypeptide chains we consider here. For brevity, we will refer to those chains by their PDB code and the chain identifier in the PDB file (e.g. 3ins.A is the A chain of insulin). The full name of each protein can be found in Table 1. Generally, we included only the accurately determined (≤2·5 Å nominal resolution: 1 Å = 0·1 nm) structures, although some lower-resolution structures, having no interior chain breaks, were included in this study, sometimes to increase the number of alternative conformations we could generate, and sometimes to increase the number of short protein chains considered. We also included 2 other PDB entries that technically did not fulfil the 2·5 Å resolution criterion: 1bds is a structure determined by n.m.r. and distance geometry having unknown accuracy, and 1hvp is a hypothetical conformation built by homology modeling. In the final analysis, these 2 caused no special problems. The 109 protein structures ranged from 21 residues for the shorter insulin chain 3ins.A to 498 residues for 8cat.A. However, we used only the smallest 86 chains as reference structures because these all had 255 or fewer residues. The limit of 255 is due to the database packing scheme we used, where each contact in each alternative encodes its sequence separation in one 8-bit byte. Even so, our total database of all contacts for all 691,165 alternatives of all the reference proteins required a few hundred megabytes of storage. Thus, the 23 largest structures (2cab, 4rhv.1, 1cse.E, 1pyp, 1rhd, 2cyp, 1abp, 5cpa, 4tln, 1pfk.A, 1hmg.A, 6ldh, 4ape, 4mdh.A, 4gpd.G, 7api.A, 8adh, 1phh, 3pgk, 1cts, 3grs, 2tta.A and 8cat.A) were used only for building alternative structures.

The 86 reference structures are those in Table 1 which have chain length less than 256 (also listed in Table 4 in

Table 2

The 19 reference proteins used in the present work and their 95 homologues

No. residues	PDB codes and chain identifiers: Reference†	Homologous‡	r.m.s.d. (Å)‖
58	4pti	5pti	0·59
62	3ebx	5ebx§	0·15
82	351c	451c	0·03
99	1pcy	2pcy 3pcy 4pcy 5pcy 6pcy	0·12
106	4fd1	1fd2 2fd2	0·20
108	5cpv	1cdp 4cpv	0·27
112	2c2c	3c2c	0·09
124	1rn3	5rsa 6rsa 7rsa	0·15
129	1lyz	1lzt 2lym 2lyz 2lz2 2lzt 3lym [3-8]lyz 1lym.A	0·35
136	1ecd	1eca 1ecn 1eco	0·06
138	3fxn	4fxn	0·21
146	1mba	2mba 3mba 4mba	0·21
153	1lh4	1lh[1-3] 1lh[5–7] 2lh[1-7]	0·12
153	1mbd	5mbn 1mb5 1mbc 1mbn 1mbo 4mbn	0·41
159	4dfr.A	7dfr	0·74
164	2lzm	1l[01-02] 1l[04-10] 1l[12-25] 1l[27-35] 1lyd 3lzm	0·12
170	9wga.A	1wgc.A 2wgc.A 7wga.A	0·21
212	9pap	1ppd	0·18
237	2cna	3cna	0·74

† The following reference proteins having homologies were excluded from the reduced training set: 1pcy, 4fd1, 5cpv, 3fxn, 1mba, 1mbd, 4dfr.A, 2lzm, 9pap, 2cna (see Table 4).

‡ Digits in square brackets mean the whole range of numbers, e.g. [3–8]lyz is 3lyz, 4lyz, . . ., 8lyz.

§ Even though neurotoxin B, 1nxb, is homologous to erabutoxins 3ebx and 5ebx, it nevertheless was excluded from this list because of noticeable shape distortion: $e_g = 1{\cdot}17$ and $e_N = 1{\cdot}54$.

‖ This is the C^{α}–C^{α} distance r.m.s.d. (eqn (9)), averaged over all the homologous structures.

order of chain length). In addition, 19 of these reference structures have one or more homologous structures, by which we denote other crystal structures of proteins having the same chain length and strong sequence identity or the same proteins in different crystal environments and/or complexed with different ligands. These were not used in any training set and served only to assess the quality and predictive power of the deduced contact potentials. In all, there are 95 homologous structures, as listed in Table 2 with their corresponding 19 reference structures. As in our previous work, we derived alternative conformations for each reference structure from all larger references and from the 23 very large structures. Consequently, the smallest references had more alternatives (16,521 for 3ins.A), but even the largest, 4rhv.2, had 2127. The total number of alternatives for all reference proteins was 691,165, an order of magnitude more than in our previous work.

(b) *Compactness*

As before, the goal is to construct a function of the interresidue contacts such that each reference structure has a lower function value than any of its alternatives, just as the native conformation of a real protein has a lower free energy than any kinetically accessible alternative conformation. Although we make no claim that the function we determine in this work resembles the real free energy, we will loosely refer to our function as the contact energy. Preliminary studies indicated that some proteins are particularly difficult to bring into agreement with our goal, perhaps because they are not adequately compact or globular, certainly necessary conditions for lattice models of proteins (Crippen, 1991). For example, if a polypeptide chain crystallizes as a dimer with many interchain contacts, it is unreasonable to use the co-ordinates of a monomer in isolation as a reference structure in developing our energy function, because the contacts that stabilize the conformation would not be included in our calculations.

In order to develop a quantitative criterion to decide the suitability of a structure for use as a reference, and generally in order to distinguish between compact and non-compact structures, we examined 2 functions of a conformer's radius of gyration, r_g, and number of contacts, N_c: (1) the ratio e_g of the radius of gyration of the putative reference structure to the minimal radius of gyration $r_g(\min)$ over the set of all its alternatives:

$$e_g = r_g / r_g(\min) \tag{1}$$

and (2) the ratio of the maximal number of contacts for all alternatives, $N_c(\max)$, to the number of contacts for the reference structure in question:

$$e_N = N_c(\max)/N_c. \tag{2}$$

Here, N_c corresponds to the discrete form of the contact function, as described below. The values of $r_g(\min)$ and $N_c(\max)$ were determined by examining all the alternatives corresponding to the given reference structure, all of which have the same number of residues, of course. We find by linear regression over all our reference structures that the minimal radius of gyration depends on the number of amino acid residues N_{res} as follows:

$$r_g(\min) = -1{\cdot}26 + 2{\cdot}79(N_{res})^{1/3} \tag{3}$$

with correlation coefficient of 0·997. Another way to estimate the minimal possible radius of gyration as a function of N_{res} is to model a globular protein as an ellipsoid of rotation (Damaschun *et al.*, 1969) with mean partial volume of 134 Å^3/residue. Then the minimal radius of gyration is achieved at unit eccentricity, i.e. spherical shape, giving the same functional form as eqn (3), but changing the coefficients from −1·26 and 2·79 to 0 and 2·46, respectively. The $r_g(\min)$ values resulting from the 2 functions differ by less than 6% over the range of N_{res} considered, but the ellipsoid model curve fits the data slightly worse. Consequently, we used the empirical eqn (3) as our estimated minimal radius of gyration. Similarly, we find that the maximal number of contacts fits the linear regression equation:

$$N_c(\max) = -53{\cdot}17 + 4{\cdot}25N_{res} \tag{4}$$

with correlation coefficient 0·992. Note that the slope value of 4·25 indirectly bears out the correctness of the cutoff distances described below for specifying contacts; we really have something like the first co-ordination sphere for each residue in a contact.

We find that the position of a given protein structure on the e_g *versus* e_N diagram (Fig. 1) accurately reflects the degree and nature of its compactness. Most interesting here is the clear evidence for the existence of 2 types of non-compactness: one characterized by noticeably larger values of r_g compared with its minimal value, and the second type marked by a definitely smaller number of

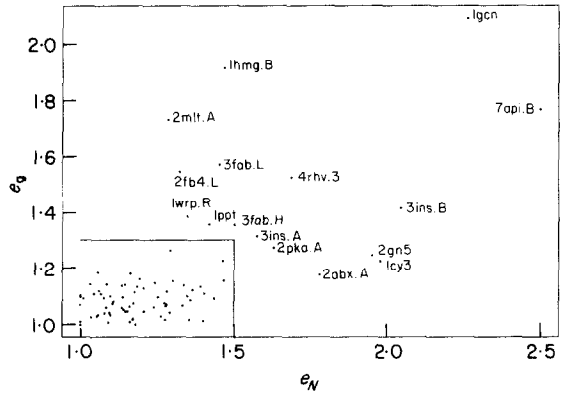

Figure 1. Diagram of e_g *versus* e_N (see the text) to characterize the compactness of the 86 reference protein structures in terms of radius of gyration and number of contacts. The 16 non-compact structures exceeding the marked limits $e_N < 1{\cdot}50$ and $e_g < 1{\cdot}30$ are indicated by their corresponding PDB codes and chain identifiers. One more extremely non-compact structure, 4rhv.4, is off scale at $e_N = 4{\cdot}52$ and $e_g = 2{\cdot}54$.

contacts N_c compared with the maximum possible for a polypeptide chain of the given length. These 2 types of non-compactness may both occur separately, e.g. high radius of gyration for 2mlt.A ($e_N = 1{\cdot}28$, $e_g = 1{\cdot}73$) and 1hmg.B ($e_N = 1{\cdot}47$, $e_g = 1{\cdot}92$); or low number of contacts for 2gn5 ($e_N = 1{\cdot}95$, $e_g = 1{\cdot}25$) and 1cy3 ($e_N = 1{\cdot}98$, $e_g = 1{\cdot}23$); and simultaneously, e.g. for 1gcn ($e_N = 2{\cdot}26$, $e_g = 2{\cdot}09$) and 7api.B ($e_N = 2{\cdot}50$, $e_g = 1{\cdot}77$), as shown in Fig. 1. Clearly, most of the proteins are rather compact, being clustered in the lower left part of the diagram, while 17 proteins obviously have non-compact conformations. We chose:

$$e_N < 1{\cdot}5 \quad \text{and} \quad e_g < 1{\cdot}3 \tag{5}$$

as the requirements for compactness. We realize that the distribution in Fig. 1 is fairly continuous throughout the diagram, and therefore these limits are somewhat arbitrary. However, we employ them in this work because such a differentiation helps us determine the desired energy function, and it is also in good agreement with visual inspections of the protein folds. Note that while we require the reference structures to be compact according to this definition, the alternative conformations have no such constraint. In fact, 10 to 20% of the alternatives for each reference protein turn out to be compact.

In order to use eqn (5) for a particular protein structure having chain length N_{res}, one needs to know r_g(min) and N_c(max). The direct way is to generate the many thousands of alternative structures and calculate r_g and N_c for each. Not only is this tedious, but for large N_{res}, there are sometimes substantial deviations from the very regular trend shown for smaller proteins. The reason is that the number of alternatives decreases as the chain length increases, simply because we are dealing with a fixed number of proteins from which to generate alternatives. Consequently, now that we have established the accurate relations given in eqns (3) and (4), we use them in all subsequent calculations to quickly obtain r_g(min) and N_c(max).

(c) *Calculations*

As in our previous study, we have evaluated conformations according to the interresidue contacts formed. The exact definition of a contact we continue to use (see Table 3) is designed to be applicable even if the sequence of a given conformation is changed. We consider only the backbone N, C′ and O atoms plus the side-chain C^β, even building in an artificial C^β if the original residue is Gly. Then a backbone–backbone contact is counted whenever $d(O, N) < 3{\cdot}2$ Å and $d(C, N) > 3{\cdot}9$ Å; a backbone–side-chain contact requires $d(N \text{ or } O, C^\beta) < 5{\cdot}0$ Å and no other atom between the interacting pair closer than 1·4 Å to the line segment joining them; and a side-chain–side-chain contact requires $d(C^\beta, C^\beta) < 9{\cdot}0$ Å and similarly no interfering atom between them. Interactions must be between residues differing by at least 3 in sequence. Backbone atoms involved in contacts are ascribed to residue type Gly, but side-chain atoms correspond to their correct residue types.

Throughout this work we have assumed the contact potential function E for a given protein conformation is a sum of the values ε assigned to the individual contacts:

$$E = \sum_{\substack{\text{contact residues} \\ i \text{ and } j}} \varepsilon(\text{class}\,(i), \text{class}\,(j), |i-j|). \tag{6}$$

where the terms depend on the same very detailed standard classification according to sequence separation and residue type classes proposed earlier (Table 2 in Crippen (1991) and Table 5, here). This classification is a plausible one that groups together helix-formers *versus* helix-breakers for short-range (i.e. sequence separation ≤ 4) interactions, and hydrophobic *versus* hydrophilic residues for long-range interactions. We assume the importance of a contact does not depend on which residue is higher in sequence, so the interaction matrices in Table 5 are all symmetric, and there are a total of 84 parameters to adjust (4 separation ranges, each having 21 interaction parameters among 7 classes of amino acid).

In the preceding work (Crippen, 1991) we required only that:

$$E(\text{reference}) \leq E(\text{alternative}) \tag{7}$$

for each reference and all alternatives of each reference. Now we demand that strict inequality hold by a margin T_k for the kth alternative given by:

$$T_k = qD_k. \tag{8}$$

Here, q is an empirically adjusted coefficient (see below), and D_k is the root-mean-square distance deviation (r.m.s.d.) between the reference and the kth alternative structure:

$$D = \left[\frac{\sum_{i<j} (d_{ij} - d'_{ij})^2}{N_{res}(N_{res}-1)/2} \right]^{1/2}, \tag{9}$$

where d_{ij} and d'_{ij} are the distances between the ith and jth C^α atoms in the reference and alternative structures, respectively. Of course, one may use the co-ordinate-based r.m.s.d. (McLachlan, 1979) instead of eqn (9), but because it makes no difference in this work, we chose the more easily calculated distance r.m.s.d. Thus, for a given reference structure and its kth alternative, we require:

$$E(k\text{th alternative}) - E(\text{reference}) \geq T_k. \tag{10}$$

The underlying idea here is to make the energy of an alternative lie above that of the corresponding reference structure by at least some minimal margin that increases linearly with their conformational difference. Test computations showed that choosing a very small positive value for q reduces eqn (10) to approximately eqn (7), makes the set of inequalities easier to solve, leads to very similar

energies for the reference structure and some of its alternatives, and leaves no room on the energy scale between the reference structure and the lowest alternative for the homologous proteins, which are expected to scatter in this range. On the other hand, too large a q caused a marked increase in the computer processor unit time required to find a solution for the set of inequalities. A reasonable compromise was $q = 3$, the value used throughout this work. Although our potential function is required to have the free energy-like property of favoring the native conformation, eqn (10) has no relation to physical energy or temperature scales. Therefore the units for E and q are arbitrary. It turned out that the method used in our previous work to solve homogeneous sets of inequalities (Jurs, 1986), as in eqn (7), could be applied to sets of inhomogeneous inequalities, as in eqn (10), and was therefore used in all that follows.

Our procedure for determining the terms consists of the following 3 steps. (1) Simply directly solving the entire set of 690,000 linear inequalities of the form in eqn (10) is hopelessly slow. At the solution, only a relatively small number of inequalities are active, as shown earlier (Crippen, 1991), particularly those inequalities arising from the more challenging compact alternatives. Therefore, we selected the first 49 alternatives for each reference that obeyed the compactness criteria of eqn (5), using eqn (3) for the minimal radius of gyration and eqn (4) for the maximal number of contacts. In the optimization procedure, all the starting values were set to the arbitrary value of $-0{\cdot}1$, and the ε terms rapidly converged to a set of first approximation values. (2) Next, we "combed" through the full list of alternatives to each reference for any alternative that violated eqn (10). Adding these to the previous list of inequalities increased the size of the problem only slightly, and the first approximation ε terms were a good starting point for calculating the second approximation. Actually, this is a very efficient way to extract all alternatives that are essential from the contact energy difference viewpoint and are missed at the first step. The clever selection of alternatives in the first 2 steps is the key to being able to treat much larger sets of inequalities than before. (3) It was found that sometimes a 3rd step of refinement of the potentials is required because some alternatives that satisfy eqn (10) before step 2 do not at the end of the step. The remedy is to return to the basic set of inequalities in the 1st step, repeat the combing, and produce a 3rd set of ε terms from the 2nd approximation. A 4th step was never required.

(d) *Two forms of contact function*

We used the above procedure to deduce contact potentials for a training set consisting of all 69 compact proteins, excluding all homologous structures (see Table 2). (Incidentally, note that Table 2 does not list neurotoxin B, 1nxb, as homologous to erabutoxins 3ebx and 5ebx, in spite of strong sequence similarity because of its noticeable shape distortion: $e_g = 1{\cdot}17$ and $e_N = 1{\cdot}54$.) However, we subsequently found that on rare occasions the resulting E for some of the homologous structures was greater than that of the lowest alternative. For example, for the reference bovine pancreatic trypsin inhibitor crystal structure 4pti, there is the homologous 5pti differing in r.m.s.d. by only 0·59 Å, yet E(5pti) is an appreciable 26·7 arbitrary units greater than E(4pti) and 9·8 above E of the lowest alternative. Since the assignment of "reference" and "homologous" structures is absolutely arbitrary, this outcome ought to be considered a violation of eqn (10). Although this happens to be the only violation of this kind, we were compelled to eliminate it.

The difficulty arises from the all-or-nothing definition of a contact, as described above. We could rewrite eqn (6) as:

$$E = \sum_{\substack{\text{contacts} \\ i,j}} V(d_{ij}, U)\varepsilon_{ij}, \tag{11}$$

where V is the value of a contact depending on d_{ij}, the relevant interatomic distance, and U, the cutoff value. The discrete contact function we have been using (Crippen, 1991) has:

$$V(d_{ij}, U) = \begin{cases} 1 & \text{if } d_{ij} \le U \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

Even slight changes in interatomic distances between 2 homologous structures may cause significantly different lists of contacts. The solution is to use a continuous contact function where V becomes a smooth sigmoidal function of d_{ij}, going from 1 below a lower cutoff distance L to 0 above an upper cutoff U:

$$V(d_{ij}, U, L) = \begin{cases} \dfrac{(d_{ij}-U)^2(2d_{ij}-3L+U)}{(U-L)^3} & \text{if } L \le d_{ij} \le \mathrm{U} \\ 1 & \text{if } d_{ij} < L \\ 0 & \text{if } d_{ij} > U. \end{cases} \tag{13}$$

Note that eqn (12) is a limiting case of the eqn (13) when $U = L$. For contacts involving side-chain atoms, we still include the effect of possible interfering atoms k near the line segment joining the interacting atoms i and j by defining the modified contact strength V_m to be:

$$V_m(d_{ij}, U, L) = V(d_{ij}, U, L) \prod_k [1 - V(d_{ijk}, U', L)], \tag{14}$$

where d_{ijk} is the distance from atom k to the line segment joining atoms i and j.

In order to determine suitable cutoff values for the continuous contact definition, we chose a limited training set of reference structures, their alternatives, and their homologous structures, namely, 4pti (12,701 alternatives and 5pti), 3ebx (12,316 alternatives and 5ebx) and 351c (10,483 alternatives and 451c). Then the cutoffs were adjusted so that each reference and its homologous structure spanned a small range of energies, while there was a large increase in energy going from the highest homologous structure to the lowest alternative. This is the only role the homologous structures played in the fitting because, otherwise, Table 2 makes it clear that homologous structures are extremely similar to their corresponding reference structures (from 0·06 to 0·74 Å r.m.s.d.), making their energies so easy to fit they were not needed in the training sets. It was found that continuity of the contact function is of critical importance only for contacts involving side-chain C^β atoms, while the contact function form for other types of contacts may remain discrete, as shown in Table 3. Similarly, we also used only the discrete form of the contact function term responsible for possible interfering atoms near the line segment joining the interacting pair of atoms (eqn (14)), as indicated by the last line of Table 3, where the continuous $U = L = 1{\cdot}4$ Å. In what follows, we will refer to this hybrid form of the function as the "continuous contact function" and to the old version as the "discrete" one. Note that in determining the ε terms, the very approximate first step of the procedure uses the discrete

Table 3
Boundary parameters of the discrete and continuous contact functions (see the text)

From atom	To atom or line	Cutoff distances (Å)‡ Discrete $L = U$	Continuous L	Continuous U
N	O	3·20	3·20	3·20
N†	C′	3·90	3·60	3·60
N or O	C^β	5·00	3·00	5·00
C^β	C^β	9·00	6·00	9·00
Any atom	Line joining 2 contact atoms	1·41	1·41	1·41

† Because the definition of a backbone–backbone contact requires a short N–O distance but a long N–C distance in order to stipulate a roughly linear hydrogen bond, the sense of these limits is reversed, compared to eqns (12) and (13).

‡ L and U denote the lower and upper distance cutoffs of the contact function, eqns (12) to (14).

form of contact function, while the more accurate continuous form was employed in the following 2 steps.

Except for treating the homologous structures, there is not a big difference between the discrete and the continuous contact functions. For example, the relationship between the discrete number of contacts and the "effective number of contacts", defined to be the sum of all contact values (eqn (13)) for the conformer according to the continuous form of the contact function, is quite linear (N_c(discrete) = 0·65N_c(continuous)+4·56) and has a correlation coefficient of 0·997.

3. Results

(a) *Complete training set (CTS)*

Our first question was whether we could satisfy equation (10) even by including in the training set all 69 compact structures having readable chain length less than 256. For each reference structure we selected the first 49 compact alternatives, which produced a set of 3381 inequalities for the whole training set. After 1521 iterations of optimization the solution for the first step was found and used as the start for the next step. On the second step, 11,336 constraints were added as described above, making altogether 14,717 inequalities. The optimization converged to a solution after 5888 iterations. Finally, on the third step only 72 new constraints were added to the 3381 from the first step (for a total of 3453 inequalities) and 1938 iterations completed the procedure. Checking the final potential against the whole data base showed perfect agreement with equation (10) for all 69 references and all their alternatives. Thus, the 14,789 constraints (=3381+11,336+72) used in all in the CTS were sufficient to predict correctly 530,062 constraints from a total of 73 proteins (including 4 non-compact structures, 3fab.L, 2fb4.L, 3fab.H and 4rhv.3 which, of course, were not in the CTS), for an average "predictive significance" of 530062/14789 = 35·8. Also, all the 95 homologous structures in Table 2 had contact energies less than the lowest alternative of the corresponding reference structure.

(b) *Reduced training set (RTS)*

Having seen that it is possible to fit all the proteins, we next tried to reduce the training set, seeking to determine the minimum number of proteins necessary to deduce a potential that could make a prediction of the same quality. Going on the theory that small and medium-sized proteins provide the most effective constraints, we first tried all reference structures having 150 residues or less. This failed in that one of the compact proteins, 2pka.B, had a number of alternatives' energies violating equation (10) and one of them was below the reference structure energy by 18·3 units.

On the other hand, when we excluded from the training set the 32 structures (Table 4) having reference energies in the CTS potential more than 100 units below the lowest alternative, we were successful. The amount of information about interresidue interactions contained in the remaining 37 structures (whose alternatives correspond to a total of 10,088 constraints) was sufficient to make correct predictions for exactly the same proteins as before with the complete training set. Note that the average "predictive significance" of a constraint in this calculation is 50% better than with the CTS: 530062/10088 = 52·54.

The values of the resulting RTS parameters seem to have clear physical meaning (Table 5). For example, for sequence separations of eight residues and more (4th separation range) the largest positive (i.e. unfavorable) values are observed for interactions between pairs of positively charged side-chains of Lys and/or Arg residues (group 5 and group 5 $\varepsilon = 9{\cdot}21$) or between pairs of negatively charged/polar Asp, Asn, Glu and Gln residues (group 7 and group 7 $\varepsilon = 4{\cdot}33$), in agreement with the obvious electrostatic repulsion between side-chains having like charges. On the other hand, the largest negative interaction parameters are for pairs of the hydrophobic residues Leu, Ile, Cys, Met and Phe (group 3 and group 3 $\varepsilon = -8{\cdot}46$) or for these hydrophobic residues and non-polar side-chains of Ala and Val (group 2 and group 3 $\varepsilon = -6{\cdot}59$), thus, reflecting the tendency of these residues to form favorable hydrophobic interactions with each other.

In general, there is an apparent correlation in contact energies of native structures and their chain lengths (Fig. 3) that fits:

$$E(\text{native}) = 47{\cdot}17 - 4{\cdot}37N_{\text{res}} \qquad (15)$$

with a correlation coefficient of −0·932. This relation may be helpful in predicting the conformation of a novel protein sequence. If the lowest proposed conformation of a protein still gives an energy well above the value expected for such a chain length, then the correct native conformation probably has not yet been suggested.

Note that all the smallest structures in the list

Table 4

Contact energy and contact energy difference for the 86 reference proteins and their 95 homologues calculated with the final (RTS) potential of Table 5

PDB code	No. res	No. alts	Contact energy (arbitrary units)						
						Homologous			
			Ref.	Alts	Diff.†	No.	Min.	Max.	Diff.‡
*3ins.A	21	16521	−59·0	−91·8	−32·8				
*2mlt.A	26	15980	−37·6	−92·2	−54·6				
*1gcn	29	15658	−11·3	−91·8	−80·5				
*3ins.B	30	15551	−51·3	−119·3	−67·9				
*1ppt	36	14920	−35·4	−86·1	−50·7				
*7api.B	36	14919	−85·2	−122·9	−37·7				
*4rhv.4	40	14506	7·8	−122·5	−130·3				
1bds	43	14199	−137·7	−121·7	15·9				
1crn	46	13895	−184·2	−166·7	17·5				
1fdx	54	13094	−233·8	−211·2	22·6				
5rxn	54	13093	−173·1	−136·7	36·5				
2ovo	56	12896	−192·5	−175·1	17·4				
4pti	58	12701	−213·3	−192·8	20·5	1	−202·8	−202·8	10·0
3ebx	62	12316	−197·7	−179·8	17·9	1	−196·8	−196·8	17·0
1cse.I	63	12220	−150·4	−133·3	17·1				
1sn3	65	12031	−204·6	−178·0	26·6				
−1ctf	68	11751	−302·9	−221·8	81·1				
1hoe	74	11198	−219·9	−196·5	23·4				
*2abx.A	74	11197	−206·0	−229.7	−23·7				
−3icb	75	11106	−358·6	−241·0	117·6				
*2pka.A	80	10660	−182·3	−254·7	−72·5				
351c	82	10483	−283·4	−274·3	36·1	1	−302·1	−302·1	54·8
1cc5	83	10395	−276·5	−255·2	21·3				
1hip	85	10222	−278·9	−255·5	23·4				
2b5c	85	10221	−303·6	−249.9	53·7				
*2gn5	87	10052	−205·5	−276·9	−71·4				
−3fxc	98	9138	−448·1	−316·2	131·9				
1hvp.A	99	9055	−373·9	−346·8	27·1				
−1pcy	99	9054	−495·5	−324·1	171·3	5	−493·6	−480·5	156·4
*1wrp.R	102	8813	−349·7	−354·1	−4·4				
5cyt.R	103	8733	−373·1	−356·0	17·1				
−4fd1	106	8498	−500·1	−357·9	142·1	2	−512·5	−492·2	134·3
1rei.A	107	8420	−332·9	−315·0	17·9				
2cdv	107	8419	−317·2	−293·2	24·1				
2ssi	107	8418	−373·4	−343·4	29·9				
1acx	108	8343	−335·8	−290·2	45·5				
−5cpv	108	8342	−600·4	−449·2	151·2	2	−620·3	−602·0	152·8
−1ccr	111	8125	−376·2	−278·3	97·9				
2c2c	112	8053	−370·4	−345·9	24·6	1	−385·0	−385·0	39·1
1hmq.A	113	7982	−375·4	−353·6	21·8				
2pab.A	114	7912	−426·2	−378·5	47·7				
2rhe	114	7911	−373·3	−369·9	3·4				
*1cy3	118	7642	−219·3	−374·1	−154·8				
−1paz	120	7509	−568·4	−394·0	174·4				
155c	121	7443	−358·9	−323·2	35·8				
1pp2.R	122	7378	−422·7	−385·1	37·5				
−1bp2	123	7314	−493·4	−389·6	103·8				
1rn3	124	7251	−399·7	−328·6	71·1	3	−443·1	−401·8	73·2
−2ccy.A	127	7067	−552·1	−438·6	113·5				
−1lyz	129	6946	−673·1	−578·0	95·1	14	−723·2	−633·3	55·3
2aza.A	129	6945	−488·1	−423·5	64·6				
−1lz1	130	6886	−743·2	−378·4	364·8				
1ecd	136	6543	−532·3	−505·9	26·5	3	−572·3	−555·4	49·5
−3fxn	138	6430	−807·5	−499·7	307·8	1	−849·7	−849·7	350·0
2hhb.A	141	6264	−528·3	−501·8	26·5				
2sns	141	6263	−418·0	−375·9	42·1				
−1mba	146	5997	−668·2	−550·0	118·1	3	−740·4	−696·2	146·2
2hhb.B	146	5996	−548·6	−493·0	55·6				
−1fx1	147	5944	−715·3	−451·3	264·0				
−2lhb	149	5843	−800·1	−521·7	278·5				
−2sod.O	151	5744	−600·4	−382·4	217·9				
2pka.B	152	5695	−462·5	−424·8	37·7				
1lh4	153	5647	−587·8	−558·6	29·3	13	−647·2	−581·8	23·2
−1mbd	153	5646	−733·8	−582·9	150·8	6	−809·5	−754·2	171·3
−4dfr.A	159	5375	−658·6	−467·3	191·3	1	−584·1	−584·1	116·8
−2lzm	164	5154	−821·0	−549·6	271·4	34	−867·1	−795·5	245·9

Table 4 *(continued)*

PDB code	No. res	No. alts	Contact energy (arbitrary units) Ref.	Alts	Diff.†	Homologous No.	Min.	Max.	Diff.‡
9wga.A	170	4895	−511·9	−472·3	39·6	3	−529·1	−499·3	27·0
–1gcr	174	4726	−816·0	−457·8	358·3				
*1hmg.B	175	4684	−363·3	−560·5	−197·2				
–2sga	181	4443	−725·7	−535·6	190·1				
2stv	184	4325	−740·1	−653·5	86·6				
3adk	194	3944	−666·1	−598·7	67·4				
–2alp	198	3795	−847·4	−634·8	212·6				
–4sbv.A	199	3758	−769·4	−618·0	151·4				
*3fab.L	207	3477	−637·1	−535·8	101·2				
–3gap.A	208	3442	−839·0	−723·5	115·5				
–9pap	212	3309	−834·1	−513·1	321·0	1	−875·1	−875·1	362·0
*2fb4.L	216	3180	−634·0	−454·6	179·4				
–2act	218	3117	−855·0	−557·4	297·6				
*3fab.H	219	3086	−653·2	−543·3	109·9				
–3rp2.A	224	2940	−1005·9	−818·8	187·0				
*4rhv.3	236	2603	−884·9	−758·1	126·8				
–2cna	237	2575	−1033·6	−643·2	390·3	1	−1035·8	−1035·8	392·6
–1est	240	2496	−963·4	−687·2	276·1				
–1tim.A	247	2320	−1088·1	−798·5	289·6				
–4rhv.2	255	2127	−993·7	−738·2	255·5				

The 17 non-compact structures are marked by an asterisk. A – marks the 32 proteins that had large contact energy differences with all the preliminary and intermediate potentials and were therefore removed from complete training set to form the reduced one.

† Ref. is the energy of the reference structure; Alts refers to the lowest contact energy over all the alternatives and Diff. is the difference between the contact energy of the reference structure and the lowest energy over all of the alternatives.

‡ Min. and Max. are the minimal and maximal contact energies over the list of homologues corresponding to a given reference structure (see Table 2). Diff. is the difference between the contact energy of the highest homologous structure and the lowest alternative.

(the first 7 in Table 4 and Fig. 3) are non-compact and violate the fitting condition, equation (10). However, only six of the remaining ten non-compact structures of larger size have violations. The energy margin between reference and lowest alternative is generally substantial (Table 4) except for the reference Bence–Jones protein 2rhe, which has energy only 3·4 units below an alternative derived from the related FAB-protein 2fb4.L (r.m.s.d. = 1·17 Å). This is an unusual situation where 2fb4.L is a close homologue of 2rhe. The next closest alternative has r.m.s.d. = 8·1 Å and energy 52·1 above the reference (Fig. 2). Otherwise, Figure

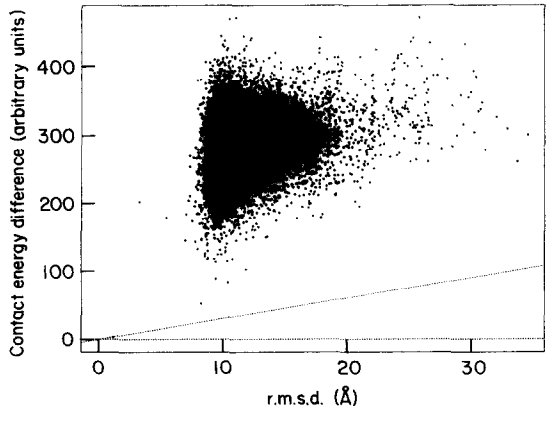

Figure 2. Contact energy difference calculated with RTS potentials *versus* r.m.s.d. plot for the 2rhe reference structure (114 residues), showing all 7911 alternatives. Zero contact energy difference and the threshold margin of 3× r.m.s.d. (eqns (8) and (10)) are shown by dotted lines. The alternative closest to the reference structure (r.m.s.d. = 1·17 Å) was generated from the crystal structure of FAB-protein 2fb4.L.

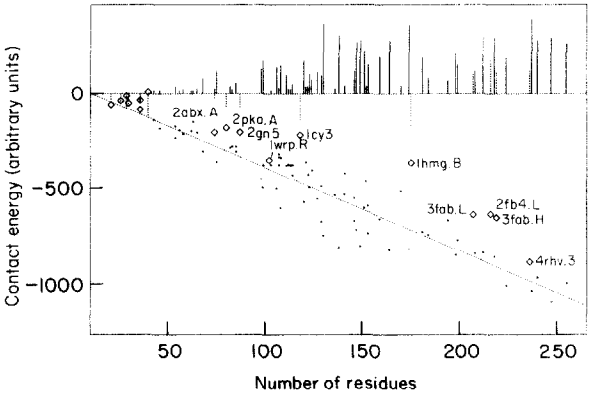

Figure 3. Contact energy of 86 reference structures *versus* number of residues. The 69 compact and 17 non-compact structures are shown by points and diamonds, respectively, the latter being marked by their PDB code. The PDB markers for the 7 smallest structures, 3ins.A, 2mlt.A, 1gcn, 3ins.B, 1ppt, 7api.B and 4rhv.4, are omitted for clarity. The upper part of the plot shows contact energy differences between the lowest energy alternative and the reference structure for each protein as lines from the zero energy level toward the respective value for the 69 compact (continuous lines) and the 17 non-compact (dotted lines) proteins. The straight line through the reference structures was determined by linear regression to have slope −4·37, intercept 47·17 and correlation coefficient −0·932.

Table 5

Contact potentials that satisfy the 73 proteins and corresponding 95 homologous structures

					Separation 3		
	G	ALICMF	VHS	P	RDEQ	TKN	YW
1	−0·69622						
2	−4·02250	4·15407					
3	4·51831	−2·15759	−1·19728				
4	8·32046	−1·52758	−2·17667	−0·25318			
5	−4·67793	0·63233	1·20919	0·55600	−3·79903		
6	−0·97071	−0·89862	0·13811	1·74729	−8·34442	−6·20602	
7	1·31923	−0·46381	−0·79571	−4·71292	−6·05593	−3·61441	−1·34488

					Separation 4		
	G	ALICMF	VHS	P	RDEQ	TKN	YW
1	0·95207						
2	−2·97196	−2·18647					
3	0·51624	−5·46954	0·59592				
4	9·43824	0·95249	1·17324	0·81616			
5	1·42327	0·88925	3·15180	−2·08650	0·22717		
6	−3·03373	−1·37106	1·46070	1·01162	2·42066	0·41988	
7	3·32498	0·98535	0·55464	0·03249	−0·27325	1·76394	−0·32876

						Separations 5 to 7	
	G	AV	LICMF	YHWST	KR	P	DNEQ
1	−1·66730						
2	0·83544	2·19482					
3	−0·80084	−1·03313	−6·92885				
4	−5·63378	−1·59914	−1·05550	−0·40860			
5	1·32582	2·20828	−2·96523	5·61379	1·27991		
6	1·94306	−0·51869	3·80413	4·24939	0·43871	−0·10825	
7	−6·21464	2·75095	−2·41791	−0·24586	3·72012	1·57452	1·98596

						Separation ≥8	
	G	AV	LICMF	YHWST	KR	P	DNEQ
1	−0·40067						
2	−0·56931	−2·62597					
3	0·92706	−6·58525	−8·46742				
4	−2·44415	−0·87669	−3·19972	2·42480			
5	−0·81858	−1·25143	−0·08153	−1·73287	9·21333		
6	1·00735	−1·33631	0·04413	0·02409	−0·69110	4·31013	
7	−3·36539	2·14130	1·90592	0·51326	−3·14586	3·39357	4·33339

Potentials are deduced with only 37 proteins, the reduced training set (RTS). Classification of amino acid residue contacts is assigned by sequence separation range and by subsets of types of residues indicated by the single-letter residue code. For each of the 4 sequence separation ranges we show the symmetric matrix of interaction parameters ε for contacts between residues of the various classes.

2 is typical of the energy distribution for all the compact proteins.

(c) *Tests of significance of a classification*

To test the significance of the classification scheme used throughout this work (Table 5) we attempted to deduce three additional potentials from the same 37 proteins of the RTS. We used the same computational protocol as described above, only with three different classification schemes.

The first test used the best (i.e. fewest adjustable ε terms) contact classification found in our earlier work (see Table 3 in Crippen, 1991). We were unable to locate a reasonable solution even after 16,000 iterations of optimization in each of the three steps. Perhaps a solution of the quality of our CTS and RTS potentials could be found after much more computing effort, but it seems unlikely.

In the second test we used the same sequence separation classifications and the same number of residue classes in each as before in the CTS and RTS potentials (Table 5), but with random assignment of residue types to classes. The first step succeeded, but the second step failed by exceeding our program's limit of 15,000 on the number of constraints while "combing" the 17th protein of the 37 in the training set. Presumably, even with a much greater limit on constraints, the calculation would fail to find a solution at great computational expense. Apparently, the classification scheme of Table 5 is not only in general agreement with conventional wisdom about residue type similarities, but the particular classification is more

important than merely the number of adjustable ε terms.

The third test was simply to interchange the residue classifications for the first and fourth separation ranges in Table 5, and then attempt to satisfy all the inequalities. Curiously enough, this succeeded, resulting in what we will refer to as the T3 parameters. With these we are able to correctly predict the same proteins as with the RTS parameters. The distribution of the contact energies of reference structures *versus* the number of residues with the T3 potential is approximately the same, with slightly different linear regression coefficients: the intercept is 25·29, the slope is −4·49, and the correlation coefficient is −0·918, compared to 47·17, −4·37 and −0·932, respectively, in equation (15). The root-mean-square difference between contact energies of compact reference structures calculated with RTS and T3 is only 67·4 arbitrary units, compared with the 1100 units for the total range of reference energies.

We were unable to repeat the systematic search for simpler classification schemes carried out in our earlier work because now we treat many more proteins, vastly more alternative conformations, and we demand in equation (10) not merely that the reference have energy less than *or equal* to each alternative, but that there be a substantial margin. Solving inhomogeneous inequalities is qualitatively different from, and more time-consuming than, solving homogeneous ones.

The RTS potentials were also checked against a representative sampling of crystal structures with fewer than 256 residues that had been added to the Protein Data Bank since 15 October 1990: 2eti (trypsin inhibitor II, 28 residues), 4tgf (human growth factor, 50 residues), 1fkf (FK506 binding protein, 107 residues) and 1cd4 (T-cell surface glycoprotein, 173 residues). These had never been seen in our laboratory until after the RTS potential had been determined. Using the same list of protein structures as before, the alternatives for each of these structures were generated. Only the structure of 4tgf growth factor (13,495 alternatives) with apparent disturbance of compactness (parameters $e_g = 1{\cdot}38$ and $e_N = 1{\cdot}51$ exceed the corresponding threshold values of 1·30 and 1·50) has a number of alternatives with contact energy less than the reference structure. The three others, 2eti (15,766 alternatives), 1fkf (8421 alternatives) and 1cd4 (4769 alternatives), demonstrate obvious satisfaction of equation (10) for all alternatives generated. While large proteins are relatively easy to fit, the trypsin inhibitor 2eti is a remarkably small structure that we can nonetheless successfully predict because it obeys our requirements for compactness and many internal contacts.

4. Discussion

It is interesting to compare our results with that of Hendlich *et al.* (1990), the most similar work outside our group that we are aware of. They derived many different potentials of mean force for C^β–C^β interactions only by surveying a database of 101 separate chains in protein X-ray crystal structures, listed in their Table 3. To make predictions of the folding for one protein, they would remove it from the database and use the remaining 100 to derive the effective energy of interaction as a function of distance between side-chains, broken down into 15 sequence separation classes and for each of these, all 210 residue pair type classes. Then they generated a set of alternative conformations in the same way we do, and compared their calculated energies of the native *versus* all its alternatives. One view of their potentials is that they consist of $210 \times 15 = 3150$ different histograms as a function of C^β–C^β distance, while we have only 84 ε terms in our Table 5. However, we adjusted our ε terms empirically to satisfy a large number of inequalities, but their histograms are not adjustable parameters. In return, we get much greater predictive power: a training set of 37 compact proteins invariably favors the native structure of 73 proteins over all alternative conformations. By way of comparison, their Table 7 lists 53 proteins that we would consider compact, ranging between 21 and 199 residues in length. Of these, the two corresponding potentials of mean force (denoted in their work as potentials S and A), derived apparently from 100 crystal structures in each case, could favor the native over the alternatives in both the S and A cases only for 34 compact proteins and two non-compact proteins.

There are a number of likely reasons for our superior predictive power. First, we derive our ε terms by comparing the native conformation with misfolded alternatives, rather than surveying only native conformations for a potential of mean force. In other words, the potential must be trained by showing it what is wrong as well as what is right. Secondly, we find it crucial to deal only with compact native conformations, as judged both from the radius of gyration and from the relative number of contacts. We cannot account for the crystal structure of an isolated non-compact polypeptide chain when its conformation is stabilized by *inter*molecular contacts in the crystal, and we suspect this has led to some of the difficulties experienced by the Sippl group, since they used both compact and non-compact native conformations. Thirdly, we find that backbone–backbone and epecially backbone–side-chain interactions are important (see the G columns in Table 5), whereas they considered only side-chain–side-chain interactions.

A priori, one might assume the classification scheme for residue–residue interactions is very important. The assumption is based on conventional wisdom about grouping together helix-formers *versus* helix-breakers for short-range interactions, and grouping according to hydrophobicity for medium and long-range interactions. Indeed, this is the line of reasoning that led to the classification scheme in Table 5, as previously set forth in Crippen (1991), and subsequently used to produce the CTS and RTS potentials. However, our classification is certainly not unique, as demonstrated by

the success of the T3 potential, where the classifications for the first and fourth separation ranges were interchanged. Although the RTS potential is very powerful in its ability to satisfy a half million inequalities, other equally good potentials could be found, possibly involving fewer parameters and possibly having even better predictive power.

In our approach, we are happy to see that some kinds of trouble simply do not arise. For example, although 1hvp.A is not an experimentally determined structure, but was rather postulated by homology modeling (Weber *et al.*, 1989), it nonetheless can be easily accounted by our contact function, whereas it could not be predicted by Hendlich *et al.* (1990). As another example, we utterly disregard any ligands or prosthetic groups in proteins, even large ones covalently attached to the polypeptide chain. Even so, we observe no correlation between the presence or absence of prosthetic groups and the quality of our predictions, just as long as the native is compact according to our criteria in equation (5).

It is worth noting that the range of contact energies for compact and non-compact alternatives calculated with the RTS potential are approximately the same. Both types of alternatives may be found among the very best (which are, of course, always higher than the reference structure energy) and the very worst. This means, in particular, that one must consider both types of alternatives when forming the set of inequalities, equation (10), rather than only compact ones. This finding runs against one's intuition that compact structures always have lower contact energy.

The major improvement of this work over our previous effort (Crippen, 1991) is the introduction of a substantial margin $T > 0$ in equation (10) that becomes small when the r.m.s.d. becomes small, coupled with a continuous contact function that guarantees a small difference in contacts for a small r.m.s.d. For example, the reference structure 5cyt.R has an alternative derived from 1ccr with r.m.s.d. = 0·47 Å, 2rhe has an alternative from 2fb4.L with r.m.s.d. = 1·17 Å, 1lyz has an alternative from 1lz1 with r.m.s.d. = 1·93 Å, and 2hhb.B has an alternative from 1mbd with r.m.s.d. = 1·92 Å. Then, quite naturally, all these "homologous" alternatives have energies only slightly above the corresponding reference structure's, but still satisfy equation (10) by a small margin. The only exception is the 2hhb.A reference structure for which the lowest energy alternative lies 55·6 units above the reference, yet has a large r.m.s.d. = 9·5 Å, compared with the second lowest alternative at 70·7 units above the reference, yet differing in conformation by only 1·92 Å. However, this result is not unexpected because a 2 Å r.m.s.d. is enough to allow a considerable change in the contacts.

It is especially interesting to note that the "novel" folding pattern found in the recently determined n.m.r. and X-ray crystal structures of 1fkf, FK506 binding protein (Michnick *et al.*, 1991; van Duyne *et al.*, 1991), is not new from the viewpoint of interatomic contact arrangements, given that we can correctly predict it on the basis of the reduced training set of 37 old protein structures. This finding allows one to hope that only minor readjustments of the contact potential will be required to keep a high level of predictive power as more proteins are considered.

In spite of the encouraging results we have obtained so far, there are two special cases we must treat in future versions of this potential. First, sequence homologues (see Table 2) are now correctly handled in the analysis without even being employed in the derivation of the potential, but we have not paid special attention to proteins having very similar conformation, yet low sequence identity. Instead, such pairs of proteins were used only in the general fashion to generate alternative conformations for each other. Presumably, we should demand that the two different sequences applied to essentially the same conformation should produce very similar contact energies. Work is in progress to at least see what the RTS potential says about such structural homologues.

The second case is that of non-compact native proteins, which we have so far simply excluded from the derivation of our potential as well as its testing. We find that for such a reference structure there are generally many alternatives having substantially lower contact energies. However, there are apparently very few examples of protein crystal structures where a polypeptide chain fails our compactness test without having significant interactions with neighboring chains. Work is in progress to treat such crystal structures as multimeric aggregates of polypeptide chains such that the multimer is compact.

The results presented here on the contact energy approach allow one to conclude that the problem of identifying the correct fold out of a large but discrete set of alternatives is basically solved. Given such a powerful tool for identifying the native fold, our next goal is to implement a method to suggest possible "native" folds for a given amino acid sequence when the correct answer is not known and when it is not just a segment out of an already determined crystal structure.

This work was supported by grants from the National Institutes of Health (GM37123) and the National Institute on Drug Abuse (DA06746). We are indebted to the reviewers for their helpful remarks and concerns, and to Steven Bryant for kindly providing his PKB software.

References

Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987). Protein data bank. In *Crystallographic Databases—Information Content, Software Systems, Scientific Applications* (Allen, F. H., Bergerhoff, G. & Sievers, R., eds), pp. 107–132, Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester.

Bryant, S. H. & Amzel, L. M. (1987). Correctly folded proteins make twice as many hydrophobic contacts. *Int. J. Pept. Protein Res.* **29**, 46–52.

Chiche, L., Grigoret, L. M., Cohen, F. E. & Kollman, P. A. (1990). Protein model structure evaluation using the solvation free energy of folding. *Proc. Nat. Acad. Sci., U.S.A.* **87**, 3240–3243.

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

Damaschun, G., Müller, J. J., Pürschel, H.-V. & Sommer, G. (1969). Berechnung der Form kolloider Teilchen aus Röntgen–Kleinwinkeldiagrammen. *Monatsh. Chemie,* **100**, 1701–1714.

Eisenberg, D. & McLachlan, A. D. (1986). Solvation energy in protein folding and binding. *Nature (London),* **319**, 199–203.

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G. & Sippl, M. J. (1990). Identification of native protein folds amongst a large number of incorrect models. *J. Mol. Biol.* **216**, 167–180.

Jurs, P. C. (1986). *Computer Software Applications in Chemistry,* pp. 198–199, John Wiley, New York.

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

McLachlan, A. D. (1979). Gene duplications in the structural evolution of chymotrypsin. *J. Mol. Biol.* **128**, 49–79.

Michnick, S. W., Rosen, M. K., Wandless, T. J., Karplus, M. & Schreiber, S. L. (1991). Solution structure of FKBP, a rotamase enzyme and receptor for FK506 and rapamycin. *Science,* **252**, 836–839.

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

Novotny, J., Rashin, A. A. & Bruccoleri, R. E. (1988). Criteria that discriminate between native proteins and incorrectly folded models. *Proteins: Struct. Funct. Genet.* **4**, 19–30.

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

van Duyne, G. D., Standaert, R. F., Karplus, P. A., Schreiber, S. L. & Clardy, J. (1991). Atomic structure of FKBP-FK506, an immunophilin-immunosuppressant complex. *Science,* **252**, 839–842.

Weber, I. T., Miller, M., Jaskolski, M., Leis, J., Skalka, A. M. & Wlodawer, A. (1989). Molecular modeling of the HIV-1 protease and its substrate binding site. *Science,* **243**, 928–931.

Edited by F. Cohen