Protein Engineering vol.10 no.8 pp.865–876, 1997

Residue–residue mean-force potentials for protein structure recognition

Boris A.Reva[1,2], Alexei V.Finkelstein[3], Michel F.Sanner[1] and Arthur J.Olson[1,4]

[1]Department of Molecular Biology, Scripps Research Institute, 10666 North Torrey Pines Road, CA 92037, USA, [3]Institute of Protein Research, Russian Academy of Sciences, 142292 Pushchino, Moscow Region, Russia

[2]On leave from the Institute of Mathematical Problems of Biology, Russian Academy of Sciences, 142292 Pushchino, Moscow Region, Russia

[4]To whom correspondence should be addressed

We present two new sets of energy functions for protein structure recognition, given the primary sequence of amino acids along the polypeptide chain. The first set of potentials is based on the positions of α- and the second on positions of β- and α-carbon atoms of amino acid residues. The potentials are derived using a theory of Boltzmann-like statistics of protein structure. 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 the interval is optimized to preserve as much detail as possible without introducing excessive error due to limited statistics. A database of 214 non-homologous proteins was used both for the derivation of the potentials, and for the 'threading' test originally suggested by Hendlich *et al.* (1990) *J. Mol. Biol.*, 216, 167–180. Special care is taken to avoid systematic error in this test. For threading, we used 100 non-homologous protein chains of 60–205 residues. The energy of each of the native structures was compared with the energy of 43 000 to 19 000 alternative structures generated by threading. Of these 100 native structures, 92 have the lowest energy with α-carbon-based potentials and, even more, 98 of these 100 structures, have the lowest energy with the β- and α-carbon based potentials.
Keywords: Boltzmann-like statistics/pairwise-residue potentials/protein structure recognition/protein threading

Introduction

The possibility of predicting a protein's structure from its amino acid sequence is limited by errors in the energy parameters (Finkelstein *et al.*, 1995b) and by the astronomical number of possible alternative structures. 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 between residues to a single point within each residue.

Physically, such simplified potentials should result from some averaging of the atomic interactions over various positions and conformations of the interacting amino acid residues in addition to the surrounding solvent molecules. However, direct calculation of such mean-force potentials is not possible today because of both methodological difficulties and the lack of reliable atom-based 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 information on proteins. Initially such potentials were used to predict secondary structure (Ptitsyn and Finkelstein, 1970; Chou and Fasman, 1974; Sternberg, 1986); now, with the rapidly increasing database of protein structures, there are many attempts to derive potentials for estimating the energy of tertiary structures.

Most of these approaches exploit Boltzmann's principle [which, has been shown to be valid for fixed and non-fluctuating native protein structures with the same exponential dependence upon occurrence-on-energy (Pohl, 1971)], namely, that frequently observed states correspond to low energy states (for reviews of applications, see Sippl, 1990, 1993, 1995; Kocher *et al.*, 1994; Godzik *et al.*, 1995; Rooman and Wodak, 1995; Jernigan and Bahar, 1996; Miyazawa and Jernigan, 1996; Thomas and Dill, 1996). However, the physical origin of Boltzmann-like statistics in fixed native protein structures, which do not form an ensemble in thermodynamic equilibrium, was analysed only recently (Finkelstein *et al.*, 1995a).

In this study, we applied this theory to derive energy functions from known protein structures. Our approach is similar to that originally used by Sippl (1990). We derive pairwise, distance-dependent, 'mean-force' potentials, treating long- and short-range interactions separately. However, our method of choosing the reference state for long-range interactions and our treatment of short-range interactions differ from those used by Sippl.

Methods

Our main task was 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 (Finkelstein *et al.*, 1995a). This theory shows that the requirement for overall thermodynamic stability of unique protein folds, taken together with a possibility of mutating the amino acid sequence to reach this stability, results in the observed Boltzmann-like statistics of the protein fold elements. As in Boltzmann statistics of liquids or solids, the correlations observed in Boltzmann-like protein statistics reflect not only the direct interaction of particles (amino acid residues) but also their indirect interactions mediated by the surrounding residues. Thus, as in Boltzmann statistics of liquids or solids, in obtaining elementary potentials one can more or less rely on the short-distance rather than the long-distance correlations.

Let us consider a large 3D database of protein structures, and define $N^s_{\alpha\beta}$ as the number of the αβ pairs occupying positions i, $i + s$ along a chain (α and β are amino acids, i is any position in a chain) and $N^s_{\alpha\beta}(r)$ as the number of such pairs having a distance r between α_i and β_{i+s} in the database.

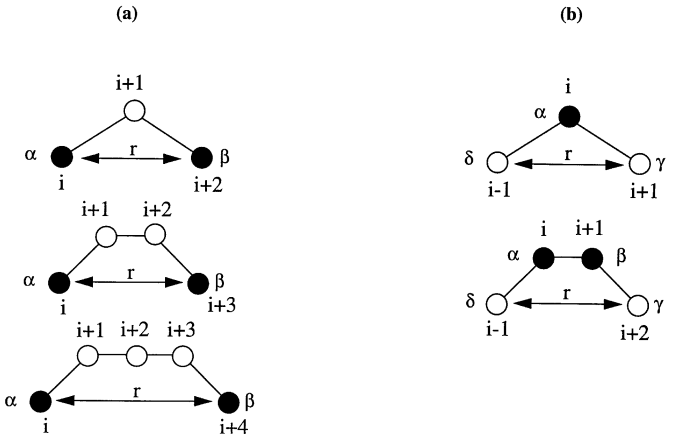

Fig. 1. Scheme of short-range interactions; residues for which potentials are derived are shown by filled circles. (**a**) Short-range interactions depending on the distance between terminal residues α and β; (**b**) short-range interactions depending on chain bending in the intervening residue α (or α and β) which affects the distance between terminal residues δ and γ.

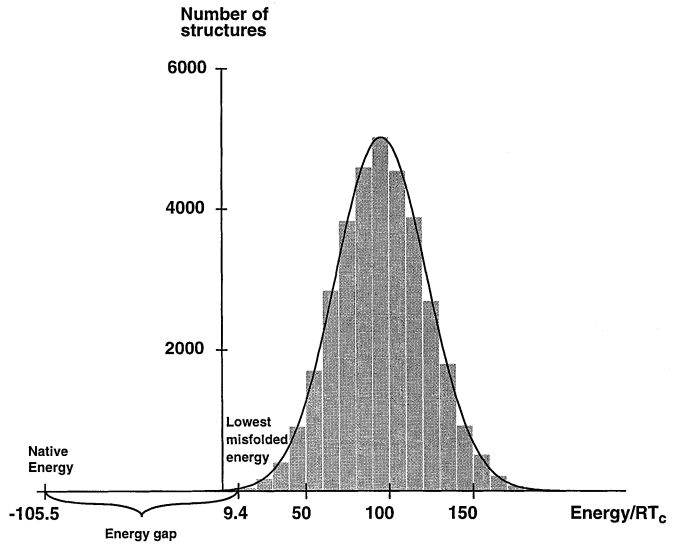

Fig. 2. Histogram and the corresponding normal distribution (thin line) of 34 234 threading energies of the ferredoxin molecule (2fd2). The normal distribution is built with an average energy of 95.4 and a standard deviation of 27.3. The difference between the average energy and the native structure energy is 200.9, which corresponds to $Z = 200.9/27.3 = 7.36$. The difference between the lowest energy of misfolded structures and the native structure energy gives the value of the energy gap (114.9) separating the native structure from misfolded ones.

According to Finkelstein *et al.* (1995a), the expected value of $N^s_{\alpha\beta}(r)$ in the limit of very large statistics is

$$N^s_{\alpha\beta}(r) = AN^s_{\alpha\beta}w^s(r)\exp[-\Delta E^s_{\alpha\beta}(r)/RT_c] \quad (1)$$

where A is a distance-independent normalization constant, $w^s(r)$ is a probability of finding i, $i + s$ residues at a distance r in the total set of globular folds [$w^s(r) = N^s(r)/\sum_r N^s(r)$, where $N^s(r) = \sum_\alpha \sum_\beta N^s_{\alpha\beta}(r)$ is the number of folds where residues i, $i + s$ are at a distance r], T_c is a 'conformational temperature' (Pohl, 1971), which is close to the characteristic temperature of freezing of native folds (Finkelstein *et al.*, 1995a) (~300 K), R is the universal gas constant and $\Delta E^s_{\alpha\beta}(r)$ is the effective interaction energy:

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

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

Thus, taking into account the proportionality $w^s(r) \sim N^s(r)$, one can write

$$\frac{N^s_{\alpha\beta}(r_1)}{N^s_{\alpha\beta}(r_2)} = \frac{N^s(r_1)}{N^s(r_2)}\exp\left(\frac{[\varepsilon^s_{\alpha\beta}(r_1)-\varepsilon^s_{\alpha\beta}(r_2)] + [\tilde{E}^s_{\alpha\beta}(r_1)-\tilde{E}^s_{\alpha\beta}(r_2)]}{RT_c}\right) \quad (3)$$

which corresponds to Equation 10 of Finkelstein *et al.* (1995a), where the term Δ_E therein would now include $\varepsilon^s_{\alpha\beta}(r_1) - \varepsilon^s_{\alpha\beta}(r_2)$, while $\tilde{E}^s_{\alpha\beta}(r_1) - \tilde{E}^s_{\alpha\beta}(r_2)$, which depends on the possible amino acid environments of the αβ pair, would contribute to both the Δ_E and $\Delta_\sigma/2RT_c$ terms in that work.

The direct residue-to-residue interaction energy estimated from Equations 1 and 2 gives

$$E^s_{\alpha\beta}(r) = -RT_c \ln\left[\frac{N^s_{\alpha\beta}(r)}{N^s_{\alpha\beta}\, w^s(r)}\right] + RT_c \ln A - \tilde{E}^s_{\alpha\beta}(r) \quad (4)$$

It is noteworthy that, since the Boltzmann-like statistics of proteins originate from amino acid mutations, the reference (zero-energy) state for the energy $\varepsilon^s_{\alpha\beta}(r)$ obtained from these statistics is a pair of 'average' amino acid residues in proteins separated by a distance s in the chain and r in space rather than an amino acid pair in vacuum or water environment (cf. Godzik *et al.*, 1995; Rooman and Wodak, 1995; Jernigan and Bahar, 1996).

Equation 4 is valid only when the expected $w^s(r)$ value is not zero. When $w^s(r) = 0$, $\varepsilon^s_{\alpha\beta}(r)$ cannot be defined from Equation 4, but must be set to infinity to make impossible any structure with the distance r between any residues.

Long-range interactions

When residues are separated 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 becomes unimportant. Moreover, the order of residues in a pair (αβ or βα) is not relevant.

Let us define $N_{\alpha\beta}(r)$ as the total number of cases where the αβ and βα pairs separated by more than s_0 chain residues occur at a distance r (or rather in an interval $r \pm \Delta/2$; the value of the resolution interval Δ will be discussed and optimized below):

$$N_{\alpha\beta}(r) = \sum_{p=1}^{P}\sum_{i=1}^{N_p-s_0}\sum_{j=i+s_0}^{N_p}(\delta_{q_i\alpha}\delta_{q_j\beta} + \delta_{q_j\beta}\delta_{q_i\alpha} - \delta_{q_i\alpha}\delta_{q_j\beta}\delta_{\alpha\beta})\,\theta\left(\tfrac{\Delta}{2}-|r_{ij}-r|\right) \quad (5)$$

where P is a number of proteins, N_p is a protein p sequence length, q_i is a residue of i type, r_{ij} is the distance between residues i and j, $\delta_{\alpha\beta} = 1$ if $\alpha = \beta$ and $\delta_{\alpha\beta} = 0$ if $\alpha \neq \beta$, $\theta(x) = 1$ if $x \geqslant 0$ and $\theta(x) = 0$, if $x < 0$.

Let us also define $N^0_{\alpha\beta}(\geqslant R_{\alpha\beta})$ as the total number of cases where residue pairs αβ and βα remote along a chain occur at non-interaction distances:

Table I. List of PDB codes of the 214 non-homologous proteins used in the threading tests[a]

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

[a]Protein chains are grouped into sets A and B as used in the 'cross-test' (Table V); the 100 target proteins used for the threading test are in italics.

Table II. Effective residue radii (Å) used in the derivation of long-range potentials[a]

Type of potential	Gly	Ala	Pro	Asn	Leu	Val	Ser	Thr	Cys	Asp	Ile	His	Gln	Glu	Met	Phe	Lys	Trp	Tyr	Arg
C_α atom	4.2	4.1	4.1	5.5	5.7	4.3	4.3	4.4	4.5	5.5	5.6	6.4	6.8	6.7	7.3	7.0	8.2	8.3	8.3	9.1
C_β atom[b]	3.9	4.9	4.9	4.9	4.9	5.0	5.0	5.0	5.0	5.0	5.0	5.1	5.1	5.2	5.6	5.7	6.7	6.8	7.1	7.6

[a]'Covalent residue radii' extracted from the database of 214 proteins are adjusted by effective van der Waals radius δ/2 (see Equation 10); δ/2 = 1.5 Å for C_α-based potentials and δ/2 = 1.2 Å for C_β-based potentials.
[b]Center of Gly is in the C_α atom.

$$N^0_{\alpha\beta}(\geqslant R_{\alpha\beta}) = \sum_{p=1}^{P} \sum_{i=1}^{N_p - s_0} \sum_{j=i+s_0}^{N_p} (\delta_{q_i\alpha}\delta_{q_j\beta} + \delta_{q_i\beta}\delta_{q_j\alpha} - \delta_{q_i\alpha}\delta_{q_j\beta}\delta_{\alpha\beta})\, \theta(r_{ij} - R_{\alpha\beta}) \quad (6)$$

where $R_{\alpha\beta}$ is the minimal distance where direct interaction between α and β residues is absent [i.e. $\varepsilon_{\alpha\beta}(r) = 0$ for $r \geqslant R_{\alpha\beta}$]; the values of $R_{\alpha\beta}$ are defined below.

Then the value of $\varepsilon_{\alpha\beta}(r)$ for the long-range interactions can be estimated as (Reva *et al.*, 1997)

$$\varepsilon_{\alpha\beta}(r) = -RT_c \ln\left[\frac{N_{\alpha\beta}(r)}{N_{\alpha\beta}\, w(r)} \div \frac{N^0_{\alpha\beta}(\geqslant R_{\alpha\beta})}{N_{\alpha\beta}\, w^0(\geqslant R_{\alpha\beta})}\right] - [\tilde{E}_{\alpha\beta}(r) - \tilde{E}_{\alpha\beta}(\geqslant R_{\alpha\beta})] \quad (7)$$

where $w(r)$ and $w^0(\geqslant R_{\alpha\beta})$ are probabilities of finding the remote residue pairs at the distance r and $r \geqslant R_{\alpha\beta}$, respectively, in the total set of globular proteins.

The term $\tilde{E}_{\alpha\beta}(\geqslant R_{\alpha\beta})$ is the average energy of the indirect interactions at $r \geqslant R_{\alpha\beta}$; because of the averaging of indirect correlations over all the distances $r \geqslant R_{\alpha\beta}$, this term is small and can be neglected. The term $\tilde{E}_{\alpha\beta}(r)$ can be neglected at small distances $r < R_{\alpha\beta}$ where the direct interactions of two residues is strong.

Thus, one can estimate $\varepsilon_{\alpha\beta}(r)$ as

$$\varepsilon_{\alpha\beta}(r) = -RT_c \ln\left[\frac{N_{\alpha\beta}(r)}{N^*_{\alpha\beta}(r)}\right] \quad (8)$$

where

$$N^*_{\alpha\beta}(r) = N^0_{\alpha\beta}(\geqslant R_{\alpha\beta})\frac{w(r)}{w^0(\geqslant R_{\alpha\beta})} = N^0_{\alpha\beta}(\geqslant R_{\alpha\beta})\frac{\sum\sum_{\alpha \geqslant \beta} N_{\alpha\beta}(r)}{\sum\sum_{\alpha \geqslant \beta} N^0_{\alpha\beta}(\geqslant R_{\alpha\beta})} \quad (9)$$

In Equation 9, the ratio of probabilities $w(r)/w^0(\geqslant R_{\alpha\beta})$ is approximated by the ratio of the total number of all the remote residue pairs found at a distance r to the total number of all the residue pairs at all the distances $r \geqslant R_{\alpha\beta}$ [sums are taken over all 20(20 + 1)/2 = 210 different residue pairs; the pairs αβ, where α < β, are taken into account in βα pairs].

Equations 8 and 9 show that the value of $\varepsilon_{\alpha\beta}$ does not change with simultaneous multiplication of all the $N_{\alpha\beta}(r)$ terms by a function depending on r (when $r \leqslant R_{\alpha\beta}$), but not on α and β. This once again shows that the above definition of $\varepsilon_{\alpha\beta}(r)$ counts the interaction energy from the interaction energy $\varepsilon_0(r)$ for some 'average' residue pair, and the function $\varepsilon_0(r)$ cannot be found from protein statistics directly. Actually,

$\varepsilon_0(r)$ can be adjusted by threading tests, but in this study we did not do this since the simplest assumption that

$$\varepsilon_0(r) = \begin{cases} 0, \text{ when } r > R_{\min} \\ +\infty, \text{ when } r \leqslant R_{\min} \end{cases} \quad (8a)$$

where $R_{\min}$ is an adjustable radius ($R_{\min}$ ⊤ 2.5–3.0 Å, see below) works well enough.

To calculate potentials using Equations 8 and 9, one needs to determine the threshold distances $R_{\alpha\beta}$. We used the estimate

$$R_{\alpha\beta} = R_\alpha' + R_\beta' \quad (10)$$

where $R_\alpha' = R_\alpha + \delta/2$ and $R_\beta' = R_\beta + \delta/2$ are 'effective' radii of residues α and β, respectively. For a 'covalent' residue radius, R_α, we simply took the maximum (over all residues of a given type α in a database) distance between the C_α (or C_β) atom and any other heavy atoms of the residue. To convert a 'covalent radius' R into something like the van der Waals radius R' of a residue, we add (δ/2) ⊤ 1.4 Å.

Short-range interactions depending on distance between residues

In this study, short-range interactions are defined as those between residues occupying positions i, $i + 2$; i, $i + 3$; i, $i + 4$ along a chain. This corresponds to $s_0 = 4$ (see Figure 1a).

To estimate these interactions, we neglect the unimportant distance-independent term ln$A$ and also the energy of indirect interactions, $\tilde{E}^s_{\alpha\beta}(r)$ (which is of a secondary importance since the residues close in a chain are also close in space) in Equation 3, and approximate the probability of finding a pair i, $i + s$ at a given distance r by the ratio of the total number of all i, $i + s$ residues pairs found at a distance r, to the total number of i, $i + s$ residue pairs found at all the distances.

Thus, for $s = 2,3,4$ we have

$$\varepsilon^s_{\alpha\beta}(r) = -RT_c \ln\left[\frac{N^s_{\alpha\beta}(r)}{N^{*s}_{\alpha\beta}(r)}\right] \quad (11)$$

where

$$N^s_{\alpha\beta}(r) \sum_{p=1}^{P} \sum_{i=1}^{N_p - s} \delta_{q_i\alpha}\delta_{q_{i+s}\beta}\,\theta\left(\tfrac{\Delta}{2} - |r_{i,\,i+s} - r|\right) \quad (12)$$

and

$$N^{*s}_{\alpha\beta}(r) = \sum_r N^s_{\alpha\beta}(r)\,\frac{\sum_\alpha \sum_\beta N^s_{\alpha\beta}(r)}{\sum_\alpha \sum_\beta \sum_r N^s_{\alpha\beta}(r)} \quad (13)$$

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

PDB code	Potential		
	$C_\beta 2$[a]	$C_\beta 2$[b]	$C_\alpha 2$[c]
1ins.A	423	82	216
1mlt.A	54	157	518
1gcn	2267	3554	535
1ins.B	173	2058	47
1ppt	39	79	784
2rhv.4	30	46	132
1bds	1	1	5
1crn	14	1	1
5rxn	2414	1	2
1fdx	28	1	2
1ovo.A	1	1	1
4pti	1	1	1
2mt2	1	1	1
2ebx	1	1	1
1cse.I	1	1	1
1sn3	10	1	1
1ctf	1	1	39
1hoe	25	43	73
2abx.A	71	2	5
3icb	3	1	2
2pka.A	1	1	1
351c	2	2	6
1cc5	12	1	1
2b5c	1	1	1
1hip	6	1	1
2gn5	35	80	895
3fxc	1	1	1
1hvp.A	3	8	1
1pcy	1	1	1
1wrp.R	1	1	1
4cyt.R	1	1	3
2ssi	1	1	1
2cdv	18	1	7
1rei.A	1	1	1
1acx	1	1	1
1cpv	1	1	1
2c2c	1	1	4
1hmq.A	1	1	1
2pab.A	1	1	1
1paz	1	1	1
155c	2	1	3
1pp2.R	1	1	1
1bp2	1	1	1
1rn3	1	1	1
2ccy.A	5	1	1
2aza.A	1	1	1
1lzl	1	1	1
3fxn	1	1	1
2hhb.A	1	1	1
2pka.B	1	1	1
2hhb.B	1	1	1
2lhb	1	1	1
2sod.O	1	1	1
1mbd	1	1	1
1lh4	1	1	1
4dfr.A	1	1	1
2lzm	1	1	1
2sga	1	1	1
3wga.A	1	1	1
2alp	1	1	1
1gcr	1	1	1
1hmg.B	14	1	1
2stv	1	1	1
3adk	1	1	1
4sbv.A	1	1	1
AV.[d]	3.0	2.0	2.5

[a]C_β-based potentials derived in Hendlich *et al.* (1990) at a resolution interval of 2 Å.

[b,c]C_β- and C_α-based potentials derived in this work at a resolution of 2 Å.

[d]Average position is defined as the geometrical mean: $\langle P\rangle = \exp\left[\sum_{i=0}^{N}(\ln P_i/N)\right]$, where P_i is the position of a protein i and N is the number of proteins.

Table IV. Average positions of the native conformation in the energy-sorted list of 65 proteins obtained with different combinations of C_α- and C_β-based potentials derived at a resolution interval of 2 Å

	C_α-based potentials	C_β-based potentials
Hendlich *et al.* (1990)	–	3.0
This work	2.5	2.0
Long-range only, derived with the reference state of Hendlich *et al.* (1990)[a]	30.2	4.9
Long-range only, derived in this work (Equation 8)	10.0	2.8
Short-range only ('direct'[b] and 'bending'[c] terms), this work	8.8	20.9
Short-range 'direct' terms only (Equation 11)	74.9	74.8
Short-range 'bending' terms only (Equations 14 and 15)	14.3	54.9

[a]The reference state of Hendlich *et al.* (1990) is calculated as: $N^*_{\alpha\beta}(r) = N^0_{\alpha\beta}(\leqslant R^*)[M(r)/M^0(\leqslant R^*)]$ (compare with our definition given by Equation 9); the value $R^* = \max\{R_{\alpha\beta}\}$ ($R^* = 18$ and 15 Å for C_α- and C_β-based potentials respectively).
[b]Short-range 'direct' terms correspond to the interactions shown in Figure 1a.
[c]Short-range 'bending' terms correspond to the interactions in Figure 1b.

For short-range interactions we distinguish between pairs αβ and βα.

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 Figure 1b); these residues determine the local chain stiffness.

To take into account these interactions, we follow the above approach and introduce two 'bending-energy' terms:

$$u^{(2)}_\alpha(r) = -RT_c \ln\left[\frac{\tilde{N}^{(2)}_\alpha(r)}{\tilde{N}^{*(2)}_\alpha(r)}\right] \tag{14}$$

and

$$u^{(3)}_{\alpha\beta}(r) = -RT_c \ln\left[\frac{\tilde{N}^{(3)}_{\alpha\beta}(r)}{\tilde{N}^{*(3)}_{\alpha\beta}(r)}\right] \tag{15}$$

where

$$\tilde{N}^{(2)}_\alpha(r) \sum_{p=1}^{P} \sum_{i=1}^{N_p-2} \delta_{q_{i+1}\alpha}\,\theta\left(\tfrac{\Delta}{2} - |r_{i,\,i+2} - r|\right) \tag{16}$$

$$\tilde{N}^{*(2)}_\alpha(r) = \sum_r \tilde{N}^{(2)}_\alpha(r)\,\frac{\sum_\alpha \tilde{N}^{(2)}_\alpha(r)}{\sum_r \sum_\alpha \tilde{N}^{(2)}_\alpha(r)} \tag{17}$$

and

$$\tilde{N}^{(3)}_{\alpha\beta}(r) = \sum_{p=1}^{P} \sum_{i=1}^{N_p-3} \delta_{q_{i+1}\alpha}\delta_{q_{i+2}\beta}\,\theta\left(\tfrac{\Delta}{2} - |r_{i,\,i+3} - r|\right) \tag{18}$$

$$\tilde{N}^{*(3)}_{\alpha\beta}(r) = \sum_r \tilde{N}^{(3)}_{\alpha\beta}(r)\,\frac{\sum_\alpha \sum_\beta \tilde{N}^{(3)}_{\alpha\beta}(r)}{\sum_\alpha \sum_\beta \sum_r \tilde{N}^{(3)}_{\alpha\beta}(r)} \tag{19}$$

Here $\tilde{N}^{(2)}_\alpha(r)$ is the number of pairs i, $i + 2$ with a distance r between i and $i + 2$ residues and the residue α in the $i + 1$ position; $\tilde{N}^{(3)}_{\alpha\beta}(r)$ is the number of i, $i + 3$ pairs with a distance r between i and $i + 3$ residues and residues α in $i + 1$ and β in $i + 2$ positions (see Figure 1b).

Sparse statistics

Above, all the potentials were obtained from equations having the general form

$$\varepsilon_x(r) = -RT_c \ln\left[\frac{N_x(r)}{N^*_x(r)}\right] \tag{20}$$

where $x = \alpha$ for the u_a potential and $x = \alpha\beta$ pair for all other $\varepsilon_{\alpha\beta}$ and $u_{\alpha\beta}$ potentials, while $N_x(r)$ and $N^*_x(r)$ are the observed and expected number of residue pairs, respectively (see Equations 8 and 13). Equation 20 is not applicable for the cases of sparse statistics when one or both of the distributions, $N_x(r)$ and $N^*_x(r)$, are equal to zero. In these cases we define potentials as follows:

$$\varepsilon_x(r) = +\infty, \text{ if } N^*_x(r) = 0 \tag{21}$$

$$\varepsilon_x(r) = RT_c N^*_x(r), \text{ if } N_x(r) = 0 \text{ and } N^*_x(r) \neq 0 \tag{22}$$

Equation 21 is obvious: in this way, we forbid inter-residue distances which, for any physical reason, are not observed in any protein structures (see above).

Equation 22 is rather arbitrary; we use it to obtain some kind of high energy and, simultaneously, to avoid an infinity which can be caused by sparse statistics, rather than the physical impossibility of particles at a distance r from each other.

The energy of a chain conformation is the sum of all the individual terms described above.

Statistical errors in potential estimates

The accuracy of phenomenological potentials depends on the size of the database used for their derivation. It is important for applications to have an estimate of the statistical error arising from the finite size of the database.

Such an estimate can be easily made in the following way: let us divide a database of protein structures into two approximately equal sub-databases, A and B, and let us derive two corresponding sets of potentials, namely sets A and B. Because of statistical fluctuations, potentials A and B will be slightly different. One can estimate an amplitude of statistical error for a potential $\varepsilon_x(r)$ as

$$\Delta\varepsilon_x = |\varepsilon^A_x - \varepsilon^B_x|/2 \tag{23}$$

where ε^A_x and ε^B_x are potentials corresponding to the databases A and B, respectively.

In the case of sparse statistics, when $N^A_x(r) = 0$ and/or $N^B_x(r) = 0$, the values of $RT_c N^{*A}_x(r)/2$ and/or $RT_c N^{*B}_x(r)/2$ are added to the value of $\Delta\varepsilon_x$.

Threading test for potentials

The accuracy of potentials is estimated using the threading test suggested by Hendlich *et al.* (1990). In this test, the energy of the native structure is compared with energies of alternative structures obtained by threading the native sequence through all possible structural conformations provided by backbones of a set of proteins. No gaps or insertions are allowed, thus a chain of N residues length can be threaded through a host

Table V. Characteristics of the native structure recognition in threading tests with potentials derived from two independent sets of proteins

Resolution (Å)	Proteins used in threading[b]	Structures used in threading[c]	Potentials derived from[d]	C_α potentials					C_β–C_α potentials[a]				
				Averaged position in energy list[e]			Averaged energy gap[f] (in RT_c)	*Z*-score[g]	Averaged position in energy list[e]			Averaged energy gap[f] (in RT_c)	*Z*-score[g]
				T	LR	SR			T	LR	SR		
0.25	50A	107A	107A	1.2	3.6	2.8	57.7	5.0	1.2	1.5	8.3	69.8	5.3
	50A	107A	107B	1.1	6.2	2.0	57.3	4.8	1.1	1.5	6.2	67.1	5.1
	50B	107B	107B	1.6	6.3	4.4	53.9	4.7	1.2	1.7	8.0	67.3	5.3
	50B	107B	107A	1.3	2.5	3.6	67.1	5.2	1.2	1.5	7.4	78.2	5.7
0.5	50A	107A	107A	1.0	3.0	1.8	68.1	5.1	1.1	1.5	3.8	80.5	5.6
	50A	107A	107B	1.0	4.6	1.8	60.1	4.7	1.1	1.5	3.4	76.2	5.5
	50B	107B	107B	1.4	4.2	3.5	60.3	4.7	1.1	1.5	5.4	73.6	5.6
	50B	107B	107A	1.3	2.0	3.3	71.9	5.2	1.1	1.3	5.7	85.7	5.9
1.0	50A	107A	107A	1.0	2.8	2.2	57.9	5.4	1.0	1.4	2.4	80.5	6.3
	50A	107A	107B	1.0	4.3	1.8	55.1	5.3	1.1	1.4	2.6	77.5	6.1
	50B	107B	107B	1.4	4.1	3.4	52.5	5.2	1.0	1.4	3.4	80.4	6.2
	50B	107B	107A	1.3	2.2	3.4	62.1	5.6	1.1	1.2	3.9	87.3	6.6
2.0	50A	107A	107A	1.1	2.9	2.5	46.2	5.4	1.1	1.4	3.0	71.9	6.3
	50A	107A	107B	1.4	5.3	3.3	36.1	5.0	1.1	1.5	4.2	65.9	6.0
	50B	107B	107B	1.5	4.1	3.8	38.8	5.1	1.1	1.3	5.1	67.5	6.2
	50B	107B	107A	1.3	2.3	5.5	49.0	5.6	1.1	1.2	4.4	77.4	6.6
3.0	50A	107A	107A	1.1	3.4	3.1	44.7	5.3	1.0	1.4	4.7	71.0	6.3
	50A	107A	107B	1.2	6.0	2.4	37.9	5.0	1.2	1.5	5.3	64.2	6.1
	50B	107B	107B	1.7	4.9	4.4	37.4	5.0	1.1	1.5	4.0	70.6	6.3
	50B	107B	107A	1.4	2.6	4.8	50.5	5.5	1.1	1.3	5.4	80.3	6.6

[a]Short-range bending potential is applied to positions of C_α atoms.
[b]50A and 50B correspond to the 50 proteins of sets A and B respectively, which were used in threading (Table I).
[c,d]107A and 107B correspond to the 107 proteins of sets A and B, respectively, which were used as sources of structures in threading and as datasets for extracting of potentials.
[e]Average position is defined as in Table III; T, LR, SR correspond to the total energy and energy of long- and short-range interactions, respectively.
[f]Energy gap is the difference between the lowest energy of alternative structures and the energy of the native structures.
[g]*Z*-score is defined as $(E_{av}–E_{nat}/\sigma$, where E_{av} is an average threading energy, E_{nat} is the native structure energy and σ is the standard deviation of threading energies.

Table VI. Average characteristics of threading tests with 100 proteins obtained for C_α- and C_β–C_α-based potentials[a] at different values of the resolution intervals

Potentials	Resolution (Å)	Average position[b]			Average energy gap[c]		Average *Z*-score[d]	Average N_Z[e]
		T	LR	SR	In RT_c	In σ units		
C_α-based	3.0	1.4	5.4	4.0	37.8	1.3	5.3	7.4×10^6
	2.0	1.3	3.7	3.3	39.5	1.4	5.3	1.5×10^7
	1.0[f]	**1.2**	**3.8**	**2.3**	**51.4**	**1.6**	**5.4**	$\mathbf{3.0\times10^7}$
	0.5	1.2	3.8	2.1	60.0	1.3	4.6	4.5×10^5
	0.25	1.3	4.4	2.7	59.2	1.2	4.7	7.8×10^5
C_β–C_α-based	3.0	1.1	1.7	4.3	65.0	2.2	6.1	2.0×10^9
	2.0	1.1	1.4	3.9	67.2	2.4	6.4	1.3×10^{10}
	1.0	**1.03**	**1.4**	**2.4**	**80.8**	**2.6**	**6.5**	$\mathbf{2.7\times10^{10}}$
	0.5	1.1	1.5	3.6	79.0	2.1	5.9	6.2×10^8
	0.25	1.2	1.6	5.8	72.4	1.8	5.7	1.2×10^8

[a,b,c]See the corresponding footnotes in Table V.

[d]Average *Z*-score is defined as $\langle Z\rangle = \sqrt{(1/102)\sum_{i=1}^{102} Z_i^2}$.

[e]N_Z values (defined by Equation 25) are averaged geometrically (see Table III).
[f]The values in bold summarize the results presented in Table VII.

protein molecule of M residues length in $M - N + 1$ different ways. Since glycine residues have no C_β atoms (which are necessary for threading with C_β atom-based potentials), we constructed virtual C_β atoms for all glycine residues of the threading database.

For a strict test one needs two sets of proteins: one for derivation of potentials and another for threading. Hendlich *et al.* (1990) used a simplified testing procedure. From the entire set of 101 proteins they chose 65 protein chains of less than 200 residues. For each of these 65 proteins, the remaining 100 proteins were used both for deriving potentials and as a source of alternative structures in threading. For comparison of our potentials with those used in their work, we used this way of testing.

However, to avoid possible systematic error, which could be caused by using the same protein structures for derivation of potentials and in threading, we also performed a more correct test. We took two independent data sets, A and B, one for deriving potentials and another for threading. In these tests we used a database of 214 non-homologous proteins (see Table I) of resolution better than 2.5 Å and with no structural defects (chain gaps, significant distortions of bond lengths, absent atoms), chosen from the list of 331 no- or low-homology proteins provided by Hobohm *et al.* (1992).

This database of 214 proteins was also used for extracting the maximum 'covalent' radii of the residues (see Equation 10 and Table II).

Results and discussion

In order to study the accuracy of our potentials, we first repeated the test done by Hendlich *et al.* (1990) with our energy functions. These have been derived for the force centers positioned both at C_α and at C_β atoms (for glycines having no C_β atom, the force center is always positioned in the C_α atom). The potentials were derived at a resolution of 2 Å as in Hendlich *et al.* (1990).

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

One can see from Table III 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), neither of the potentials gives a satisfactory ranking. The conformations of these molecules are probably stabilized by interactions within molecular complexes. For larger proteins our new potentials show noticeably better accuracy than those used by Hendlich *et al.* (1990).

To analyse the contribution of different energy terms in the recognition of protein structure, in Table IV we compare the averaged positions of the 65 native structures given separately by each of the energy terms. For long-range (LR) energy terms where the reference ('zero energy') state is important, we compute the results for two definitions of the reference state: one is given by Equation 9 in this work and the other (see footnote to Table IV) is that used by Sippl (1990) and Hendlich *et al.* (1990).

The results in Table IV show that long-range energies derived using the reference state of Equation 9 are significantly more accurate than those derived using the reference state of Hendlich *et al.* (1990).

One can also see that for both C_α- and C_β-based potentials the main contribution to protein structure recognition arises primarily from long-range interactions and bending energy. The C_β-based distance-dependent potentials are more accurate than the C_α-based potentials because they approximate better the relative positions of side chains; the bending energy is more accurate for C_α-based potentials (Kocher *et al.*, 1994) because the positions of C_α atoms can more accurately approximate the chain bending.

As bending energy terms are more effective using the C_α atoms, in the following tests we used a combination of potentials (Kocher *et al.*, 1994): C_β atom-based long-range and 'direct' short-range interactions with C_α-based bending energies (below we refer to these as 'C_β potentials').

The database used by Hendlich *et al.* (1990) is relatively small, so it is of interest to see what results one obtains using a larger database. For this purpose, 214 proteins (see Table I) were selected from the low-homology protein list of Hobohm *et al.* (1992). However, before repeating the threading tests on

Table VII. Characteristics of the native conformation position in the energy-sorted list for 100 proteins obtained with C_α- and C_β–C_α-based potentials derived at a resolution interval of 1.0 Å[a]

PDB code	Threadings	C_α potentials			C_β–C_α potentials			PDB code	Threadings	C_α potentials			C_β–C_α potentials		
		Position	Energy gap	*Z*-score	Position	Energy gap	*Z*-score			Position	Energy gap	*Z*-score	Position	Energy gap	*Z*-score
1tgx.A	43887	1	8.2	4.3	1	1.1	3.9	2end	28692	1	49.6	5.6	1	49.6	5.6
4mt2	43673	2	–4.2	4.0	1	44.3	7.5	4fxn	28528	1	93.4	6.5	1	93.4	6.5
1cse.I	43248	1	21.3	4.5	1	32.1	5.8	1pbx.A	27879	1	24.5	4.9	1	24.5	4.9
1ptx	43036	1	24.1	6.2	1	35.3	6.7	1nhk.L	27717	1	59.4	5.5	1	59.4	5.5
1cks.B	40095	1	.1	3.7	1	31.3	4.9	1lpe	27556	1	61.6	4.3	1	61.6	4.3
lzaa.C	38631	14	–9.7	3.0	3	–18.8	3.4	3sdh.A	27396	1	31.7	4.4	1	31.7	4.4
1cyo	38006	1	29.9	4.4	1	33.0	4.6	1lba	27237	1	91.1	6.3	1	91.1	6.3
1mol.A	36763	4	–5.5	3.3	1	22.8	4.3	2fal	27237	1	84.4	5.8	1	84.4	5.8
1fxi.A	36350	1	49.8	5.6	1	49.8	6.0	8atc.B	27237	1	69.2	5.8	1	69.2	5.8
7pcy	35939	1	43.8	5.7	1	89.1	7.7	1ash	27079	1	74.0	4.8	1	74.0	4.8
2hpe.A	35734	1	38.9	5.2	1	71.9	7.2	2hbg	27079	1	61.5	5.6	1	61.5	5.6
1aya.A	35327	1	34.4	5.7	1	69.1	6.6	2mta.C	27079	1	70.2	5.7	1	70.2	5.7
2kau.B	35327	1	6.3	3.9	1	35.7	5.4	1osa	26924	1	117.4	5.8	1	117.4	5.8
1lts.D	34923	1	46.6	5.5	1	42.0	5.4	1rtm.1	26772	1	74.6	6.2	1	74.6	6.2
1cmb.A	34722	1	10.7	3.4	1	22.5	4.2	1sxc.A	26471	1	83.0	6.5	1	83.0	6.5
9rnt	34722	1	28.1	4.7	1	81.1	7.9	1wht.B	26172	1	42.8	5.3	1	42.8	5.3
256b.A	34324	1	14.6	3.6	1	66.2	5.3	1mls	26023	1	47.2	5.1	1	47.2	5.1
2fd2	34324	1	74.1	5.8	1	114.9	7.4	2rn2	25875	1	72.8	5.4	1	72.8	5.4
1bet	34126	1	48.6	5.8	1	35.9	6.5	1hlb	25582	1	53.6	4.7	1	53.6	4.7
2cdv	34126	118	–40.1	2.5	1	15.9	3.8	1mup	25582	1	39.7	4.6	1	39.7	4.6
3sic.I	34126	1	41.2	5.8	1	64.0	7.5	1gpr	25436	1	83.3	6.3	1	83.3	6.3
1cew.I	33930	2	–16.7	4.2	1	20.0	5.5	1hjr.A	25436	1	86.6	6.6	1	86.6	6.6
1rtp.1	33737	1	19.1	3.6	1	63.8	5.7	1mnc	25436	1	90.3	6.5	1	90.3	6.5
1ccr	33354	1	5.4	3.9	1	23.3	4.9	131l	24869	1	93.7	5.6	1	93.7	5.6
2tgi	33163	1	31.2	5.5	1	37.4	6.4	1cpc.A	24869	1	42.1	5.5	1	42.1	5.5
1dyn.A	32973	89	–26.0	2.3	1	22.0	4.6	1mmo.G	24869	1	37.8	4.0	1	37.8	4.0
2chs.A	32784	1	31.2	4.7	1	69.1	5.9	2cpl	24590	1	98.0	6.6	1	98.0	6.6
2hmz.A	32973	8	–6.8	3.1	1	23.5	4.5	1rcf	23904	1	88.3	6.4	1	88.3	6.4
1gmf.A	31853	1	37.9	5.2	1	69.4	6.3	1cpc.B	23495	1	64.6	6.0	1	64.6	6.0
2rsl.B	31667	1	59.1	5.2	1	65.7	5.7	1lki	23495	1	102.4	6.6	1	102.4	6.6
2pf1	31482	1	23.4	4.6	1	57.6	5.9	2scp.A	23225	1	98.0	6.1	1	98.0	6.1
1bp2	31115	1	46.3	5.3	1	66.8	6.3	4gcr	23225	1	115.5	8.4	1	115.5	8.4
1htm.D	31115	1	14.9	3.4	1	20.1	4.2	3cd4	22695	1	49.0	4.6	1	49.0	4.6
4fgf	30932	1	26.6	4.3	1	19.1	4.4	1cau.A	22301	1	1.0	3.6	1	1.0	3.6
7rsa	30932	1	34.8	4.9	1	86.3	7.3	1cau.B	21913	1	29.2	4.2	1	29.2	4.2
2acg	30751	1	71.9	7.6	1	130.8	9.0	153l	21784	1	17.3	4.2	1	17.3	4.2
1ttb.A	30396	1	48.9	5.4	1	53.8	6.1	1lts.A	21784	1	114.8	7.2	1	114.8	7.2
2ccy.A	30396	1	19.9	3.8	1	62.8	5.1	2sas	21784	1	91.1	5.8	1	91.1	5.8
3chy	30219	1	50.9	4.6	1	105.4	6.4	1gky	21656	1	149.0	7.3	1	149.0	7.3
1msc	30044	2	–2.3	4.0	8	–23.8	3.5	1knb	21656	1	99.2	6.6	1	99.2	6.6
1rcb	30044	1	20.2	4.2	1	75.6	5.9	1dsb.A	21408	1	97.8	6.0	1	97.8	6.0
2aza.A	30044	1	44.1	5.2	1	49.8	5.8	1isc.A	20919	1	134.4	8.5	1	134.4	8.5
1hmt	29699	1	5.9	3.5	1	25.1	4.5	1tss.A	20676	1	52.4	4.6	1	52.4	4.6
1htp	29699	1	19.1	4.6	1	55.1	5.7	1cus	20315	1	136.6	8.7	1	136.6	8.7
1lis	29699	1	34.9	5.2	1	65.3	6.6	2alp	20195	1	102.1	7.5	1	102.1	7.5
1poc	29192	1	29.4	5.0	1	68.8	6.5	1iae	19958	1	99.7	7.3	1	99.7	7.3
1rsy	29024	1	9.4	3.6	1	21.5	4.7	1sac.A	19489	1	134.0	7.7	1	134.0	7.7
1snc	29024	1	14.4	3.9	1	29.1	4.6	1cfb	19372	1	80.6	5.5	1	80.6	5.5
1eca	28857	1	52.4	5.6	1	89.5	7.5	1huc.B	19372	1	115.4	6.5	1	115.4	6.5

[a]See footnotes to Table IV.
The energy gaps are given in RT_c units in this table.

the new database, we tried to estimate how the limited size of a database influences the accuracy of the potentials. There are two sources of errors: random errors, coming from poor statistics, and systematic error, which arises when the same protein set is used both for derivation of potentials and for the threading test.

Random errors necessitate optimization of the size of the resolution interval Δ in order to obtain the most accurate potentials: a wider interval will resolve less detail of the potential, a narrower interval will have poorer statistics, and therefore larger random errors (to estimate the values of these errors, see the corresponding section in Methods).

To reduce the probability of systematic error and also to validate using the same set of protein structures both for the derivation of potentials and for the threading test, we carried out the following experiment. The database of 214 proteins was divided into two subsets, A and B, of 107 proteins each. The potentials derived from set A were used to thread proteins of set A and, separately, of set B, and the proteins of both sets A and B were tested with the potentials derived from set B. For threading we chose 50 proteins of residue length 60–205 from each of sets A and B. Averaged characteristics of the native structure positional rank obtained in these tests are given in Table V.

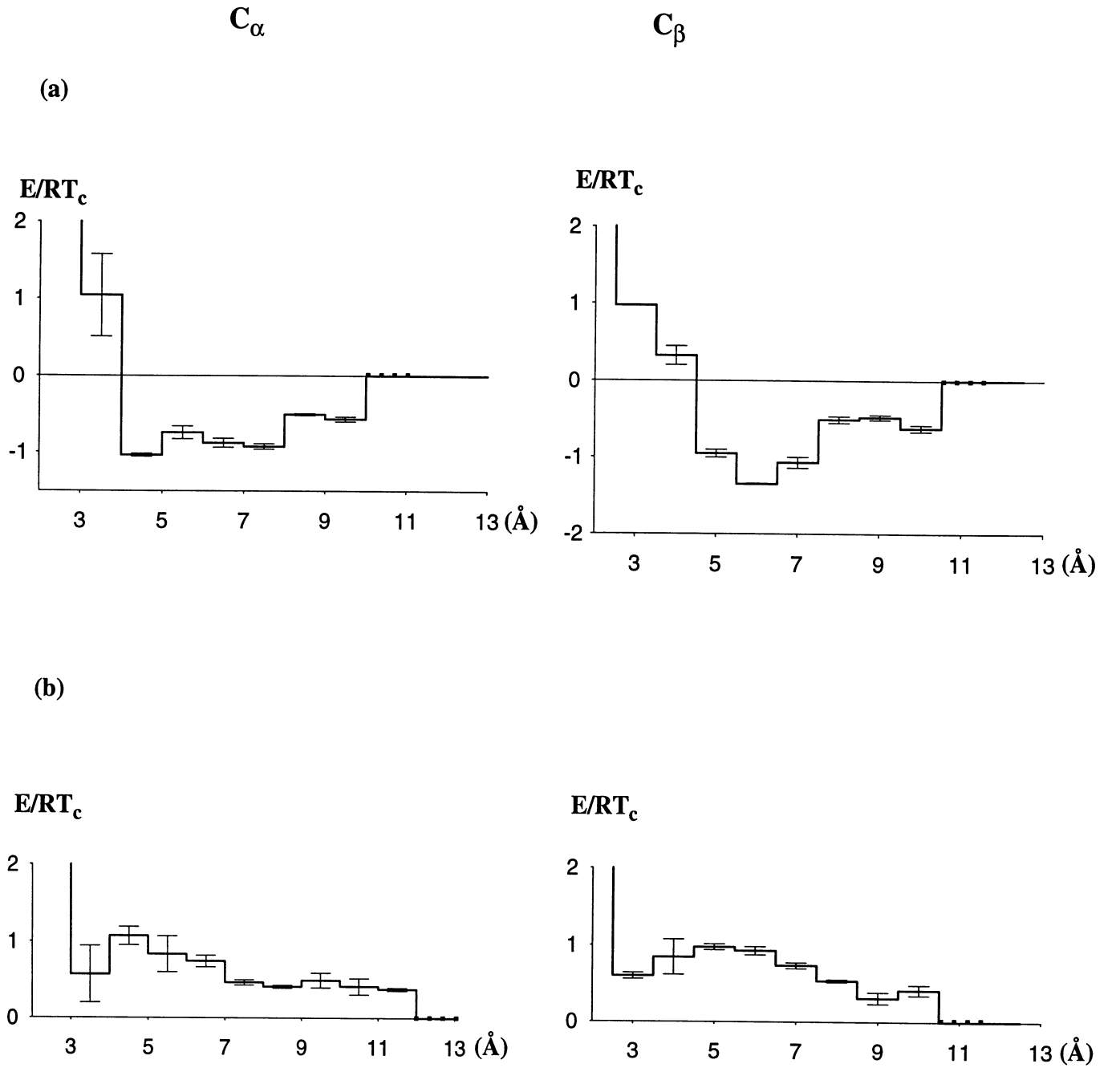

Fig. 3. Long-range potentials for (**a**) Val–Ile and (**b**) Asp–Asp residue pairs derived from the database of 214 proteins (Table I) at a resolution of 1 Å. Inaccuracies of potentials caused by limited statistics are shown by thin error bars; the estimates were obtained using Equation 23. Errors of amplitude less than $0.05RT_c$ are not shown. Long-range potentials are infinitely high at $r \leqslant R_{\min} = 3.0$ Å for C_α-based potentials and $r \leqslant R_{\min} = 2.5$ Å for C_β-based potentials. The dots show that part of the potential which is taken as zero at $r \geqslant R_\alpha + R_\beta$.

One can see that the averaged characteristics obtained in the 'cross' tests and in the 'direct' tests are close to each other for both C_α- and C_β-based potentials. Some of the observed differences could be caused by statistical fluctuations in the databases rather than by significant systematic deviations between potentials derived from the 'self' and 'other' databases.

These results enabled us to do more complete testing on the total set of 214 proteins, using them as in Hendlich *et al.* (1990) for both derivation of potentials and for threading.

A typical example of the energy distribution for the ferredoxin molecule (2fd2) in this threading experiment is shown in Figure 2.

The results of these experiments, threading 100 proteins chains, are given in Tables VI and VII.

Table VI shows how accuracy of the potentials depends on the size of the resolution interval. This table shows that measures such as (i) average ranking, (ii) average energy gap and (iii) average relative deviation of the native structure energy from the mean energy of alternative structures (*Z*-score; see the definition in Table V) are optimal in an interval of ~1.0 Å for both C_α- and C_β-based potentials. One can see that

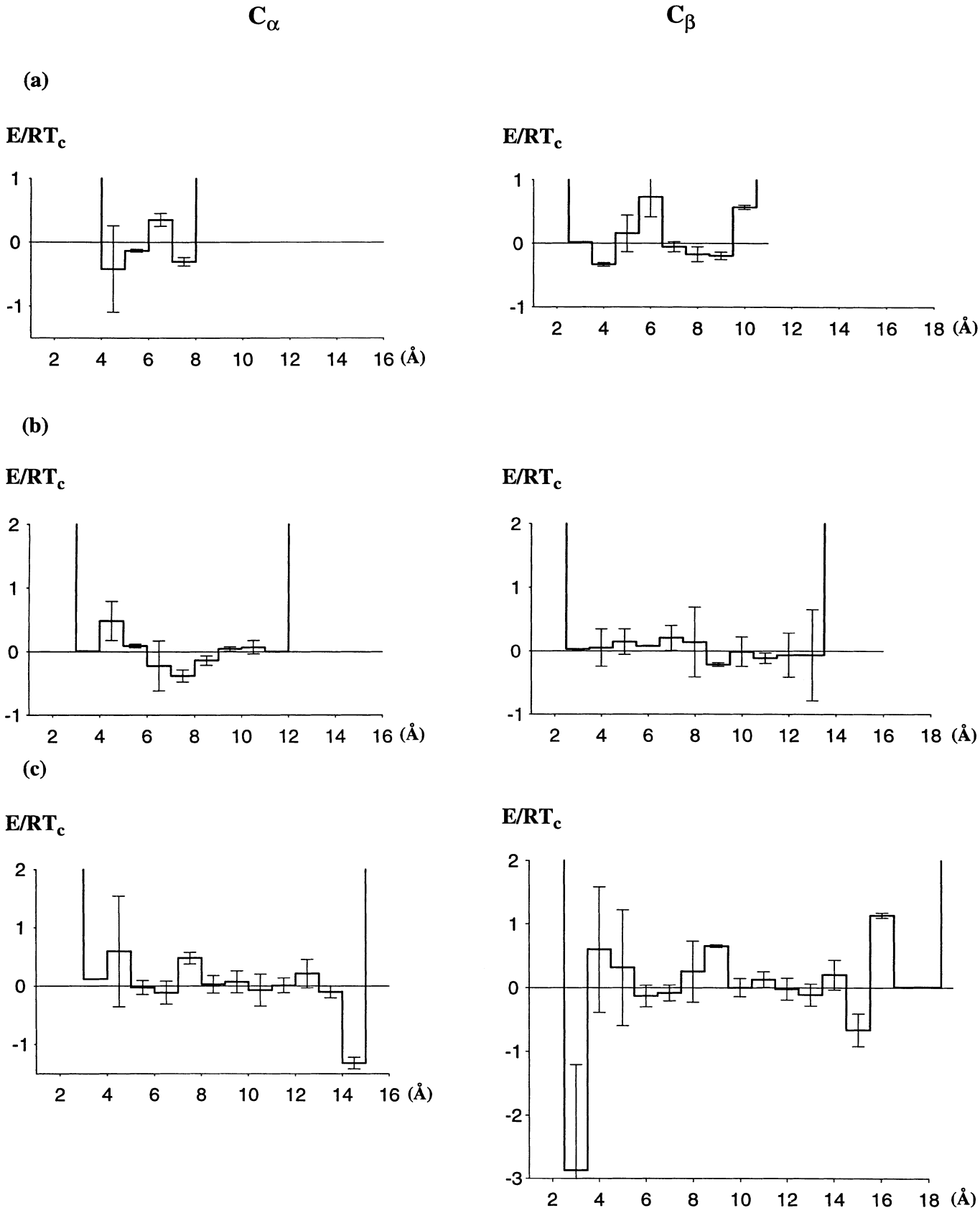

Fig. 4. Short-range distance-dependent pairwise potentials. The potentials are given for the Ala–Ser residue pair; they are derived from the database of 214 proteins at a resolution of 1 Å: (**a**), (**b**) and (**c**) correspond to *i*, *i* + 2; *i*, *i* + 3 and *i*, *i* + 4 types of short-range interactions respectively (see Figure 1a). Statistical inaccuracy of potentials is shown by error bars; errors of amplitude less than $0.05RT_c$ are not shown. Potentials are infinitely high at $r \leqslant R_{min}$ and $r \geqslant R_{max}$. For C_α-based potentials R_{min} = 4, 3 and 3 Å and R_{max} = 8, 12 and 15 Å for *i*, *i* + 2, *i*, *i* + 3 and *i*, *i* + 4 types of short-range interactions, respectively. For C_β-based potentials the corresponding values are R_{min} = 2.5, 2.5 and 2.5 Å and R_{max} = 10.5, 13.5 and 18.5 Å.

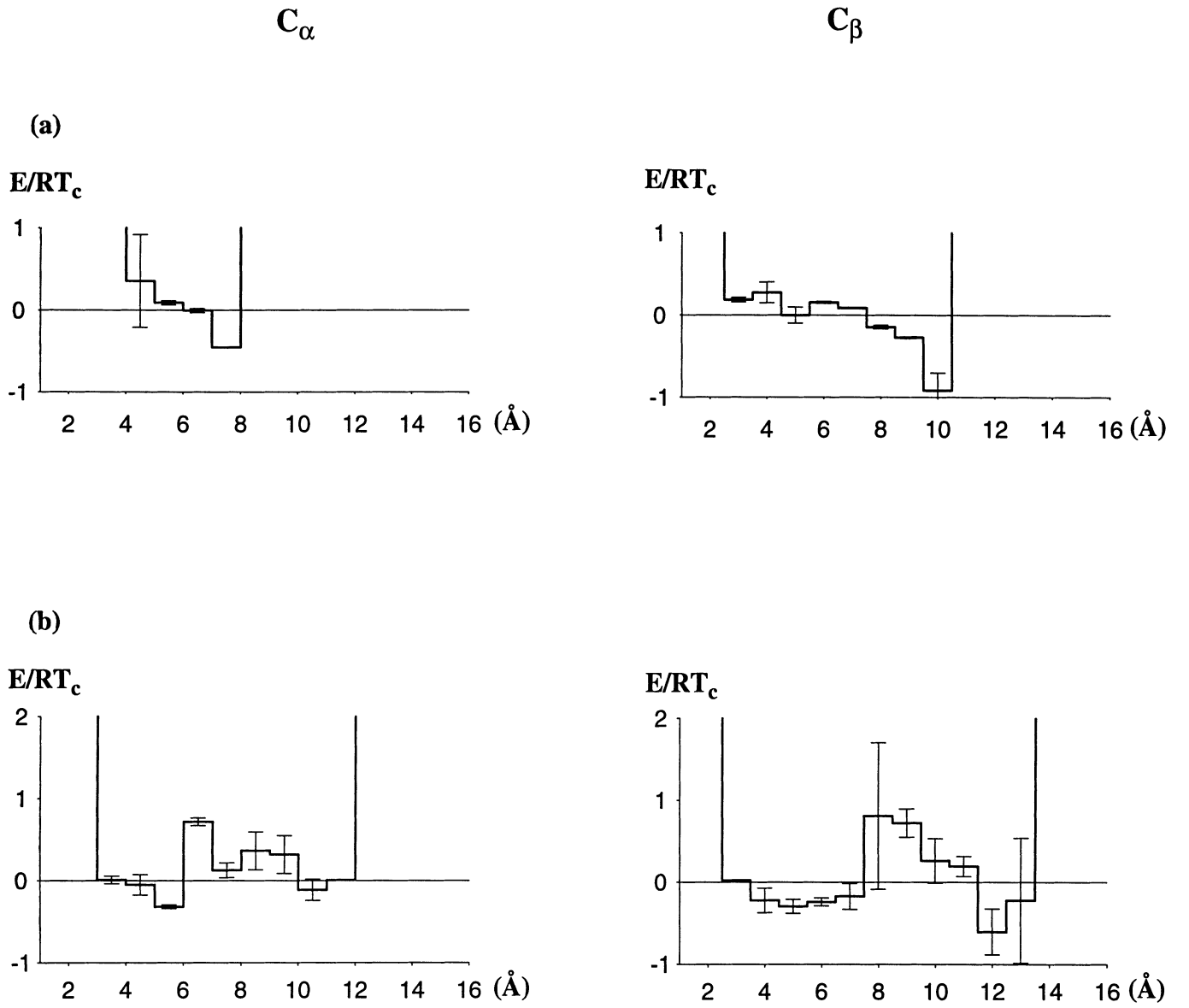

Fig. 5. Short-range bending potentials derived from the database of 214 proteins at a resolution of 1 Å: (**a**) and (**b**) correspond, respectively, to Ser residue and to the Ala–Ser residue pair occupying an intervening position (see Figure 1b). Error bars show statistical inaccuracy of potentials; errors of amplitude less than $0.05RT_c$ are not shown. Potentials are infinitely high at $r \leqslant R_{min}$ and $r \geqslant R_{max}$; values of R_{min} and R_{max} are the same as in i, $i + 2$ and i, $i + 3$ types of short-range interactions.

short-range potentials are more sensitive to resolution than long-range potentials, but their contribution to the overall accuracy at the optimal resolution is not worse than that of long-range potentials.

Plots of typical potentials derived from the data set of 214 proteins at a resolution of 1 Å are given in Figures 3–5.

One can note a significant difference between long-range (Figure 3) and short-range (Figures 4 and 5) potentials. Long-range potentials change relatively smoothly with distance and, in essence, have one energy minimum for attractive (usually hydrophobic) residue pairs and no minimum for repelling pairs; the C_β potentials look 'sharper' than C_α potentials.

Short-range potentials are characterized by more abrupt changes; they can have more than one local minimum, separated by barriers. Also, it is worth noting that because of the hard-core inter-atom repulsion, both long-range and short-range potential wells are bounded at R_{min} = 2.5 Å for C_β-based potentials and 3 Å for C_α-based potentials; a prohibition against chain breaking additionally restricts short-range potentials at long distances.

The statistical error estimates, calculated using Equation 23 with the potential sets A and B, are shown in Figures 3–5 by the corresponding error bars.

One can see differences in the amplitudes of the statistical errors, which are moderate for long-range interactions and vary significantly for short-range interactions. Thus, one can expect an improvement in the accuracy of short-range potentials with an increase in the size of the protein database.

The detailed results of the threading experiment for 100 proteins with C_α- and C_β-based potentials, derived at a resolution interval of 1.0 Å, are given in Table VII. The potentials successfully recognize the native structure: 92 proteins for C_α-based potentials and 98 proteins for C_β-based potentials were evaluated with the lowest energy for their native structures.

Since all of the above energy estimates were made with approximate energy functions, there is always a chance of finding a structure with lower energy than a given native one considering more extensive ensembles of structures.

Table VII shows large energy gaps between the native and competing folds for almost all the protein chains tested. However, these gaps depend on the number of alternatives tested. Since the energies of alternative structures have virtually a Gaussian distribution (Figure 2), one can estimate the probability of finding a structure with energy less than a given native one as

$$p(Z) = (1/\sqrt{2\pi}) \int_{Z}^{+\infty} e^{-\frac{t^2}{2}} dt \qquad (24)$$

where Z is a Z-score (see Table V). Thus, to find an energy lower than the energy of a given native structure, one needs to look through N_Z random structures:

$$N_Z = 1/p(Z) \qquad (25)$$

Having Z-score values obtained for C_α and C_β potentials and assuming that structures obtained in threading give representative ensembles of misfolded protein-like structures, we found N_Z values for each of the 100 proteins tested. The geometric averaging over N_Z values gives $<N_Z> \top 3\times10^7$ for C_α potentials and $<N_Z> \top 3\times10^{10}$ for C_β potentials. Given an average chain length of 134 residues, these numbers show that one can predict a protein fold only if the average number of possible backbone conformations per residue does not exceed $10^{10.5/134} = 1.2$. For globular folds where backbone conformations are not independent, this crucial number is not yet known (for a coil, where backbone conformations are independent, there are at least three conformations per residue: α_R, α_L and β). Since the backbone conformations used for threading represent only a portion of the globular folds and since they are not necessarily compact, the above estimates indicate that our potentials are adequate for recognition of the native fold within some restricted set of folds, rather than for distinguishing the native fold from of all possible folds.

Conclusion

We have developed a consistent approach to derive phenomenological energy functions using the previously established theory of Boltzman-like statistics of protein structure.

We have tested the approach by deriving potentials using the positions of C_α and C_β atoms. The energy function includes both long-range interactions between residues which are remote along a chain and short-range interactions between 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 (~1 Å) is optimized to preserve as much detail as possible without introducing excessive error due to limited statistics.

Our studies show that long- and short-range interactions are equally important in protein structure recognition. Since statistics for the short-range interactions are poorer than those for the long-range interactions, short-range interactions become a 'bottle-neck' for the improvement of potential function accuracy.

In estimating the role of simplified pairwise potentials for the protein folding problem, one should not presume to explain with them all of the details of protein structure. However, these potentials can be useful for efficient discrimination of the tiny fraction of most favorable conformations from the vast majority of the other conformations.

Acknowledgements

This work was supported by NIH Grant PO1GM38794 (to A.J.O.). The research of A.V.F. 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. This is manuscript No. 10641-MB from the Scripps Research Institute.

References

Chou,P.Y. and Fasman,G.D. (1974) *Biochemistry*, **13**, 211–222.
Finkelstein,A., Badretdinov,A. and Gutin,A. (1995a) *Proteins*, **23**, 142–150.
Finkelstein,A., Badretdinov,A. and Gutin,A. (1995b) *Proteins*, **23**, 151–162.
Godzik,A., Kolinski,A. and Skolnick,J. (1995) *Protein Sci.*, **4**, 2107–2117.
Hendlich,M., Lackner,P., Weitckus,S., Floeckner,H., Froschauer,R., Gottsbacher,K., Casari,G. and Sippl,M. (1990) *J. Mol. Biol.*, **216**, 167–180.
Hobohm,U., Scharf,M., Schneider,R. and Sander,C. (1992) *Protein Sci.*, **1**, 409–417.
Jernigan,R. and Bahar,I. (1996) *Curr. Opin. Struct. Biol.*, **6**, 195–209.
Kocher,J.A., Reoman,M.J. and Wodak,S.J. (1994) *J. Mol. Biol.*, **235**, 1598–1613.
Miyazawa,S. and Jernigan,R. (1996) *J. Mol. Biol.*, **256**, 623–644.
Pohl,F.M. (1971) *Nature New Biol.*, **234**, 277–279.
Ptitsyn,O.B. and Finkelstein,A.V. (1970) *Biofizika*, **15**, 757–767.
Reva,B.A., Finkelstein,A.V., Sanner,M.F. and Olson,A.J. (1997) In *Proceedings of Pacific Symposium on Biomolecular Computations*, World Scientific Publishing, Singapore, pp. 373–384.
Rooman,J. and Wodak,S. (1995) *Protein Engng*, **8**, 849–858.
Sternberg,M.J. (1986) *Anti-Cancer Drug Des.*, **1**, 169–178.
Sippl,M.J. (1990) *J. Mol. Biol.*, **213**, 859–883.
Sippl,M.J. (1993) *J. Comput.-Aided Mol. Des.*, **7**, 473–501.
Sippl,M.J. (1995) *Curr. Opin. Struct. Biol.*, **5**, 229–235.
Thomas,P.D. and Dill,K.A. (1996) *J. Mol. Biol.*, **257**, 457–469.

Received February 18, 1997; revised April 2, 1997; accepted April 15, 1997