Distance-Dependent, Pair Potential for Protein Folding: Results From Linear Optimization

Dror Tobi[1] **and Ron Elber**[1,2]*
[1]*Department of Biological Chemistry, The Hebrew University, Jerusalem, Israel*
[2]*Department of Computer Science, Cornell University, Ithaca, New York*

ABSTRACT **The results of an optimization of a folding potential are reported. The complete energy function is modeled as a sum of pairwise interactions with a flexible functional form. The relevant distance between two amino acids (2 – 9 Å) is divided into 13 intervals, and the energy of each interval is optimized independently. We show, in accord with a previous publication (Tobi et al., Proteins 2000;40:71–85) that it is impossible to find a pair potential with the above flexible form that recognizes all native folds. Nevertheless, a potential that rates correctly a subset of the decoy structures was constructed and optimized. The resulting potential is compared with a distance-dependent statistical potential of Bahar and Jernigan. It is further tested against decoy structures that were created in the Levitt's group. On average, the new potential places native shapes lower in energy and provides higher Z scores than other potentials. Proteins 2000; 41:40–46.** © 2000 Wiley-Liss, Inc.

Key words: fold recognition; score function; Z scores; decoy structures

INTRODUCTION

Recently we showed that linear optimization of parameters produces contact potentials that are comparable with contact potentials obtained with other approaches.[1] We further conjecture that it is impossible to find a sum of pair potentials with a cutoff of 9 Å that "solves" the protein-folding problem. An energy function of the form $U=\Sigma_{k>l}u_{kl}(r_{kl})$ fails to detect some native folds. (The distance r_{kl} is between the geometric centers of a pair of amino acids, and the pair potential u_{kl} (r_{kl}) is an arbitrary function of r_{kl} chosen to make the energy of the native fold as low as possible). $u_{kl}(r_{kl})$ depends on the type of the two amino acids in contact u_{kl} (r_{kl}). For 20 amino acids we assume 210 types of contacts. We comment that the folding potential can recognize a large class of proteins: small and large, with and without disulfide bonds, and proteins that are part of aggregates. Hence, the recognition potential is going beyond what is expected from a folding model in the chemical physics sense.

In the present manuscript we report the best pair potential that we were able to find. The proposed energy function improves the average position of the native shape and the Z score compared with other potentials that we examined.

The linear programming (LP) approach to determine potentials for protein folding was introduced by Maiorov and Crippen.[2] Vendruscolo and co-workers[3] used LP extensively to design contact potentials. A contact potential assigns a single "contact" energy to two amino acids that are close to each other. Alternatively, it is set to zero if the two amino acids are not close enough. Recently, we used the LP approach[1] to create yet another set of contact potentials (similar to the studies of Vendruscolo et al.[3]). We further analyzed in detail the information content of contact matrices and explore the feasibility problem of general pair potentials. Here we expand our study to a general pair potential with explicit distance dependence.

In the LP approach we solve the following set of inequalities:

$$E(S_n, X_j; P) - E(S_n, X_n; P) > 0 \quad j = 1, \ldots, J \qquad n = 1, \ldots, N \quad j \neq n. \quad (1)$$

The sets of coordinates $\{X_j\}$ are decoy structures. S_n and X_n are the sequence and the coordinate set of a native protein. The vector P contains the parameters (the unknowns) that we want to determine to satisfy the set of inequalities in Eq. (1). Eq. (1) states that the energy of the native sequence embedded in the native structure must be lower than the energies of the native sequence embedded in all other shapes. The parameters must be chosen to satisfy this "statement." To make the determination of the parameters relatively easy, the following expansion is used:

$$E(S_m, X_j; P) = \sum_{\lambda} p_\lambda f_\lambda(S_m, X_j). \quad (2)$$

The components of the vector P are the p_λ-s that we want to determine. At present the basis set in which we expand the energy function, f_λ (S_m, X_j) is not specified. When Eq. (2) is substituted into Eq. (1), we obtain a set of J **linear** inequalities for all the p_λ parameters. Linear inequalities can be solved efficiently.

Grant sponsor: NIH NCRR; Grant sponsor: DARPA.

The Supplementary Material referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/index.html.

*Correspondence to: Ron Elber, Department of Computer Science, Cornell University, 4130 Upson Hall, Ithaca, NY 14853. E-mail: ron@cs.cornell.edu

Received 23 March 2000; Accepted 18 May 2000

The requirement that the energy of the native fold is lower than the energies of other structures is most sensitive to the tail of the distribution, i.e., to decoy structures with energies near the energy of the native shape. This is in contrast to other optimization methods to determine parameters[4–6] that use the first and second moments of the energy distribution to define a target for optimization. Averages over the whole distribution are less sensitive to the addition of a few new constraints at the tail (a few more decoy structures). However, a few new constraints can change Eq. (1) from being feasible to being infeasible (!) or profoundly modify the volume of possible parameters. A disadvantage of the LP approach is that it is considerably more expensive, considerably more computer resources are required to obtain one solution. Nevertheless, the promise of the LP approach for very large sets encouraged us to use it to further advance the quality of folding potentials when millions and tens of millions of constraints are included.

FORMULATION OF THE POTENTIAL

Our model potential (TE-13) is based on Eqs. (1) and (2). Below we explain its functional form. A crucial step is the choice of the "basis set" f_λ (S_m, X_j) that expands the true potential. Here it consists of pair interactions. We denote by $u_{\alpha\beta}$ (r) a step potential between a pair of amino acids. The distance between the geometric centers of two amino acid side chains, r, is divided into 13 steps between 2 and 9 Å. The first step along r is between 2 and 3 Å, and the rest of the 12 steps are for half an angstrom each. Each of the $u_{\alpha\beta}$ (r) (as a function of the index β) is 1 only at one of the windows (steps), and it is zero elsewhere. That is

$$u_{\alpha\beta}(r) = \left\{ \begin{array}{ll} 1 & q(\beta) - \Delta_\beta < r \le q(\beta) + \Delta_\beta \\ 0 & \text{otherwise} \end{array} \right\}$$

$$q(\beta) = 2.5, 3.25, 3.75, \ldots, 8.75$$

$$\Delta_\beta = 0.5, 0.25, 0.25, \ldots, 0.25 \quad (3)$$

The coordinate at the center of each window is q (β), 2Δ is its width. The total potential energy is

$$E(S_m, X_j; P) = \sum_{\alpha,\beta} p_{\alpha\beta} n_\alpha u_{\alpha\beta}(r) \equiv \sum_\lambda p_\lambda n_\lambda^{(j)}. \quad (4)$$

The index α parameterizes the types of the two interacting amino acids. n_α is the number of contacts of a specific type found when threading the complete sequence S_m into the known shape X_j (e.g., the number of lysine-serine contacts in the shape X_j). n_α and $u_{\alpha\beta}$ are combined together to form $n_\lambda^{(j)}$, the number of contacts of a specific type and at a specific distance (λ) of structure j. For each of the λ-s there is a corresponding independent parameter p_λ.

The set of inequalities (for the p_λ-s) that we need to solve is:

$$\left\{ \sum_\lambda p_\lambda (n_\lambda^{(j)} - n_\lambda^{(n)}) > 0 \right\}_{j=1}^{J} \qquad n = 1, \ldots, N \quad j \neq n. \quad (5)$$

In practice (to avoid a trivial solution $P = 0$) the inequalities are solved as larger than a small number (within the machine accuracy), which is equivalent to Eq. (5). Because the interaction between the amino acids is symmetric (e.g., the pair potential of glycine − proline is the same as the pair potential of proline − glycine) the total number of p_λ-s is (20×21/2)×13=2,730. For example, the interaction between cysteine and alanine at a distance between 3.5 and 4.0 Å is −7.353, and at a distance between 4.0 and 4.5 − it is: −7.379 . A few plots with distance-dependent pair potentials are shown in Figure 1. The complete potential table is available in an electronic form from the authors. It is also included as supplementary material (http://www.interscience.wiley.com/jpages/0887-3585/suppmat/index.html).

The number of parameters that we need to determine is significantly larger than the number of parameters that are used in contact potentials because contact potentials use only a single step (2,730 versus 210). We need to show that this large number of parameters can be determined in a unique way and that they indeed provide a better potential.

LEARNING SET

The decoy structures were selected by the combination of two separate procedures. In the first part we used directly structures and sequences from the Protein Data Bank. Five hundred seventy-two structures and sequences were used in gapless threading of all sequences into all structures to create 28,143,009 inequalities.[1] The 572 proteins were selected according to diversity of protein shapes. Nevertheless, the set includes a number of homologous proteins. This is another possible advantage of the LP approach that is less sensitive to the presence of related proteins in the same learning set compared with determination of statistical potentials.

The large number of inequalities can be understood as follows. Let the length of the sequence S_m be l; let the number of structural sites in X_j be k ($k \ge l$). If a short sequence is threaded into a long structure ($k - l + 1$), decoy structures are obtained. We have in our set also short proteins (30–50 amino acids) that contribute a large number of inequalities. The above set of inequalities can be solved with a single step function, as we did in our recent article.[1] The solution with a single step suggests that gapless threading does not pose a significant challenge to the potential design problem.

A more challenging set of decoy structures was generated as follows. We followed up the work presented in our previous manuscript. We have used the program MONSSTER by Skolnick et al.[7] to generate a set of additional 4,299,167 decoy structures for 75 proteins (Table I). This set was created in cycles of potential optimizations and stochastic conformation searches. First, a potential optimized on the set of 572 structures was implemented in the MONSSTER program and used in a Monte Carlo simulation of folding of 75 proteins (Table 1). The resulting set of structures was used to optimize the energy further and to initiate more folding simulations.

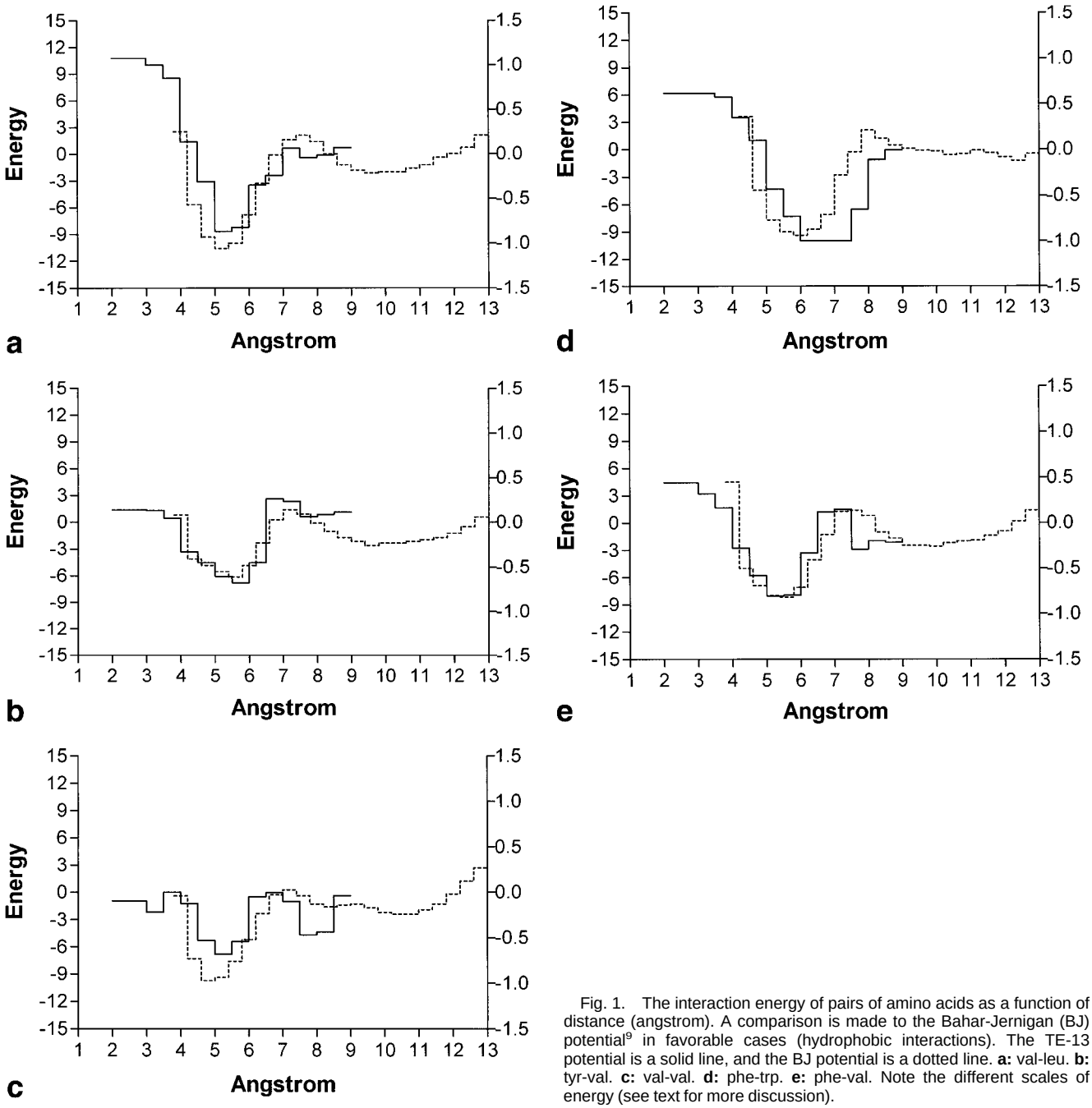

Fig. 1. The interaction energy of pairs of amino acids as a function of distance (angstrom). A comparison is made to the Bahar-Jernigan (BJ) potential[9] in favorable cases (hydrophobic interactions). The TE-13 potential is a solid line, and the BJ potential is a dotted line. **a:** val-leu. **b:** tyr-val. **c:** val-val. **d:** phe-trp. **e:** phe-val. Note the different scales of energy (see text for more discussion).

The last set of decoy structures plus the set that was obtained by gapless threading experiments led to a feasible solution using the 13 step potential (in our previous manuscript the same decoy structures led to an infeasible solution with a potential based on seven steps). It is important to emphasize that with one more cycle (generating a total of 4,939,145 MONSSTER structures) the set of inequalities becomes infeasible even with 13 steps. Hence, our conjecture from the previous publication that it is not possible to find a folding potential based on a sum of pair potentials applies also to the present case. We further comment that the current potential uses a reduced representation of the amino acids (a single point at the side chain geometric center). The reduced representation does not justify the application of yet finer steps. The size of a single amino acid is already a few angstroms, significantly larger than the step size of 0.5 Å.

A cycle before the infeasibility, the optimized potential recognizes a large number of native shapes. We (of course) recognize that it is not the "ultimate" answer to protein folding potentials. Nevertheless, this potential that has a highly flexible functional form and was trained on tens of millions inequalities is likely to be an improvement compared with other contact energies.

TABLE I. Seventy-Five Proteins That Were Used in Iterative Folding Simulations†

1aac	1vih	1rgeA
1abv	1wkt	1rip
1acp	1ycqA	1rro
1acx	2abd	1skz
1ag2	2cbp	1smpl
1ah9	2chsA	1sphA
1ail	2gf1	1tig
1bdo	2gmfA	1tlk
1beo	2hts	1tsg
1cd8	2kauA	1tul
1cei	2mcm	1vcc
1cewl	2msbA	1jpc
1cmcA	2pspA	1kjs
1coo	2tgi	1kptA
1difA	2trxA	1krn
1emn	3sicl	1kum
1exg	4cpv	1lklA
1fbr	4rhn	
1fkb	5icb	
1fow	9rnt	
1frd	1molA	
1frrA	1ngr	
1fvl	1npoA	
1hdj	1nre	
1hfh	1paz	
1hiwA	1pdr	
1hoe	1poa	
1iba	1ppa	
1igmL	1mai	

†Each of the 75 proteins was used in short folding simulations, with the program MONSSTER[7] to collect decoy structures. Based on the decoy structures of the last iteration, a new contact potential was created by solving the inequalities of Eq. (5). The new potential was used in the forthcoming iterations.

COMPUTING THE POTENTIAL PARAMETERS

To solve the set of 28,143,009 + 4,299,167 = 32,442,176 inequalities, we used the program BPMPD.[8] Only a fraction of the total number of inequalities (between 50,000 and 60,000) could be loaded into the program (and to the computer memory) at a time. We select the expressions that were smaller than a cutoff C [Eq. (6)]. The cutoff was tuned so the number of expressions was between 50,000 and 60,000.

$$\left\{ \sum_{\lambda} p_{\lambda}(n_{\lambda}^{(j)} - n_{\lambda}^{(n)}) < C \right\}_{j=1}^{=50,000} \tag{6}$$

This subset of inequalities is solved exactly to produce a new set of parameters $\{p_\lambda\}$. The new parameter set is used to evaluate the correctness of the inequalities. If all the new constraints are satisfied [Eq. (5)], then we found our solution; if not, the process is iterated.

Our energy function, which can change abruptly on a scale of 0.5 Å, is very flexible, perhaps even too flexible. It is reasonable to assume that the "true" energy will be quite smooth and vary gradually over distance. Therefore, we imposed the condition of "maximally smoothed energy" to select a particular solution from parameter sets that solve the inequalities.

Consider the set of parameters $p_\lambda \equiv p_{\alpha\beta}$. For a fixed α (the types of the two interacting amino acids), $p_{\alpha\beta}$ is a function of the distance only (β). A maximally smoothed parameter set is defined as the set that minimizes the double sum below:

$$\sum_{\alpha} F(\alpha) = \sum_{\alpha=1}^{210} \sum_{\beta=1}^{12} (p_{\alpha,\beta} - p_{\alpha,\beta+1})^2 = \text{minimum}. \tag{7}$$

The minimum of F ($p_{\alpha\beta}$) provides us with a potential for the α pair with the minimum "noise." The solution of the inequalities is performed in conjunction with the optimization of a single quadratic function— $\Sigma_\alpha F(\alpha)$. A quadratic function can be optimized efficiently with linear constraints and BPMPD has this option. The solution of Eqs. (5) and (7) yields the potential that we report here and call TE-13. The "zero" of our potential is the state of no contacts. Hence, the energy is set to zero when none of the amino acids interact with each other (they interact only with water molecules).

COMPARISON TO ANOTHER DISTANCE-DEPENDENT, CONTACT POTENTIAL

The functional form of distance-dependent potentials can also be extracted by other techniques. A recent investigation of a distance-dependent potential is of Bahar and Jernigan.[9] A distance-dependent statistical potential for pairs of amino acids was computed. There is no obvious reason why a statistical potential will be the same as a potential constructed with the LP approach. Thomas and Dill[10] showed that statistical potentials do not play the same role as the energies defined in the present manuscript. Betancourt and Thirumalai[11] discussed further the extraction of potentials from statistical data. Nevertheless, parameters for contact potentials that were obtained by different approaches tend to agree on many common features. Vendruscolo and co-workers[3] and Tobi et al.[1] showed this similarity in the past. Therefore, we compared our data with the model of Bahar and Jernigan (BJ).[9]

We first comment on the energy scale. The energy scale of the TE-13 is arbitrary. Hence, if the set of parameters $\{p_\lambda\}$ is a solution of the inequalities, so is $\{\mu p_\lambda\}$ where μ is an arbitrary positive constant. The computational setup, which puts an upper bound on the values of the parameters, determines the scale. Clearly, the energy scale of the TE-13 and the BJ potentials is not necessarily the same. The individual parameters could be stretched up and down (by a single constant number) to find a good match. For the qualitative comparison outlined below, a convenient factor is 10 (TE-13 values are larger by a factor of 10 compared with BJ values). Therefore, in the figures listed below we used different scales for the two potentials. We also comment on the definition of the contacts. BJ did not use the side-chain centers to define a contact but rather terminal atoms. However, for the examples reported here the BJ potential performed better by using the side-chain centers, which is the option we used for all potentials.

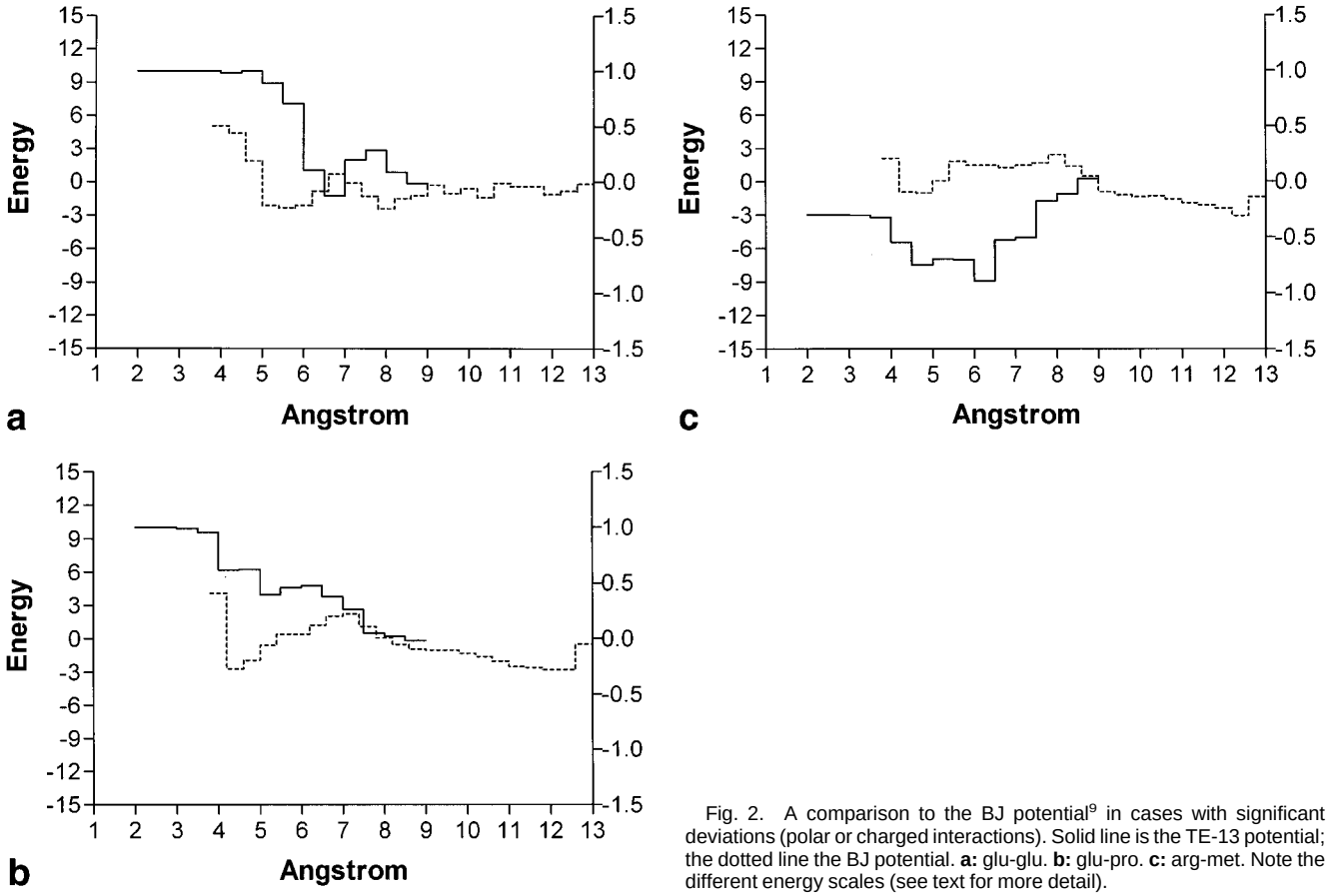

Fig. 2. A comparison to the BJ potential[9] in cases with significant deviations (polar or charged interactions). Solid line is the TE-13 potential; the dotted line the BJ potential. **a:** glu-glu. **b:** glu-pro. **c:** arg-met. Note the different energy scales (see text for more detail).

In Figure 1a we show the pair potential for the interaction of leu and val. The similarity of the two functional forms is surprisingly high. Other surprisingly similar pair potentials are of tyr and val (Fig. 1b). In general, hydrophobic pair interactions seem to share many common features. For example, the val-val, phe-trp, and phe-val potentials (Fig. 1c–e) share similar shapes.

However, although these similarities are common, they are not the rule. In Figure 2 we summarize a number of cases with significant deviations. Our glu-glu potential (Fig. 2a) is repulsive (positive) at distances shorter than six angstroms. The BJ potential is negative until distances of approximately 3 Å. glu-pro is uniformly positive in TE-13 but includes a significant well at 4 Å in the BJ potential (Fig. 2b). As a last example, consider arg-met potentials (Fig. 2c). The TE-13 is considerably more attractive than BJ.

The contact potential (only one step instead of 13) that we computed with the LP approach[1]] was also quite different from other statistical potentials (although similar in performance). However, when performing eigenvalue analysis, we found that the dominant contribution to the energy came from a "hydrophobicity" eigenvector with amplitudes similar for a number of potentials.[1] We were unable to perform a similar analysis for the present case. It is not clear what matrix we should consider. An attempt to define the matrix of the interactions averaged over all distances provided senseless results.

INDEPENDENT SET OF DECOY STRUCTURES

TE-13 was "trained" on 572 protein structures from the Protein data Bank (PDB), and an additional set of decoys was constructed by a Monte Carlo program[7] The large sample of PDB structures makes it difficult to generate another nonhomologous set of decoy shapes to have an independent test of TE-13. Therefore, it is useful that the Levitt's group deposited a set of test structures on the web. We have used a subset of all the tests to examine the prediction capacity of the newly generated, distance-dependent potential. We consider only the multiple decoy sets. We did not include proteins that were part of our folding computations of 75 proteins. It is important to emphasize that from the proteins that were used to generate the set of decoys only the native structures of 1ctf and 1igd were included in the learning set of 572 proteins (but of course not the decoy structures!).

In Table II we provide the results for the TE-13 together with the results of five other contact potentials. The results are divided between the different classes of the tests as described on the web site.[12] In Table II we report the number of decoys included in the test, the position of the true native structure with respect to the decoy structures, and the Z scores for the individual tests. Note that we have not shown

TABLE II. Scoring of Different Potential Using a Subset of the Levitt's Decoys[12†]

1.

4state_reduced	Ref. 12a						
Protein	No. of decoys	TE-13	MJ	GKS	BT	HL	BJ
1ctf	631	1/4.20	1/3.73	1/3.01	1/3.86	1/3.47	1/4.12
1r69	676	1/4.06	1/4.11	1/3.83	1/4.47	1/4.26	1/5.33
1sn3	661	6/2.70	2/3.17	6/2.64	6/2.97	4/2.64	151/0.73
2cro	675	1/3.48	1/4.29	1/4.06	1/3.92	1/3.65	1/4.60
4pti	688	7/2.43	3/3.16	1/3.87	5/2.65	2/2.96	300/0.18
4rxn	678	16/1.97	1/3.09	2/2.86	1/3.01	5/2.53	1/3.11

2.

fisa	Ref. 12b						
Protein	No. of decoys	TE-13	MJ	GKS	BT	HL	BJ
1fc2-C	501	16/1.67	282/−0.11	404/−0.85	314/−0.28	246/0.07	75/0.95
1hddc-C	501	1/4.35	1/3.03	2/3.23	2/3.27	2/3.29	1/3.81
2cro	501	1/4.00	2/3.47	3/2.90	5/2.82	3/2.84	1/4.54

3.

fisa_casp3	Ref. 12b						
Protein	No. of decoys	TE-13	MJ	GKS	BT	HL	BJ
1bg8-A	1201	3/2.98	13/2.27	9/2.47	3/2.81	34/1.94	1/2.98
1bl0	972	3/2.80	532/−0.42	561/−0.49	437/−0.13	511/−0.31	1/3.05
1jwe	1408	1/6.04	1/3.53	1/2.72	5/2.46	2/2.91	12/2.14

4.

lattice_ssfit	Ref. 12c						
Protein	No. of decoys	TE-13	MJ	GKS	BT	HL	BJ
1ctf	2001	1/6.17	1/5.35	1/4.90	1/6.99	1/5.82	1/5.54
1dkt-A	2001	2/3.92	32/2.41	4/3.55	5/3.49	3/3.48	10/2.24
1fca	2001	36/2.25	5/3.40	2/3.53	2/3.92	19/2.37	1/3.60
1nkl	2001	1/4.51	1/5.09	1/6.32	1/7.28	2/4.26	16/2.16
1pgb	2001	1/4.13	3/3.78	27/2.22	2/3.82	3/4.35	1/4.53
1trl-A	2001	1/3.63	4/2.91	19/2.38	17/2.59	6/3.24	27/1.88

5.

1mds	Ref. 12d						
Protein	No. of decoys	TE-13	MJ	GKS	BT	HL	BJ
1ctf	498	1/4.13	1/3.86	1/3.72	1/3.15	1/3.10	1/3.57
1dtk	216	5/1.88	13/1.71	50/0.70	122/−0.08	2/2.66	214/−3.42
1fc2-C	501	14/2.04	501/−6.24	501/−7.55	501/−5.11	501/−5.11	319/−.031
1igd	501	2/3.11	1/3.25	2/2.56	1/3.76	1/3.57	1/3.45
1shf-A	438	1/4.13	11/2.01	84/0.89	16/1.82	18/1.76	1/4.12
2cro	501	1/3.96	1/5.07	1/3.93	1/4.01	1/4.21	1/5.75
2ovo	348	1/3.62	2/3.25	15/1.70	31/1.29	1/2.67	150/0.14

[†]Only the multiple decoy sets are considered. Structures of proteins that were included in the Monte Carlo training set (75 proteins) are removed. The tests are separated into families of decoys and different proteins. For each protein we report (a) the name and (b) the number of decoy structures that were considered in a test family. We also report two scores for each potential: the location of the native structure in the energy scale ("1" is the best) and the Z score. The potentials considered are of Tobi and Elber (the present manuscript) (TE-13), of Miyazawa and Jernigan (MJ),[14] of Godzik, Kolinski, and Skolnick (GKS),[15] of Betancourt and Thirumalai,[11] of Hinds and Levitt,[13] and of Bahar and Jernigan.[9]

6. Average Position and Z Score (Over All Proteins) of the Native Structure for All Potentials

—	TE-13	MJ	GKS	BT	HL	BJ
⟨native⟩ ⟨Zscore⟩	4.96/3.53	56.64/2.82	68/2.36	59/2.65	54.8/2.67	51.6/2.75

the tests for lmds/1b0n-B (which is an open structure) and lmds/1bba (which is a protein with an unusual core of proline residues) on which all potentials failed.

To test the potential quality we report the relative position of the true native and also computed the Z score. It is defined as follows $Z = \langle E \rangle - E_n / \sqrt{\langle E^2 \rangle - \langle E \rangle^2}$ where the

average is over all decoy structures and E_n is the energy of the native state. The arithmetic average of the Z scores of all the tests of individual proteins provides a single number that is convenient as a measure of prediction capacity. For the TE-13 the average Z score is 3.53, significantly higher than the Z scores of other contact potentials, which are below 2.9.

Perhaps more striking is the consistent behavior of the TE-13 throughout a large set of tests. 1fc2-c and 1bl0 present significant challenges to the other contact potentials (all have *negative* Z scores!) but score reasonably high (16 and third) by using the TE-13 potential. It is encouraging that the TE-13 is showing a more uniform and consistent behavior, even when it is doing worse than other potentials. For example, the native shape was placed 16 for 4rxn, whereas the other potentials have better scoring. Not only is the average Z score better but also the average position of the native state (TE-13 places the average native at the fifth location, whereas the other potentials locate the average native around position 50). The number of proteins solved exactly by the HL potential[13] is eight, MJ[14] solved 11, GKS[15] and BT[11] solved nine proteins. The TE-13 potential solved 14 proteins.

CONCLUSIONS AND FINAL REMARKS

We have shown that the linear programming protocol, in conjunction with a large number of decoy structures can produce a new set of parameters with improved prediction capacity. The hydrophobic components of our potential are remarkably similar to the statistical potential derived by Bahar and Jernigan. Significant differences in polar interactions lead to a better prediction (on the average) of the native shape and better Z scores. We also showed (another useful features of the LP approach) that the distance-dependent contact potential is still not good enough. It is possible to create protein-like shapes that lead to infeasibility. It is impossible to find a set of parameters such that the native structures will be selected from a large set of decoys.

Of course, in principle it is possible to decrease further the step size of the contact potential in an attempt to find a solution. Such an increase may increase the number of inequalities that we are able to solve. However, on the basis of physical consideration (the size of an amino acid is a few angstrom), it is unreasonable to seek a reduced potential with a distance resolution that is <0.5 Å. Even when moving from 7 to 13 steps, the number of parameters increases significantly with only minute improvements in prediction ability. A possible promising direction in further refinement of protein folding potentials is the inclusion of three and higher body terms.

The complete set of structures is available from www.tc.cornell.edu/CB10/loopp as a part of the loopp program.

ACKNOWLEDGMENTS

This research was supported by NIH NCRR grant to the Cornell Theory Center, and by DARPA seed money to R.E.

REFERENCES

1. Tobi D, Shafran G, Linial N, Elber R. On the design and analysis of protein folding potentials. Proteins Struct Funct Genet, in press.
2. Maiorov VN, Crippen GM. Contact potential that recognizes the correct folding of globular proteins. J Mol Biol 1992;227:876–888.
3. Vendruscolo M, Najmanovich R, Domany E. Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading? Proteins Struct Funct Genet 2000;38:134–148.
4. Socci ND, Onuchic JN, Wolynes PG. Protein folding mechanisms and the multidimensional folding funnel. Proteins Struct Funct Genet 1998;32:136–158.
5. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–170.
6. Klimov DK, Thirumalai D. Criterion that determines the foldability of proteins. Phys Rev Lett 1996;76:4070–4073.
7. Skolnick J, Kolinski A, Ortiz AR. MONSSTER: a method for folding globular proteins with a small number of distance constraints. J Mol Biol 1997;265:217–241.
8. Meszaros CS. Fast Cholesky factorization for interior point methods of linear programming. Comput Math Appl 1996;31:49–51.
9. Bahar I, Jernigan RL. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. JMB 1997;266:195–214.
10. Thomas PD, Dill KA. Statistical potentials extracted from protein structures: how accurate are they? J Mol Biol 1996;257:457–469.
11. Betancourt MR, Thirumalai D. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. Protein Sci 1999;2:361–369.
12. References for Levitt tests. (a) Park B, Levitt M. Energy functions that discrimination x-ray and near-native folds from well-constructed decoys. J Mol Biol 1996;258:367–392. (b) Simons KT, Kooperberg C, Huang ES, Baker D, Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. J Mol Biol 1997;268:209–225.
13. Hinds DA, Levitt M. A lattice model for protein structure prediction at low resolution. Proc Natl Acad Sci USA 1992;89:2536–2540.
14. Miyazawa S, Jernigan RL. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term for simulation and threading. J Mol Biol 1996;256:623–644.
15. Godzik A, Kolinski A, Skolnick J. Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. Protein Sci 1995;4:2107–2117.