ACCURATE MEAN-FORCE PAIRWISE-RESIDUE POTENTIALS FOR DISCRIMINATION OF PROTEIN FOLDS

Boris A. Reva[1,2], Alexei V. Finkelstein[3], Michel F. Sanner[1] and Arthur J. Olson[1]

[1]Department of Molecular Biology, The Scripps Research Institute, 10666 North Torrey Pines Road, Ca 92037, USA;

[2]on leave from Institute of Mathematical Problems of Biology Russian Academy of Sciences, 142292, Pushchino, Moscow Region, Russian Federation;

[3]Institute of Protein Research, Russian Academy of Sciences, 142292, Pushchino, Moscow Region, Russian Federation

We present two new sets of energy functions for protein structure recognition. The first set of potentials is based on the positions of alpha- and the second on positions of beta-carbon atoms of amino acid residues. The potentials are derived using a theory of Boltzmann-like statistics of protein structure by Finkelstein *et al*[1]. The energy terms incorporate both long-range interactions between residues remote along a chain and short-range interactions between near neighbors. Distance-dependence is approximated by a piecewise constant function defined on intervals of equal size. The size of this interval is optimized. A database of 222 non-homologous proteins was used both for the derivation of the potentials, and for the “threading” test originally suggested by Hendlich *et al*[2]. For threading, we used 102 non-homologous protein chains of 60 to 200 residues. The energy of each of the native structures was compared with the energy of 45 to 20 thousand alternative structures generated by threading. Of these 102 native structures 94 have the lowest energy with alpha-carbon-based potentials, and even more, 100 of these 102 structures, have the lowest energy with the beta-carbon-based potentials.

1 INTRODUCTION

The possibility of predicting protein structure from amino acid sequence is limited by errors in the energy parameters[3] and the combinatorial complexity of the problem. Prediction is a feasible task only with energy functions that allow fast and efficient sorting over many conformations. To this end, a residue-residue approximation is usually used which attributes all atomic interactions to single points placed one per residue.

Physically, such potentials should come as the result of averaging over all interactions at the atomic level between amino acid residues, and between residues and solvent molecules. However, direct calculations of such mean-force potentials are not currently possible both because of methodological difficulties and the lack of reliable atomic energy functions. Therefore, there is significant interest in finding alternative ways to derive simplified energy functions.

There have been several attempts to derive energy functions from structural data on proteins. Initially such potentials were used to predict secondary structure[4-6]; now with the rapidly increasing protein database, there are many attempts to derive potentials for estimating the energy of the tertiary structure (see for review Refs. 9-14).

Most of the approaches exploit Boltzmann's principle: that frequently observed states correspond to low energy states[7-15]. However, the physical origin of Boltzmann-like statistics in proteins, which form unique 3D structures rather than ergodic ensembles of separate residues, was analyzed only recently[1].

In this study we apply the results of that analysis to derive energy functions from known protein structures. Our approach is similar to the one originally used by Sippl[8]. We derive pairwise, distance-dependent, "mean-force" potentials, treating separately long-range and short-range interactions. However, our method of choosing the reference state for long-range interactions and our treatment of short-range interactions differ from the approach used by Sippl.

2 METHODS

Our task is to estimate the energy of interaction, $\varepsilon_{\alpha\beta}(r)$, for a pair of residues α and β (α, β = Gly, Ala,...), where the inter-residue distance r is defined from positions of the C_α (or C_β) atoms. Our estimates of $\varepsilon_{\alpha\beta}(r)$ follow from the theory of Boltzmann-like statistics of protein structures[1]. This theory shows that the requirement for overall thermodynamic stability of unique protein folds results in the observed Boltzmann-like statistics of their elements.

Let us consider a large database of protein structures, and define $N^s_{\alpha\beta}$ as the number of all $\alpha\beta$–pairs occupying positions $i, i+s$ along a chain (i is any position); and $N^{1s}_{\alpha\beta}(r)$ as the number of these pairs at a distance r.

According to Ref.1, the expected value of $N^{1s}_{\alpha\beta}(r)$ is:

$$N^{1s}_{\alpha\beta}(r) = AN^s_{\alpha\beta}M^s(r)\exp\left[-\Delta E^s_{\alpha\beta}(r)/RT_c\right] \tag{1}$$

Here A is a distance- and residue-independent normalization constant; $M^s(r)$ is the probability of finding $i, i+s$ residues at a distance r in the total set of globular folds, (i.e. $M^s(r)$ is proportional to the number of folds where residues $i, i+s$ are at a distance r), T_c is the characteristic temperature of freezing of native folds ($\sim 300K$), R is the gas constant, and $\Delta E^s_{\alpha\beta}(r)$ is the effective interaction energy:

$$\Delta E^s_{\alpha\beta}(r) = \varepsilon^s_{\alpha\beta}(r) + E^s_{\alpha\beta}(r)\ , \tag{2}$$

where $\varepsilon^s_{\alpha\beta}(r)$ is the energy of direct interaction between residues α and β at a distance r, and $E^s_{\alpha\beta}(r)$ is the mean (averaged over all the possible environments of the pair $\alpha\beta$ in stable protein structures) energy of indirect interaction of α and β, i.e. the interaction mediated by all the surrounding residues.

Thus,

$$\frac{N^{1s}_{\alpha\beta}(r_1)}{N^{1s}_{\alpha\beta}(r_2)} = \frac{M^s(r_1)}{M^s(r_2)}\cdot\exp\left(-\frac{\left[\varepsilon^s_{\alpha\beta}(r_1) - \varepsilon^s_{\alpha\beta}(r_2)\right] + \left[E^s_{\alpha\beta}(r_1) - E^s_{\alpha\beta}(r_2)\right]}{RT_c}\right), \tag{3}$$

which corresponds to Eq.10 of Ref.1 where the term Δ_E therein would now include $\varepsilon^s_{\alpha\beta}(r_1) - \varepsilon^s_{\alpha\beta}(r_2)$, while $E^s_{\alpha\beta}(r_1) - E^s_{\alpha\beta}(r_2)$, which depends on the possible amino acid environment of the αβ pair, would contribute to both Δ_E and $\Delta_\sigma/2RT_c$ terms.

The direct residue-to-residue interaction energy can be estimated as

$$\varepsilon^s_{\alpha\beta}(r) = -RT_c \ln\left[\frac{N^{1s}_{\alpha\beta}(r)}{N^s_{\alpha\beta} \cdot M^s(r)}\right] + RT_c \cdot \ln A - E^s_{\alpha\beta}(r) \tag{4}$$

2.1 *Long-range interactions*

When residues are remote in the chain ($s > s_0$ » 1), so that they can be at a distance where they do not interact, the precise value of s is not important. Moreover, the order of residues in a pair (αβ or βα) is also not important. Then the value of $\varepsilon_{\alpha\beta}(r) = \varepsilon_{\beta\alpha}(r)$ for the long-range interactions can be estimated as

$$\varepsilon_{\alpha\beta}(r) = -RT_c \ln\left[\frac{N^1_{\alpha\beta}(r)}{N_{\alpha\beta} \cdot M(r)} \Big/ \frac{N^0_{\alpha\beta}(\geq R_c)}{N_{\alpha\beta} \cdot M^0(\geq R_c)}\right] - [E_{\alpha\beta}(r) - E_{\alpha\beta}(\geq R_c)] \tag{5}$$

Here $N^1_{\alpha\beta}(r)$ are the number of cases where the remote (separated by more than s_0 chain residues) αβ and βα pairs occur at a distance r (or rather in an interval $r \pm \Delta/2$; the value of the resolution interval Δ will be optimized below); R_c is the minimal distance where direct interactions between any pair of residues is absent, i.e. $\varepsilon_{\alpha\beta}(r \geq R_c) = 0$; $N^0_{\alpha\beta}(\geq R_c)$ are the number of cases when $r_{\alpha\beta} \geq R_c$ or (more precisely, $r_{\alpha\beta} \geq R_c + \Delta/2$, for a given resolution Δ); $N_{\alpha\beta}$ is the total number of the remote αβ and βα pairs; $M(r)$ and $M^0(\geq R_c)$ are the probabilities of finding the remote residue pairs at the distances r (or rather from $r - \Delta/2$ to $r + \Delta/2$) and $r \geq R_c$, respectively, in the total set of globular folds. The term $E_{\alpha\beta}(\geq R_c)$ is the average energy of the indirect interactions at $r \geq R_c$; because of the averaging over the distances $r \geq R_c$, this term is small and can be neglected. The term $E_{\alpha\beta}(r)$ can be neglected at small distances $r < R_c$ where a direct interaction of two residues is strong.

Thus,

$$\varepsilon_{\alpha\beta}(r) = -RT_c \ln\left[\frac{N^1_{\alpha\beta}(r)}{N^{*1}{}_{\alpha\beta}(r)}\right] \tag{6}$$

where

$$N^{*1}{}_{\alpha\beta}(r) = N^0_{\alpha\beta}(\geq R_c)\frac{M(r)}{M^0(\geq R_c)} = N^0_{\alpha\beta}(\geq R_c)\frac{\sum_{\alpha\beta} N^1_{\alpha\beta}(r)}{\sum_{\alpha\beta} N^0_{\alpha\beta}(\geq R_c)} ; \tag{7}$$

here the ratio of probabilities $M(r)/M^0(\geq R_c)$ is approximated by the ratio of the total number of all remote residue pairs found at a distance r, to the total number of all residue pairs at all the distances $r \geq R_c$; (sums are taken over all the $20 \cdot (20+1)/2 = 210$ kinds of residue pairs).

In formula (5), $N^1_{\alpha\beta}(r)$ represents the pairwise distribution, which depends on the energy of interaction between residues α and β; $N^{*1}_{\alpha\beta}(r)$ represents the pairwise distribution extrapolated to the distances of inter-residue interactions from the non-interaction region. Thus, equation (5) is a potential of mean-force as it is defined in statistical physics[16].

2.2 Short-range interactions depending on distance between residues

In this study, short-range interactions are defined as the ones between residues occupying positions “i,i+2”, “i,i+3”, “i,i+4” along a chain, that corresponds to $s_0 = 4$ (see Fig.1a).

.

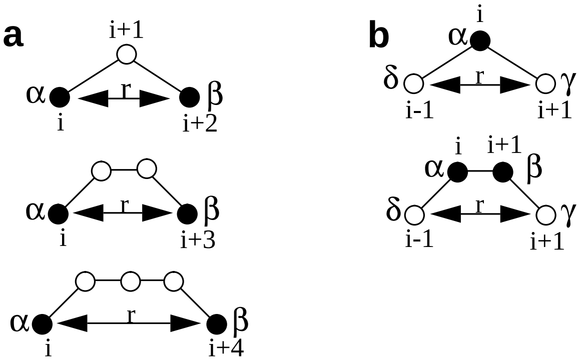

Fig. 1: A scheme of short range interactions; residues for which potentials are derived are shown by filled circles. (a) SR-interactions depending on inter-residues distances; (b) SR-interactions depending on chain bending.

To estimate these interactions, we simply neglect the distance-independent term, $\ln A$, and the energy of indirect interactions, $E^s_{\alpha\beta}(r)$, in equation (3) to obtain:

$$\varepsilon^s_{\alpha\beta}(r) = -RT_c \ln\left[\frac{N^{1s}_{\alpha\beta}(r)}{N^{*1s}_{\alpha\beta}(r)}\right] \tag{8}$$

where

$$N^{*1s}_{\alpha\beta}(r) = N^s_{\alpha\beta} \cdot M^s(r) = \sum_r N^{1s}_{\alpha\beta}(r) \frac{\sum_\alpha \sum_\beta N^{1s}_{\alpha\beta}(r)}{\sum_\alpha \sum_\beta \sum_r N^{1s}_{\alpha\beta}(r)} \quad ,s=2,3,4 \tag{9}$$

The value $N^{1s}_{\alpha\beta}(r)$ is calculated as a number of αβ pairs having positions $i, i+s$ in a chain and the distance between $r-\Delta/2$ to $r+\Delta/2$ in space in the total set of globular folds.

The definition of potentials given in (6) and (8) are different in two ways: the reference state for short-range interactions is chosen as the average energy of an interacting pair rather than a state where interaction is absent (the latter is not possible for neighbor residues); also for short-range interactions we distinguish between pairs αβ and βα.

2.3 Short-range interactions depending on chain bending

The distance between two residues in positions $i, i+s$ also depends on residues which occupy intervening positions (see Fig1b): these residues determine the local chain stiffness.

To take into account these interactions we introduce two “bending-energy” terms:

$$u^1_\alpha(r) = -RT_c\ln\left[\frac{\tilde{N}^1_\alpha(r)}{\tilde{N}^{*1}_\alpha(r)}\right] \text{ and } u^2_{\alpha\beta}(r) = -RT_c\ln\left[\frac{\tilde{N}^1_{\alpha\beta}(r)}{\tilde{N}^{*1}_{\alpha\beta}(r)}\right] \quad (10)$$

where

$$\tilde{N}^1_\alpha(r) = \sum_\delta\sum_\gamma \tilde{N}^{1s=2}_{\delta\alpha\gamma}(r) \quad (11)$$

$$\tilde{N}^{*1}{}_\alpha(r) = \tilde{N}^{s=2}_\alpha \cdot M^{s=2}(r) = \sum_\delta\sum_\gamma\sum_r \tilde{N}^{1s=2}_{\delta\alpha\gamma}(r)\,\frac{\sum_\delta\sum_\gamma N^{1s=2}_{\delta\gamma}(r)}{\sum_\delta\sum_\gamma\sum_r N^{1s=2}_{\delta\gamma}(r)} \quad (12)$$

and

$$\tilde{N}^1_{\alpha\beta}(r) = \sum_\delta\sum_\gamma \tilde{N}^{1s=3}_{\delta\alpha\beta\gamma}(r) \quad (13)$$

$$\tilde{N}^{*1}{}_{\alpha\beta}(r) = \tilde{N}^{s=3}_{\alpha\beta} \cdot M^{s=3}(r) = \sum_\delta\sum_\gamma\sum_r \tilde{N}^{1s=3}_{\delta\alpha\beta\gamma}(r)\,\frac{\sum_\delta\sum_\gamma N^{1s=3}_{\delta\gamma}(r)}{\sum_\delta\sum_\gamma\sum_r N^{1s=3}_{\delta\gamma}(r)} \quad (14)$$

In formulae (12)-(13) $\tilde{N}^{1s=2}_{\delta\alpha\gamma}(r)$ and $\tilde{N}^{1s=3}_{\delta\alpha\beta\gamma}(r)$ are the numbers of cases for which a residue α or, correspondingly, a residue pair αβ intervenes in a δγ-pair at distances r-Δ/2 to r+Δ/2, see Fig.1b; (index s shows a separation between δ and γ); $\tilde{N}^{s=2}_\alpha$ and $\tilde{N}^{s=3}_{\alpha\beta}$ are the total numbers of cases for which a residue α or, correspondingly, a residue pair αβ occurs in intervening positions. Thus, the same δαγ fragment contributes to both $u^1_\alpha(r)$ and $\varepsilon^{s=2}_{\delta\gamma}(r)$ potentials. However, the correlation between these potentials is negligible since the number of different amino acid types, 20, is great.

All the potentials (the chain-bending potentials, the short-range distant-dependent potentials, and the long-range potentials) have the form of a piecewise constant function of the distance. The optimal size of the resolution intervals Δ of these functions is established below.

3 RESULTS AND DISCUSSIONS

In order to study the accuracy of our potentials we repeated the test done by Sippl and coworkers[2]. In this test 101 proteins of different sizes and structural classes were used to derive potentials and to evaluate their accuracy using the "threading" method. Among the 101 proteins, 65 of length less than 200 residues were chosen; for each of these proteins, potentials were derived using the remaining 100 original proteins. Then, for any selected protein of length L, all the segments of length L from the 100 protein set were used as alternative structures. Energies of these structures were estimated with the corresponding potentials.

We repeated this test with our energy functions (Eqs.6,8,10) using the same set of proteins. Potentials were derived for both C_β and for C_α atoms (for the threading test with C_β potentials, absent C_β atom positions in Gly-residues were replaced by positions of corresponding C_α atoms).

The cut-off distance $R_c = 14\overset{\circ}{A}$ was chosen. Below $14\overset{\circ}{A}$ one can expect a direct interaction of long side chains. Above this distance, any direct interaction is absent. The maximal distance between the residues participating in short-range interactions (4 residues along a chain) is in concordance with the interaction cut-off $R_c = 14\overset{\circ}{A}$.

Positions of the native conformations in the energy-sorted list for 65 proteins obtained with different potentials are given in Table 1:

Table 1: Position of the native conformation in the energy-sorted list for 65 proteins obtained with different potentials.

PDB code	$C_\beta 2^a$	$C_\beta 2^b$	$C_\alpha 2^b$	$C_\beta 1^b$	$C_\alpha 1^b$
1ins.A	423	207	1994	96	23
1mlt.A	54	363	303	1	494
1gcn	2267	1992	547	2481	188
1ins.B	173	263	881	845	341
1ppt	39	96	571	43	643
1rhv.4	30	7	109	19	6
1bds	1	1	1	1	1
1crn	14	1	1	1	1
5rxn	2414	1	278	1	44
1fdx	28	1	5	1	2
1ovo.A	1	1	1	1	3
4pti	1	1	2	1	1
2mt2	1	1	66	1	26
2ebx 1cse.I	1	1	1	1	1
1sn3	10	1	1	1	1
1ctf	1	1	1	1	1
1hoe	25	1	151	1	20
2abx.A	71	2	1	2	2
3icb	3	1	1	1	1
2pka.A	1	1	1	1	1
351c	2	1	2	1	1
1cc5	12	1	1	1	1
2b5c	1	1	1	1	1
1hip	6	1	1	1	1
2gn5	35	15	113	322	66

PDB code	$C_\beta 2^a$	$C_\beta 2^b$	$C_\alpha 2^b$	$C_\beta 1^b$	$C_\alpha 1^b$
3fxc	1	1	1	1	1
1hvp.A	3	1	4	1	1
1pcy 1wrp.R	1	1	1	1	1
4cyt.R	1	1	1	1	2
2ssi	1	1	1	1	1
2cdv	18	1	9	1	1
1rei.A 1acx 1cpv 2c2c 1hmq.A	1	1	1	1	1
2pab.A 1paz	1	1	1	1	1
155c	2	1	2	1	2
1pp2.R 1bp2 1rn3	1	1	1	1	1
2ccy.A	5	1	1	1	1
2aza.A 1lz1 3fxn 2hhb.A 2pka.B	1	1	1	1	1
2hhb.B 2lhb 2sod.O 1mbd 1lh4	1	1	1	1	1
4dfr.A 2lzm 2sga 3wga.A 4dfr.A	1	1	1	1	1
2alp 1gcr	1	1	1	1	1
1hmg.B	14	1	1	1	1
2stv 3adk 4sbv.A	1	1	1	1	1
Avr:[c]	3.0	1.7	2.7	1.6	2.1

a. C_β atom based potentials derived in[2] at the resolution interval of 2Å.

b. C_β atom and C_α atom based potentials derived according to Eqs. (6), (8) and (10) at the resolutions of 2Å and 1Å, respectively.

c. Average position is defined as the mean geometrical: $\langle P \rangle = \exp\left(\sum_{i=0}^{65} \frac{\ln(P_i)}{65}\right)$ where P_i is the position of a protein i.

One can see from the Table 1 that for short non-globular chains (hormones 1ppt, 1gcn; the individual insulin chains 1ins.A and 1ins.B; the membrane attacking peptide 1mlt and a small component of the rhinovirus protein coat 2rhv.4), whose conformations are probably stabilized by interactions within molecular complexes, neither of the approaches give satisfactory ranking; for larger proteins the new potentials show significantly better accuracy than those used by Sippl and coworkers in the previous work[2].

In Table 2 we compare contributions of different energy terms into protein structure recognition. For long-range (LR) energy terms we also considered the alternative definition of the reference state used in the Refs. 2 and 8.

The results in Table 2 show that long-range energies derived with the reference state of eq.(7) are significantly more accurate than the ones derived with the reference state of the Ref. 2. One can also see that the main contribution in protein structure recognition is achieved with only four energy terms (long-range, [eq.6], and three short-range ones, [eq.(8), s=2,3,4]). This could be another reason for improvement of the native structure ranking in comparison to the Ref.2, where fifteen energy terms were used, since the more energy functions derived from the limited database, the bigger the statistical error.

One can also see that the accuracy of potentials derived from a particular set of proteins depends on the size of the resolution Δ, and that C_β based potentials are always more accurate, than C_α based ones.

Table 2: Average positions of the native conformation in the energy sorted list of 65 proteins obtained with different combinations of C_β based potentials.

C_β based potentials	Resolution in (Å)	
	2.0	1.0
used in the work[2]	3.0	-
derived by Eqs. (6),(8),(10)	1.7	1.6
LR[§] derived with the reference state of the work[2]	5.4	4.1
LR derived by Eqs.(6), (7)	2.6	2.8
SR distance-dependent only derived by Eq.(8)	8.8	7.5
SR bending energy only derived by Eq.(10)	44.5	28.4
SR derived by Eqs.(8) and (10)	5.8	4.4
LR and SR distance-dependent of Eqs.(6) and (8)	1.7	1.6
LR and SR bending energy of Eqs.(6) and (10)	2.0	2.1

[§] The reference state of the work[2] is calculated as: $N^{*1}_{\alpha\beta}(r) = N^{0}_{\alpha\beta}(\le R_c)\dfrac{M(r)}{M^{0}(\le R_c)}$, compare to the definition of Eq.(7).

The accuracy of the statistics-derived potentials must also depend on the size of the database. The database used in the Ref.2 was relatively small, so it was of-interest to see the results obtained by using a larger one. For this purpose we used a list of low-homology (less than 25%) proteins provided by Hobohm and Sander[17].

Table 3: List of PDB codes of 222 non-homologous proteins used in the threading tests.

131l	153l	1abr.B	1add	1amg	1amp	1aor.A	1aoz.A	1arb	1ars
1ash	1atp.E	1aya.A	1bam	1bet	1bnh	1bp2	1buc.A	1cau.A	1cau.B
1ccr	1cew.I	1cfb	1chm.A	1cks.B	1clc	1cmb.A	1cpc.A	1cpc.B	1crl
1cse.I	1csh	1ctn	1ctt	1cus	1cyg	1cyo	1dhr	1dlc	1dpb
1dsb.A	1dyn.A	1eca	1ede	1eft	1fnc	1fru.A	1fxi.A	1gky	1gmf.A
1gof	1gpb	1gpr	1grj	1hdc.A	1hdg.O	1hjr.A	1hlb	1hle.A	1hmt
1hsl.A	1htm.D	1htp	1huc.B	1hur.A	1hvd	1iae	1inp	1irk	1isc.A
1ivd	1knb	1lba	1ldm	1lga.A	1lis	1lki	1lpb.B	1lpe	1lts.A
1lts.D	1mat	1min.B	1mld.A	1mls	1mmo.B	1mmo.D	1mmo.G	1mnc	1mol.A
1mpp	1mrj	1msc	1mup	1nar	1nba.A	1nch.A	1ndh	1nfp	1nhk.l
1omp	1osa	1oyc	1pbe	1pbp	1pbx.A	1pfk.A	1pgs	1phg	1pii
1plq	1poc	1pox.A	1ppi	1ppn	1ptx	1pya.B	1qor.A	1rcb	1rcf
1rib.A	1rsy	1rtm.1	1rtp.1	1rva.A	1sac.A	1sbp	1scs	1scu.A	1scu.B
1ses.A	1snc	1sxc.A	1tca	1tgx.A	1thv	1tie	1tph.1	1trk.A	1tss.A
1ttb.A	1wht.B	1wsy.B	1xyl.A	1ypt.B	1ytb.A	1zaa.C	256b.A	2acg	2acq
2alp	2aza.A	2bbk.H	2blt.A	2cba	2ccy.A	2cdv	2chs.A	2cpl	2ctc
2dkb	2dnj.A	2dri	2ebn	2end	2er7.E	2fal	2fd2	2gbp	2gst.A
2hbg	2hhm.A	2hpd.A	2hpe.A	2kau.B	2kau.C	2liv	2mad.l	2mnr	2mta.C
2nac.A	2pf1	2pgd	2pia	2pol.A	2por	2prk	2rn2	2rsl.B	2sas
2scp.A	2sil	2tgi	2tmd.A	3aah.A	3cd4	3chy	3est	3gap.B	3gly
3grs	3pga.1	3sdh.A	3sic.I	3tgl	4blm.A	4enl	4fgf	4fxn	4gcr
4mt2	6taa	7icd	7pcy	7rsa	8abp	8acn	8atc.A	8atc.B	8cat.A
8tln.E	9rnt								

From this list of 472 proteins we chose those with resolution better than 2.5Å and with no structural defects (chain gaps, significant distortions of bond lengths, missing residues), resulting in a database of 222 non-homologous proteins (see Table 3).

For threading we chose those having from 60 to 200 residues, resulting in 102 sequences. For each of these sequences we extracted potentials from the remaining 221 proteins and then used structural backbones of these proteins as alternative conformations for threading.

The results of these threading tests are presented in Tables 4 and 5. A comparison of Tables 4and 1 shows that accuracy of the potentials improves with the database size. Besides, Table 4 shows that the most accurate potentials are derived at an optimal resolution interval, Δ, used for approximating energy functions: a bigger interval will resolve fewer details of the potential, a smaller one will yield poorer statistics, and therefore larger errors.

The average ranking as well as average relative deviation of the native structure energy from the mean energy of alternative structures (“Z”-score, see the definition in legend to Table 4) are optimal when Δ=1.0-0.5Å for both C_α and C_β atom based potentials.

Table 4: Average characteristics of the threading test obtained for C_α and C_β atoms based potentials at different resolutions Δ.

Resolution (in Å)	**Positions**			**Z-score**			**Positions**			**Z-score**		
	T	**LR**	**SR**	**T**	**LR**	**SR**	**T**	**LR**	**SR**	**T**	**LR**	**SR**
	C_α atom based potentials						C_β atom based potentials					
3.0	1.5	2.5	24.1	5.2	4.5	2.9	1.13	1.5	2.6	6.7	5.5	4.3
2.0	1.3	2.4	4.7	5.5	4.5	3.6	1.08	1.5	1.9	6.9	5.5	4.8
1.0	1.16	2.4	2.4	5.7	4.6	4.1	1.06	1.5	1.3	7.2	5.6	5.3
0.5	1.18	2.4	2.1	5.9	4.6	4.4	1.09	1.4	1.4	7.4	5.6	5.5
0.25	1.18	2.4	2.8	5.8	4.5	4.3	1.08	1.4	1.8	7.3	5.5	5.3

T, LR, SR stand for average position of the native structure for the total, the long range and the short-range energies; the average position of 102 native structures is found by the formula given in Table 1; the Z-score is defined as $(E_{avr} - E_{nat})/\sigma$, where E_{avr} is the average energy, E_{nat} is the native structure energy and σ is the standard deviation of energies of alternative structures from E_{avr}.

For both types of potentials, long-range interactions give approximately two thirds of the total energy of the native structure. They provide equal accuracy in all the range of Δ from 3 to 0.25Å. Short-range interactions give virtually the same contribution in recognition of the native structure, but only at the resolution Δ of 1Å. When Δ is bigger than 1Å, the details of the short-range potentials are poorly resolved; when Δ is smaller than 1Å, statistical errors increase and become a limiting factor for precision of the complete energy function.

The C_β atom-based potentials are more accurate than the C_α ones because they better approximate the relative positions of centers of residues.

Table 5 gives the details of the threading experiment for 102 proteins with C_α and C_β based potentials, derived at the resolution interval of 1.0A.

Table 5: Characteristics of the native conformation position in the energy sorted list for 102 proteins obtained for C_α and C_β based potentials derived at 1.0Å resolution.

PDB code	Threadings	C_α potential		C_β potential		PDB code	Threadings	C_α potential		C_β potential		PDB code	Threadings	C_α potential		C_β potential	
		P[a]	Z[b]	P[a]	Z[b]			P[a]	Z[b]	P[a]	Z[b]			P[a]	Z[b]	P[a]	Z[b]
1tgx	45359	1	3.8	11	3.5	2mad	31907	1	5.8	1	6.6	1mls	26760	1	5.9	1	8.8
4mt2	45137	13	3.8	1	6.0	4fgf	31907	1	4.6	1	6.1	2rn2	26605	1	5.7	1	6.6
1cse	44696	1	4.3	1	6.1	7rsa	31907	1	6.0	1	6.6	1hlb	26298	1	5.4	1	7.7
1ptx	44476	2	5.0	1	6.7	2acg	31718	1	9.2	1	12.0	1mup	26298	1	4.9	1	6.7
1cks	41423	1	4.3	1	4.8	1ttb	31347	1	6.0	1	6.5	1gpr	26145	1	6.5	1	8.5
1zaa	39903	714	2.0	29	2.9	2ccy	31347	1	4.5	1	6.2	1hjr	26145	1	6.8	1	9.2
1cyo	39254	1	4.9	1	5.5	3chy	31162	1	6.0	1	8.7	1mnc	26145	1	6.3	1	8.5
1mol	37963	2	3.3	1	5.1	1msc	30979	1	4.3	1	4.7	131l	25550	1	6.5	1	8.2
1fxi	37534	1	5.2	1	6.3	1rcb	30979	1	5.4	1	7.2	1cpc	25550	1	6.3	1	9.0
1nch	37107	1	3.9	1	6.4	2aza	30979	1	6.1	1	7.8	1mmo	25550	1	4.1	1	5.9
7pcy	37107	1	6.5	1	8.5	1hmt	30618	1	5.4	1	6.0	2cpl	25257	1	6.9	1	8.1
2hpe	36894	1	5.9	1	7.8	1htp	30618	1	4.7	1	5.8	1tie	24968	1	7.0	1	7.2
1aya	36473	1	6.1	1	7.1	1lis	30618	1	6.0	1	7.6	1rcf	24538	1	7.1	1	8.8
2kau	36473	1	4.5	1	6.2	1poc	30087	1	5.3	1	6.8	1cpc	24111	1	6.9	1	9.7
1lts	36055	1	5.7	1	6.3	1rsy	29911	1	4.2	1	5.9	1lki	24111	1	7.6	1	9.7
1cmb	35847	19	3.0	1	3.7	1snc	29911	1	4.2	1	5.3	2scp	23829	1	5.8	1	8.1
9rnt	35847	1	4.5	1	6.8	1eca	29736	1	5.9	1	9.0	4gcr	23829	1	8.7	1	9.3
256b	35435	1	3.9	1	5.3	2end	29563	1	5.9	1	7.1	3cd4	23275	1	5.1	1	6.0
2fd2	35435	1	6.7	1	7.8	4fxn	29391	1	7.9	1	9.2	1hur	23000	1	7.2	1	8.9
1bet	35230	1	5.6	1	5.9	1pbx	28710	1	6.1	1	9.3	1ytb	23000	1	7.6	1	8.9
2cdv	35230	1	3.6	1	4.2	1nhk	28540	1	6.2	1	7.8	1cau	22863	1	4.9	1	6.9
3sic	35230	1	6.4	1	9.4	1lpe	28371	1	4.5	1	6.7	1cau	22460	1	4.8	1	6.4
1cew	35027	2	4.8	1	5.6	3sdh	28203	1	5.6	1	8.1	153l	22326	1	4.9	1	7.1
1rtp	34827	1	4.3	1	6.8	1lba	28036	1	7.0	1	8.5	1lts	22326	1	6.7	1	7.9
1ccr	34430	2	3.7	1	5.3	2fal	28036	1	6.7	1	9.2	2sas	22326	1	6.1	1	8.1
2tgi	34232	1	5.3	1	6.6	8atc	28036	1	6.6	1	8.3	1gky	22193	1	8.2	1	8.5
1dyn	34035	1	3.5	1	5.3	1ash	27870	1	5.7	1	7.3	1knb	22193	1	6.5	1	8.0
2chs	33839	1	5.2	1	6.4	2hbg	27870	1	6.9	1	10.8	1dsb	21935	1	6.5	1	8.4
2hmz	33840	1	4.2	1	5.9	2mta	27870	1	6.4	1	7.8	1isc	21426	1	8.2	1	10.7
1gmf	32868	1	5.4	1	6.9	1osa	27707	1	6.1	1	7.4	1tss	21173	1	4.6	1	5.6
2rsl	32674	1	6.1	1	7.8	1rtm	27547	1	6.1	1	7.3	1cus	20797	1	8.9	1	10.5
2pf1	32481	2	3.7	1	5.5	1grj	27230	1	5.3	1	6.7	2alp	20672	1	7.2	1	9.6
1bp2	32098	1	5.4	1	5.8	1sxc	27230	1	7.1	1	9.0	1bam	20425	1	7.5	1	7.9
1htm	32098	1	3.4	1	4.1	1wht	26916	1	5.0	1	4.9	1iae	20425	1	7.3	1	8.6

a. Position of the native conformation’s energy in the energy sorted list.

b. Z-score defined as $(E_{avr} - E_{nat})/\sigma$, where E_{avr} is the average energy, E_{nat} is the native structure’s energy and σ is the standard deviations of energies of alternative structures.

The potentials successfully recognize the native structure: only 8 proteins for C_α atom based potentials and only 2 for C_β ones are not in the lowest energy for their native structures. It is important to note the large “Z-scores”: the bigger the relative deviation of the native energy from the mean energy, the higher probability that the native structure will have the lowest energy among any other competing conformations. We have also checked if the results of threading are biased by the fact that potentials are extracted from the same protein set which is subsequently used as a source of templates for threading: the set of set of 222 proteins was divided in half. The first 111 proteins (set A) was used to extract the potentials; the second set of 111 proteins (set B) was used for threading experiments with these potentials. The obtained ranking of native structures is essentially the same as reported in Table 5.

4 CONCLUSION

In this work we have developed a consistent approach to derive phenomenological energy functions using the theory of Boltzman-like statistics of protein structure.

We have tested the approach to derive pairwise, distance-dependent potentials using the positions of C_α or C_β atoms. The energy function includes both long-range interactions between residues which are remote along a chain, and short-range ones between near chain neighbors. The distance dependence of the energy functions is approximated by a piecewise constant function defined on intervals of equal size. The size of this interval is optimized to preserve as much detail as possible without introducing excessive error due to limited statistics.

Results of these tests demonstrate that our new approach to derive potentials performs better than the previous one used by Sippl and co-workes[8,2]. Our's is more accurate in treatment of some important details of both short- and long-range potentials and therefore performs better. It is noteworthy that a similar improvement of performance has been obtained by Sippl[15] at the cost of adding of a "surface" term in the energy function and atomic description of residues, i.e. for the cost of inclusion of many additional statistical information.

The ability of our potentials to recognize protein structure was also checked on 102 non-homologous proteins 60 to 200 residues in length. Each of the sequences had to choose among a corresponding set of alternative structures obtained by threading the sequence through the backbones of 222 proteins. Most of the 102 protein sequences (94 for C_α-atom based potentials and 100 for C_β-ones) recognized their native structures.

Our studies also show that long-range and short-range interactions are equally important in protein structure recognition. As the statistics of short-range interactions are poorer than those of long-range ones, short-range interactions become the "bottle-neck" for improving the accuracy of statistical potential functions.

In our tests the best ranking of native structures was achieved for potentials approximated at a resolution of 1Å, which is obviously far from a detailed representation of the actual energy functions. We can further improve the potentials by enlarging the database.

In estimating the role of simplified pairwise potentials for the protein folding problem, one should not expect to explain all of the details of protein structure. However, these potentials can be useful for efficient discrimination of a small number of the favorable conformations from a vast number of unfavorable ones.

ACKNOWLEDGMENTS

This work was supported by NIH Grant PO1GM38794 (to AJO). The research of AVF was supported in part by an International Research Scholar's Award No. 75195-544702 from the Howard Hughes Medical Institute and by NIH Fogarty Research Collaboration Grant No. TW00546. Authors are thankful to Drs. B.Duncan and G.Morris for careful reading and improvement of the manuscript.

REFERENCES

1. Finkelstein, A., Badretdinov, A., Gutin, A. *Proteins* **23**, 142, (1995)
2. Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G., Sippl.,M. *J.Mol.Biol.* **216**, 167, (1990)
3. Finkelstein, A., Badretdinov, A., Gutin, A. *Proteins* **23**, 151, (1995)
4. Ptitsyn O.B., Finkelstein A.V. *Biofizika* (USSR) **15**, 757, (1970)
5. Chou, P.Y. and Fasman, G.D. *Biochemistry* **13**, 211, (1974)
6. Sternberg, M.J. *Anti-Cancer Drug Design* **1(3)**, 169, (1986)
7. Pohl, F.M. *Nature New Biology* **234**, 277, (1971)
8. Sippl, M. *J.Mol.Biol.* **213**, 859, (1990)
9. Sippl, M.J. *Current Opinion in Structural Biology* **5,** 229, (1995)
10. Jernigan, R., Bahar, I. *Current Opinion in Structural Biology* **6**, 195, (1996)
11. Rooman, J., Wodak, S. *Protein Engineering* **8**, 849, (1995)
12. Godzik, A., Kolinski, A., Skolnick, J. *Protein Science* **4**, 2107 (1995)
13. Miyazawa, S., Jernigan, R. *J.Mol.Biol.* , **256**, 623 (1996)
14. Thomas, P.D., Dill, K.A. *J.Mol.Biol.* **257**, 457, (1996)
15. Sippl, M.J. *J.Comput.-Aided Mol.Design* **7,** 473 (1993)
16. Mayer, J.E., Geppert Mayer, M., *Statistical Mechanics.* Wiley & Sons, 1977
17. Hobohm, U., Scharf, M., Schneider, R., Sander, C. *Protein Science* **1** , 409, (1992)