PROTEINS: Structure, Function, and Genetics 40:71–85 (2000)

On the Design and Analysis of Protein Folding Potentials

Dror Tobi,[1] Gil Shafran,[2] Nathan Linial,[2] and Ron Elber[1,3,4]*
[1]*Department of Biological Chemistry, The Hebrew University, Givat Ram, Jerusalem, Israel*
[2]*Department of Computer Science, The Hebrew University, Givat Ram, Jerusalem, Israel*
[3]*Department of Computer Science, Cornell University, Ithaca, New York*
[4]*Department of Physical Chemistry, The Hebrew University, Givat Ram, Jerusalem, Israel*

ABSTRACT **Pairwise interaction models to recognize native folds are designed and analyzed. Different sets of parameters are considered but the focus was on 20 × 20 contact matrices. Simultaneous solution of inequalities and minimization of the variance of the energy find matrices that recognize exactly the native folds of 572 sequences and structures from the protein data bank (PDB). The set includes many homologous pairs, which present a difficult recognition problem. Significant recognition ability is recovered with a small number of parameters (e.g., the H/P model). However, full recognition requires a complete set of amino acids. In addition to structures from the PDB, a folding program (MONSSTER) was used to generate decoy structures for 75 proteins. It is impossible to recognize all the native structures of the extended set by contact potentials. We therefore searched for a new functional form. An energy function *U*, which is based on a sum of general pairwise interactions limited to a resolution of 1 angstrom, is considered. This set was infeasible too. We therefore conjecture that it is not possible to find a folding potential, resolved to 1 angstrom, which is a sum of pair interactions. Proteins 2000;40:71–85.**

Key words: contact matrices; linear programming; optimization; decoy structures

INTRODUCTION

It is obvious that a good scoring function or a suitable potential is an essential ingredient of an ab initio attempt to fold proteins. One approach (which we do not employ here) is to find a potential energy function using physical chemistry principles, trying to mimic the way proteins fold in nature. Another approach, more limited in scope, is to find an energy function that will set the native conformation to be the lowest in energy. In the last scheme, we do not care how misfolded conformations are related to each other as long as they are higher than the native fold. The potential energy is expected to be "correct" in only one point (the native coordinate set) when compared to the rest of the states.

By limiting the task, and examining only energy differences between folded states and misfolded conformations (called "decoys"), the design problem of the potential might be easier. Considering only known (experimental) structures can further reduce the problem size to the set of experimentally solved structures (at most couple of thousands), and is called "threading." By focusing on the recognition problem we hope to find a potential that will do this limited task better than other potentials that were designed with additional features in mind.

We therefore consider in the present manuscript two separate tasks: (a) generating potentials to recognize structures from the protein data bank (PDB), and (b) designing a potential that differentiates between protein-like structures generated by a computer program[1] and true native shapes. To address task (b) decoys are computationally prepared for 75 proteins.

At present, there are a number of alternate "scoring functions" that are used to recognize native folds.[2–18] A detailed comparison between folding potentials, which were extracted using alternative approaches, is of considerable interest. Is the computed potential unique? In what ways are the potentials different and what information is considered essential to make a reasonable folding potential? Specific characteristics that repeat in different derivations are more likely to be an inherent feature of the true potential. As we will show, such characteristics indeed exist in the differently derived scoring functions. Nevertheless, significant room for improvement remains.

We also discuss the limits of the contact model. Can we identify a bound to the capacity of the model? That is, can we generate a set of decoys for which no parameters are able to recognize the native fold? Domany and co-workers studied this question in the past[13] and concluded that a square well contact potential is insufficient to recognize native folds. Based on the following numerical experiments, we conjecture that a folding potential, which is based on arbitrary pairwise function with 1 angstrom resolution, does not exist.

In the next section, we describe the computational protocol that was used to generate the new potentials. In the Analysis of Potentials section, we study the potentials' properties. Of special interest are the similarities between different contact matrices, which were derived using differ-

Grant sponsors: Israel Science Foundation; NIH Resource at the Cornell Theory Center; and FIRST award from the Israel Science Foundation.

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

Received 29 March 1999; Accepted 6 January 2000

ent techniques. Conclusions and future directions are provided in the last section.

POTENTIAL DESIGN: PROTOCOL AND OBSERVATIONS

Computational Approach

We state the limited design problem as follows: Let X be the coordinate vector that represents the protein. A single interaction site is used to represent an amino acid. The single point can be the position of the C_α, or the geometric center of the side chain. We further denote the native fold by $\hat{X}_n$, and a set of decoy structures by $\{X_i\}_{i=1}^N$. The statement that the energy of the native fold is lower that the energy of any of the N decoy structures, is written as:

$$\{U(X_i) - U(\hat{X}_n) > \varepsilon_i\}_{i=1}^N \tag{1}$$

The potential energy is $U(X)$, and ε_i is a nonnegative constant to be provided. In the calculations below, it is set to zero or to a small positive number. Clearly any other choice of ε_i will weaken the condition that the native state must be the lowest in energy. This is in conjunction with an additional inequality that the average potential is larger than a critical value (see below). We seek a parameterization of the energy function such that the inequalities can be solved (easily), preferably for a large number of decoy structures. Computationally, it is convenient to consider a potential that depends linearly on its parameters.

$$U(X, P) = \sum_{l=1}^{L} p_l S_l(X) \tag{2}$$

The vector of parameters is denoted by P with elements $\{p_l\}_{l=1}^L \cdot \{S_l(X)\}_{l=1}^L$ are basis set functions that are used to expand the potential and L is the number of basis functions used. Equation 1 is written as:

$$\{\sum_{l=1}^{L} p_l[S_l(X_i) - S_l(\hat{X}_n)] > \varepsilon_i\}_{i=1}^N \tag{3}$$

The unknowns, which we want to determine, are the parameters $\{p_l\}_{l=1}^L$. The inequalities in Eq. 3 are linear in the $\{p_l\}_{l=1}^L$, and are therefore accessible to efficient computations. The functions $S_l(X)$ are not specified yet. An arbitrary potential can be expanded by these functions if the $\{S_l(X)\}$ form a complete basis set. However, we need to choose a specific functional form so that a finite number of terms will provide a useful approximation. We employ a model consistent with previous derivations of folding potentials[2,5–7,9,10] that sets $S_l(X)$ to a constant over a given range.

$$S_l(X) = S_{nm}(|x_n - x_m|) = S_{nm}(r_{nm}) \equiv S_{nm}$$
$$S_{nm} = \begin{Bmatrix} 1 & r_{cut\_low} < r_{nm} < r_{cut\_up} \\ 0 & otherwise \end{Bmatrix} \tag{4}$$

The nm refers to amino acid indices and replaces the single expansion index l. x_j is the vector representing the position of amino acid j, and r_{nm} is the distance between the two amino acids. Equation 4 suggests a very simple picture for the folding potential. If the two amino acids n and m are sufficiently near (but not too close), then a contact is "on," otherwise it is "off." The potential is given by a sum of the contacts that are " on." Such a contact potential is a widely used approximation in studies of folding. We employ the same functional form to make it easier to compare to other people's results. Of course, following Eq. 3, any functional form can be used as long as it is linear in its parameters. We also explored a considerably more elaborate potential once it becomes clear that the contact potential is insufficient.

We are not the first investigators to use inequalities to design folding potentials. The first pioneering work was of Maiorov and Crippen; they provided the foundation to follow up works (including ours).[4] Our study differs from the original work by using an energy function that is more widely employed today. We also considered a larger set of decoy structures. The larger set requires a few more twists to the optimization protocol that are discussed below. Finally, the extensive comparison to other potentials was not possible at that time, and the study of the limits of a general pair potential is also new.

Other related studies are of Domany and co-workers.[10] The functional form of the potentials was similar to the works of other investigators and our studies. However, significant differences remain. The present manuscript is using a different algorithm to solve the inequalities, and it also considers the optimization of a width function in conjunction with the solution of the inequalities. A detailed analysis and comparison between alternative parameter sets is also provided, attempting to capture and to understand the common features of different parameterizations.

Another approximation in the current study, which (again) is widely employed, is the application of contact types. According to this model, contacts of the same types of amino acids provide the same value of energy regardless of their location on the peptide chain. Hence, three (and higher) body effects are ignored. If we index contact types by a, we can re-write Eq. 3 as:

$$\{\sum_{\alpha=1}^{210} p_\alpha(n_\alpha^i - n_\alpha^n) > \varepsilon_i\}_{i=1}^N \tag{5}$$

where p_α is the energy associated with a contact of type α, and n_α^i is the number of contacts of type α in structure i. There are 210 types of contacts for 20 amino acids. A solution of Eq. 5 provides a set of 210 parameters. The parameters determine the potential for a model of $S_l(X)$, which as we argued before, is widely used.[5–7,9,10] It is also possible to reduce the number of parameters, by dividing the amino acids into groups and assigning a single parameter to each group. The extreme model of this type is the HP parameterization for which we demonstrate surprisingly high, but nevertheless low capacity.

Does a Sum of General Pairwise Interactions Recognize Native Folds?

The "true" pair interaction may have considerably more complex functional form than the widely used square-well potential. We therefore construct a flexible form of the potential that can adjust into numerous possible forms. For any of the contact types, we divide the distance between the two amino acids into seven segments. Seven independent parameters, $\{p_{\alpha q}\}_{q=1}^{7}$, describe the energy of each of the individual segments. Moreover, to test that the choice of the segments does not have a significant influence on our results we consider separately three choices of the distance partitioning:

	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Step 7
Potential 1	2–3	3–3.5	3.5–4	4–4.5	4.5–5	5–7	7–9
Potential 2	2–3	3–4	4–5	5–6	6–7	7–8	8–9
Potential 3	2–3	3–3.5	3.5–4	4–5	5–5.5	5.5–7	8.9

Physically it is hard to justify higher resolution when the basic model of an amino acid is a single interaction site. The new formulation yields another set of inequalities that are similar to Eq. 5:

$$\{\sum_{\alpha=1}^{210}\sum_{k=1}^{7} p_{\alpha k}(n_{\alpha k}^{i} - n_{\alpha k}^{n}) > \varepsilon_{ik}\}_{i=1}^{N} \quad (6)$$

Equation 6 has 1,470 linear parameters to optimize, but beside the obvious increase in the number of unknowns is essentially the same as Eq. 5. The height of each of the steps is optimized independently, providing a new potential with a resolution of about 1 angstrom.

This highly detailed set of steps and potentials may seem unnecessary, because many useful contact potentials are available that are using less parameters. However, even the detailed potentials are insufficient for the set of structures at hand (shown later) when we examine structures that are not in the PDB (but in many respects are protein-like). Although the above forms are not the most general possible for pair interactions, they are likely to provide a sound approximation to a general pairwise potential with a resolution of about 1 angstrom. Hence, we conjecture that ab initio attempts to fold proteins based on pairwise interactions are bound to fail. Nevertheless, we emphasize that for threading through carefully selected PDB structures the common approaches that employ 20×20 matrices may be sufficient.

Derivation of Potential Parameters Using Linear Programming

Each of the linear inequalities in Eq. 5 divides the parameter space into two, a part that is allowed, and a part that is not. An inequality can have different effects on the solution. It may: (1) restrict further the space allowed for the parameter set (most desirable result), (2) have no effect on the space already restricted (e.g., adding an inequality $X > 5$ after having already $X > 4$), and (3) impose an impossible condition (such as $X - 5 > Y$ and $Y > X$) making the inequalities unsolvable.

TABLE I. Small Learning Set†

1aak	1llc	256b	3cd4	7tim
1abe	1lld	2act	3chy	8adh
1abm	1lpe	2alp	3cla	8atc
1ace	1lts	2aza	3dfr	8atc
1acx	1lts	2bb2	3est	8dfr
1ak3	1lz1	2ca2	3gap	8fab
1ake	1mba	2ccy	3gly	8gch
1ala	1mcp	2cna	3hla	9rnt
1alc	1mpp	2cts	3icd	9wga
1ald	1nn2	2cyp	3lad	
1bab	1npx	2dnj	3pfk	
1bab	1nsb	2end	3pgk	
1bbh	1ofv	2er7	3pgm	
1bbp	1ova	2fcr	3pmg	
1bbt	1paz	2fox	3rp2	
1bbt	1pbx	2fx2	3rub	
1bbt	1pbx	2gbp	3rub	
1bia	1phd	2gmf	3sdp	
1bmv	1phh	2hbg	3sic	
1bmv	1pii	2lhb	3sic	
1c2r	1pp2	2liv	4bp2	
1ccr	1ppa	2ltn	4cpv	
1cd8	1ppf	2mcg	4dfr	
1cmc	1ppl	2mcm	4enl	
1col	1ppn	2mev	4fab	
1cox	1prc	2mev	4fgf	
1dri	1prc	2mev	4gcr	
1dwc	1prc	2msb	4gpd	
1eca	1prc	2pab	4lzm	
1etu	1pyp	2pf1	4mdh	
1ezm	1r09	2pfk	4pep	
1f3g	1r09	2pgd	4rcr	
1fc2	1r09	2pka	4rcr	
1fcb	1rbp	2plv	4rcr	
1fha	1rcb	2plv	4sbv	
1fkb	1rhd	2por	4sdh	
1fnb	1rnh	2prk	4tgl	
1gal	1rro	2reb	4tms	
1gd1	1rve	2ren	5cpa	
1gky	1sdy	2rsp	5cyt	
1gox	1sgt	2scp	5fbp	
1gp1	1spa	2sga	5ptp	
1gpb	1stp	2sns	5rub	
1gpr	1thm	2snv	5tim	
1grc	1tie	2sod	5tmn	
1hge	1tim	2stv	5tnc	
1hge	1tlk	2tbv	6gst	
1hrh	1tnf	2tmv	6ldh	
1ifc	1ton	2tpr	6q21	
1igm	1trb	2trx	6taa	
1ipd	1ula	2wrp	6xia	
1ith	1vsg	3adk	7acn	
1lap	1wsy	3apr	7api	
1ldn	1wsy	3blm	7cat	
1lh2	1yea	3c2c	7nn9	

†A subset of the Hinds and Levitt database that includes only 229 proteins of the 246 proteins originally included in the database reported in Bryant and Lawrence.[6] We Provide a List of the PDB Identifiers.

After collecting the effects of many inequalities, an allowed "volume" in parameter space is defined. A point in the allowed volume corresponds to a possible solution of

TABLE II. Large Learning Set†

1aac	1bdo	1ddt	1gai	1knb	1nre	1rcb	1u9aA	2gf1	2tysB	5timA
1aak	1beo	1deaA	1gal	1kptA	1nsbA	1regX	1ula	2gmfA	2vik	5tmnE
1ab8A	1berA	1dec	1garA	1krn	1nulA	1rgeA	1vaoA	2gstA	3adk	5tnc
1ab9C	1bgk	1def	1gca	1kuh	1nzyA	1rgp	1vcc	2hbg	3aprE	6ldh
1abe	1bglA	1difA	1gcb	1kum	1oacA	1rhgA	1vhh	2hts	3blm	6q21A
1abmA	1bgw	1dik	1gd1O	1kveB	1octC	1rie	1vhrA	2kauA	3btoA	6rxn
1abv	1bia	1div	1gdoA	1lay	1ofv	1rip	1vid	2kauB	3c2c	6taa
1acp	1ble	1dkzA	1gds	1lba	1ois	1rlaA	1vih	2kauC	3chy	6xia
1acx	1bme	1dnpA	1geo	1lbd	1orc	1rlr	1vjs	2lhb	3cla	7aatA
1ad2	1bmfA	1dsbA	1gky	1lbeA	1ordA	1rpa	1vnc	2liv	3dfr	7apiA
1ad3A	1bmtA	1dvh	1gln	1lbu	1ospO	1rro	1wba	2ltnA	3ecaA	7nn9
1aep	1bmv1	1dwcH	1gof	1lci	1otfA	1rsy	1whi	2masA	3est	7rsa
1af5	1bmv2	1dxgA	1gox	1lcpA	1ovaA	1rtm1	1wkt	2mbr	3gapA	7timA
1afp	1bnb	1eaf	1gpb	1ldnA	1pauA	1rveA	1wsyA	2mcg1	3gly	8atcA
1afwA	1bncA	1ebdC	1gpc	1lfaA	1pax	1rvvA	1xaa	2mcm	3grs	8atcB
1ag2	1bndA	1eca	1gpmA	1lgr	1paz	1ryt	1xnb	2mev1	3hlaA	8catA
1ah6	1bor	1ecl	1gpr	1lh2	1pbwA	1sap	1ycqA	2mev2	3icd	8dfr
1ah9	1brsD	1ecpA	1grj	1lis	1pbxA	1sdyA	1yea	2mev3	3ladA	8fabB
1aihA	1btl	1ecrA	1gsa	1lit	1pbxB	1seiA	1yge	2msbA	3minB	8gch
1ail	1bucA	1eczA	1gtqA	1lklA	1pdc	1sfe	1yrnA	2nIIB	3pfk	8icoA
1air	1bvp1	1efnB	1gtrA	1lla	1pdo	1sgt	1ytbA	2ora	3pgk	9rnt
1ajqA	1bw4	1efuA	1gzi	1llc	1pdr	1shcA	1yua	2pabA	3pgm	9wgaA
1ak3A	1c2rA	1efuB	1han	1lldA	1pfkA	1sig	1znbA	2pf1	3pmgA	
1akeA	1ccr	1emn	1hbq	1llp	1pgs	1sis	256bA	2pfkA	3pte	
1ako	1cd8	1enh	1hcb	1lrv	1pgtA	1skz	2abd	2pgd	3rp2A	
1akz	1cdg	1eriA	1hcl	1lst	1phc	1sluA	2abk	2phlA	3rubL	
1ala	1cei	1etu	1hdj	1ltsA	1phh	1smnA	2ace	2phy	3rubS	
1alc	1cewl	1exg	1hfh	1lxa	1phr	1smpl	2act	2pkaB	3sdhA	
1ald	1cfe	1extA	1hiwA	1lz1	1pii	1spa	2alp	2plh	3sdpA	
1alkA	1cfpA	1ezm	1hleA	1mai	1pkm	1sphA	2azaA	2plv2	3sicl	
1alo	1chd	1f3g	1hoe	1maz	1pkp	1sriA	2bb2	2plv3	3tgf	
1amj	1chkA	1fbr	1hpm	1mba	1plq	1sryA	2bbkL	2polA	4aahA	
1amm	1chmA	1fc2D	1hpt	1mcpH	1pmc	1stu	2bct	2prd	4bp2	
1anv	1ckmA	1fcbA	1hrdA	1mhlC	1pne	1svpA	2bds	2prk	4cpal	
1aol	1cmcA	1fcdA	1hrhA	1mioA	1poa	1tadA	2bgu	2pspA	4cpv	
1aonA	1cnsA	1fha	1httA	1mkaA	1poc	1tbd	2bnh	2reb	4enl	
1aonO	1cof	1fid	1hxn	1mla	1pp2R	1tcp	2bpa1	2ren	4fabL	
1aorA	1coo	1finB	1iba	1mldA	1ppa	1tfr	2ca2	2rn2	4fgf	
1apyA	1cox	1fkb	1ifc	1mml	1ppfE	1thjA	2cba	2rslA	4gpd1	
1apyB	1cpo	1flel	1igd	1mngA	1pplE	1thm	2cbh	2rspA	4kbpA	
1ar5A	1cpt	1fnb	1igmL	1molA	1ppn	1thw	2cbp	2scpA	4lzm	
1arb	1crkA	1fow	1iphA	1mpp	1ptq	1tie	2ccyA	2secl	4mdhA	
1auuA	1cseE	1fps	1irk	1mrj	1pvuA	1tif	2chsA	2sga	4mt2	
1avpA	1csel	1frd	1isuA	1msk	1pyaB	1tig	2cpl	2sil	4pep	
1axh	1csh	1fre	1itg	1mtyG	1pyc	1timA	2cyp	2sns	4rhn	
1axn	1csmA	1froA	1ithA	1mxb	1pydA	1tlk	2dnjA	2sodO	4sbvA	
1ayl	1ctf	1frrA	1jbc	1nbaA	1pyp	1tml	2dri	2stv	4sgbl	
1babA	1ctj	1fruA	1jetA	1nbtA	1pysA	1tnfA	2dtr	2tbvA	4tgl	
1babB	1ctt	1frvA	1jpc	1nfn	1pytA	1ton	2end	2tct	4tms	
1bbhA	1cuk	1frvB	1jswA	1ngr	1qasA	1trb	2eng	2tgi	5cpa	
1bbpA	1cxsA	1fua	1jud	1nn2	1qorA	1trkA	2er7E	2tmdA	5cytR	
1bbt2	1cyo	1fuiA	1kapP	1nox	1r091	1tsg	2erl	2tmvP	5fbpA	
1bbt3	1daaA	1fvl	1kfd	1noyA	1r092	1tul	2fcr	2tprA	5icb	
1bco	1dar	1fxd	1kjs	1npoA	1r093	1tum	2fox	2trxA	5ptp	
1bcpA	1dcoA	1gadO	1klo	1npx	1ra9	1tys	2fx2	2ts1	5rubA	

†The set of 572 proteins that was used extensively in the present work. The potential we derived using linear programming techniques solves this set exactly. None of the other sets reported in this work solves the complete set.

the inequalities in Eq. 5. It is therefore important to appreciate that the solution so obtained is not unique. Is it possible to obtain a more "focused" solution? One way of narrowing further the available space is by maximizing (or minimizing) a function in conjunction with the solution of the inequalities. Akutsu and Tashimo[11] considered a protocol similar to the maximization of $\Sigma_{i=1}^{N} \varepsilon_i$ where the ε_i's are treated as independent nonpositive parameters.

TABLE III. The Set of 75 Proteins That Were Used to Generate Decoy Structures with the MONSSTER Program†

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

†An optimized seven step potential was used as the contact potential during the simulation to generate decoys. Self-consistency cycles (updating the contact potential) were made three times. In each cycle 100 folding trajectories were computed for each of the proteins. Structures were saved 200 times during the folding trajectory and were filtered to avoid structures with bad contacts. The total number of decoys generated was 4,299,167. The last set was not feasible. An analysis of properties of multi-step potentials will be published elsewhere.

This protocol does not decrease the parameter space but intend to accommodate infeasible solutions. If the ε_i's can be negative the inequalities (Eq. 5) can be satisfied even if the energy of the native structure is not the lowest. To determine our favorite potential (on a feasible set of inequalities), we use a different protocol.

Consider a solution to Eq. 5, $\{p_\alpha\}_{\alpha=1}^{210}$ in which the ε_i's are set to 0. In addition to the previous solution, there are an infinite number of solutions that are trivially associated with it, $\{\lambda p_\alpha\}_{\alpha=1}^{210}$ where λ is an arbitrary positive constant. Furthermore, for a given set of ε_i, $\{\varepsilon_i\}_{i=1}^{N}$ and a solution $\{p_\alpha\}_{\alpha=1}^{210}$ there is another trivial solution, $\{\lambda p_\alpha\}_{\alpha=1}^{210}$, for a new (trivial) set of ε_i, $\{\lambda\varepsilon_i\}_{i=1}^{N}$. Energy scaling does not affect the ordering, and the relative spacing between the energies of the different configurations.

To address the problem of an "open"-unbound solution we closed the parameters by a box between -10 and $+10$. This leaves us (of course) with still an infinite number of solutions enclosed in a box and related by scaling. To further focus the solution (if desired) we need to optimize a function in conjunction with the solution of the inequalities.

We consider two choices of function to optimize. The first is a constant (zero). In that case, the interior point algorithm that we used stops at the center of the feasible volume. This choice of (no) function to optimize makes the present approach formally similar to the studies of Maiorov and Crippen[4] and Domany et al.[10] and produces sound results.

Nevertheless, an appropriate choice of a target can be helpful in getting a "better" energy function. The inequalities alone define a volume in parameter space. This volume can be made smaller if further physical considerations are taken into account, or even set to a single point of a unique minimum. We therefore seek constraints that will pick an "optimal" point and at the same time could be implemented in a computationally convenient form.

Consider the energy gap function Δ_i

$$\Delta_i = \sum_\alpha (n_\alpha^i - n_\alpha^n)p_\alpha \tag{7}$$

We comment that the optimization of the average gap, $\langle\Delta_i\rangle = 1/N\Sigma_{i=1}^{N}\Sigma_\alpha(n_\alpha^i - n_\alpha^n)p_\alpha$ does not produce gap distributions significantly different (up to a trivial scaling) from what we already obtained without the optimization. We therefore employed only the target function to be discussed below to optimize the potential.

The optimized function is added in two steps. We first add one more inequality to the set: $\Sigma_{i=1}^{N}\Delta_i > \Gamma$ where Γ is a constant. By adding a single constraint on the average energy gap Γ we eliminated the possibility of zero as a solution, but not much more, because all other solutions are related by scaling. What gives the present algorithm a different flavor is the minimization of the variance in conjunction with the solution of the inequalities. That is, we seek a solution for

$$\frac{1}{N}\sum_{i=1}^{N}\Delta_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N}\Delta_i\right)^2 = \text{minimum} \tag{8}$$

TABLE IVA. Potential Designed on a Small Set[†]

	ala	arg	asn	asp	cys	gln	glu	gly	his	ile	leu	lys	met	phe	pro	ser	thr	trp	tyr	val
ala	−0.17	0.03	0.33	0.58	−0.48	−0.14	0.29	0.43	0.66	−0.79	−0.13	0.18	−0.82	−0.49	0.87	0.18	−0.45	−0.04	−0.08	−0.54
arg	0.03	1.01	0.82	−0.40	0.80	0.50	−0.91	0.12	−0.62	0.31	0.28	0.56	0.40	−0.89	−0.59	0.84	−0.19	0.23	−1.64	0.41
asn	0.33	0.82	−0.68	0.28	−1.89	0.87	0.78	0.46	0.95	0.48	1.17	0.00	1.92	0.36	0.08	−0.78	0.13	1.45	0.43	0.71
asp	0.58	−0.40	0.28	−0.24	−0.32	−0.26	0.18	−0.55	0.63	0.19	0.41	−0.71	0.08	0.50	0.20	0.32	0.49	−0.05	0.04	0.31
cys	−0.48	0.80	−1.89	−0.32	−2.94	−0.84	0.16	0.14	−2.34	−0.26	−0.25	0.45	−1.99	−2.52	−0.19	0.47	−1.38	−2.09	−0.65	−0.46
gln	−0.14	0.50	0.87	−0.26	−0.84	0.73	0.48	−0.08	0.83	0.31	−0.22	0.86	0.16	−0.65	0.24	0.18	−0.20	−1.14	−0.64	0.37
glu	0.29	−0.91	0.78	0.18	0.16	0.48	1.03	0.88	−0.17	−0.47	−0.05	−0.34	0.06	−0.12	−0.18	0.15	0.81	−0.41	−0.16	0.34
gly	0.43	0.12	0.46	−0.55	0.14	−0.08	0.88	−0.30	0.19	0.31	0.23	0.82	0.08	0.90	−0.81	−0.05	0.10	0.51	−0.56	−0.12
his	0.66	−0.62	0.95	0.63	−2.34	0.83	−0.17	0.19	−3.30	−0.37	−0.17	0.79	−0.52	−1.03	−1.01	1.17	0.80	1.10	−1.02	0.67
ile	−0.79	0.31	0.48	0.19	−0.26	0.31	−0.47	0.31	−0.37	−0.27	−0.58	−0.08	0.03	−0.72	−0.33	0.49	−0.29	−0.63	−1.34	−0.98
leu	−0.13	0.28	1.17	0.41	−0.25	−0.22	−0.05	0.23	−0.17	−0.58	−1.42	0.43	−0.36	−0.89	0.08	0.80	0.30	−0.44	0.25	−1.03
lys	0.18	0.56	0.00	−0.71	0.45	0.86	−0.34	0.82	0.79	−0.08	0.43	1.01	0.77	0.32	0.42	−0.04	0.38	−0.38	−0.20	0.17
met	−0.82	0.40	1.92	0.08	−1.99	0.16	0.06	0.08	−0.52	0.03	−0.36	0.77	−1.08	−1.37	−0.75	−0.20	−0.40	−0.33	−0.19	−0.93
phe	−0.49	−0.89	0.36	0.50	−2.52	−0.65	−0.12	0.90	−1.03	−0.72	−0.89	0.32	−1.37	−2.38	1.41	−0.79	0.32	−0.80	−0.79	−0.43
pro	0.87	−0.59	0.08	0.20	−0.19	0.24	−0.18	−0.81	−1.01	−0.33	0.08	0.42	−0.75	1.41	0.06	0.60	0.47	−2.24	−0.68	0.32
ser	0.18	0.84	−0.78	0.32	0.47	0.18	0.15	−0.05	1.17	0.49	0.80	−0.04	−0.20	−0.79	0.60	0.03	0.32	−0.08	−0.83	0.52
thr	−0.45	−0.19	0.13	0.49	−1.38	−0.20	0.81	0.10	0.80	−0.29	0.30	0.38	−0.40	0.32	0.47	0.32	−0.38	−0.61	0.67	0.24
trp	−0.04	0.23	1.45	−0.05	−2.09	−1.14	−0.41	0.51	1.10	−0.63	−0.44	−0.38	−0.33	−0.80	−2.24	−0.08	−0.61	−0.97	−0.85	−0.41
tyr	−0.08	−1.64	0.43	0.04	−0.65	−0.64	−0.16	−0.56	−1.02	−1.34	0.25	−0.20	−0.19	−0.79	−0.68	−0.83	0.67	−0.85	−2.22	−0.50
val	−0.54	0.41	0.71	0.31	−0.46	0.37	0.34	−0.12	0.67	−0.98	−1.03	0.17	−0.93	−0.43	0.32	0.52	0.24	−0.41	−0.50	−0.53

[†]Two 20 × 20 contact matrices that were designed based on the training set of Table I. The first set was obtained using a constant function in the optimization (only feasibility was considered) whereas in the second case we optimize the function $[\langle\Delta^2\rangle-\langle\Delta\rangle^2]^{1/2}/\langle\Delta\rangle$ in addition to exact solution of the inequalities. The second matrix is therefore expected to yield more "isolated" native states. Nevertheless, in practice the recognition capacity of the optimized potential was lower than the recognition ability of the matrix without the optimization. The matrices are available electronically from the authors.

TABLE IVB. (Continued)

	ala	arg	asn	asp	cys	gln	glu	gly	his	ile	leu	lys	met	phe	pro	ser	thr	trp	tyr	val
ala	−0.05	0.41	−0.19	0.77	−0.45	−0.43	0.03	0.40	1.61	−0.95	−0.05	0.35	−0.67	0.22	0.71	−0.11	−0.49	−0.44	0.12	−0.71
arg	0.41	0.29	0.90	0.05	0.15	0.29	−1.13	0.51	−1.56	0.00	−0.69	0.09	0.54	0.02	1.23	1.11	−0.76	−0.03	−0.71	0.02
asn	−0.19	0.90	0.30	0.29	−1.11	1.42	1.58	0.11	0.61	−0.16	0.65	−0.08	2.78	0.19	0.15	−0.27	0.03	0.98	−0.09	0.28
asp	0.77	0.05	0.29	−0.66	0.21	−0.03	0.22	−0.81	−0.29	0.62	0.63	−0.89	0.31	0.49	0.16	−0.50	0.85	1.33	−0.61	−0.04
cys	−0.45	0.15	−1.11	0.21	−2.43	−0.56	0.11	1.03	−3.06	0.04	−0.82	0.91	−0.34	−3.01	−1.09	0.42	−1.55	−1.84	−2.38	−0.45
gln	−0.43	0.29	1.42	−0.03	−0.56	1.89	0.99	−0.15	0.31	0.78	−0.79	0.68	−0.59	−0.72	−0.68	0.34	−0.46	−1.01	−0.79	0.35
glu	0.03	−1.13	1.58	0.22	0.11	0.99	1.34	0.28	−0.16	−0.60	−0.19	−0.40	−0.52	−0.32	0.48	0.07	0.74	0.56	−0.62	−0.19
gly	0.40	0.51	0.11	−0.81	1.03	−0.15	0.28	0.89	0.00	0.16	−0.37	0.44	−0.43	0.72	−0.48	0.23	−0.02	0.22	−0.36	−0.26
his	1.61	−1.56	0.61	−0.29	−3.06	0.31	−0.16	0.00	−3.85	−0.54	−0.19	1.05	−0.07	−1.40	−1.30	2.37	2.47	1.06	−1.36	0.80
ile	−0.95	0.00	−0.16	0.62	0.04	0.78	−0.60	0.16	−0.54	0.15	−0.83	−1.27	0.09	−1.18	0.50	0.89	−0.24	−1.76	−1.14	−0.73
leu	−0.05	−0.69	0.65	0.63	−0.82	−0.79	−0.19	−0.37	−0.19	−0.83	−1.65	0.50	−0.05	−1.16	−0.27	0.51	0.55	−0.12	0.41	−0.65
lys	0.35	0.09	−0.08	−0.89	0.91	0.68	−0.40	0.44	1.05	−1.27	0.50	1.70	0.85	0.11	0.28	−0.58	0.02	0.22	0.72	0.19
met	−0.67	0.54	2.78	0.31	−0.34	−0.59	−0.52	−0.43	−0.07	0.09	−0.05	0.85	−1.21	−0.49	−0.63	−1.19	0.79	−0.13	−0.54	−1.16
phe	0.22	0.02	0.19	0.49	−3.01	−0.72	−0.32	0.72	−1.40	−1.18	−1.16	0.11	−0.49	−1.97	2.05	−0.18	−0.30	−1.51	−1.13	−0.62
pro	0.71	1.23	0.15	0.16	−1.09	−0.68	0.48	−0.48	−1.30	0.50	−0.27	0.28	−0.63	2.05	−0.53	1.73	1.42	−3.30	−1.35	0.11
ser	−0.11	1.11	−0.27	−0.50	0.42	0.34	0.07	0.23	2.37	0.89	0.51	−0.58	−1.19	−0.18	1.73	−0.72	−0.22	−1.00	0.43	0.47
thr	−0.49	−0.76	0.03	0.85	−1.55	−0.46	0.74	−0.02	2.47	−0.24	0.55	0.02	0.79	−0.30	1.42	−0.22	−0.56	−0.38	0.82	0.57
trp	−0.44	−0.03	0.98	1.33	−1.84	−1.01	0.56	0.22	1.06	−1.76	−0.12	0.22	−0.13	−1.51	−3.30	−1.00	−0.38	−1.63	−1.30	−0.78
tyr	0.12	−0.71	−0.09	−0.61	−2.38	−0.79	−0.62	−0.36	−1.36	−1.14	0.41	0.72	−0.54	−1.13	−1.35	0.43	0.82	−1.30	−1.55	−1.11
val	−0.71	0.02	0.28	−0.04	−0.45	0.35	−0.19	−0.26	0.80	−0.73	−0.65	0.19	−1.16	−0.62	0.11	0.47	0.57	−0.78	−1.11	−0.41

which is a quadratic function of the parameters and still accessible to linear programming techniques. Note also that the additional inequality prevents a collapse of the solution to $\{\Delta_i = 0\}_{i=1}^{N}$ which is a global minimum for the optimization of the variance. Hence, the Γ inequality is an attempt to keep the average gap larger than a given value, which determined an energy scale. In practice, we start at an average gap very close to Γ and we remain in the same set-up after optimizing the above variance.

Details of the Computations

The solution of linear inequalities is a field with an extensive body of applied research. There are numerous software packages that are appropriate for our needs saving us considerable programming time and decades of fine-tuning. We benefited from the linear programming (LP) package BPMPD,[19] which we used on a two CPU Silicon Graphics Origin 2000 with 1.5GB of memory. Memory was the prime limiting factor in our computations. We downloaded only 170,000 inequalities to the memory at a time. Nevertheless, because the number of parameters that we wanted to solve (typically 210) was much smaller than the number of inequalities, the following protocol converged rapidly: We sorted the inequalities according to the magnitudes of $\{U(X_i) - U(X_n)\}_{i=1}^{N}$, and considered only the 170,000 inequalities with the lowest values. A new potential for the 170,000 inequalities was obtained using the LP procedure. The newly calculated potential was used to re-

TABLE VA. "Training" Potential on a Large Set[†]

	ala	arg	asn	asp	cys	gln	glu	gly	his	ile	leu	lys	met	phe	pro	ser	thr	trp	tyr	val
ala	−0.15	−0.16	1.01	0.62	−0.96	0.16	−0.02	0.23	0.29	−0.85	−0.72	0.96	−0.85	−0.79	0.08	0.24	0.33	0.24	−0.50	−0.36
arg	−0.16	0.65	0.14	−0.57	−0.25	1.24	−0.70	0.84	0.36	−0.22	−0.04	1.24	0.67	0.43	−0.31	0.91	−0.44	−0.14	0.54	0.53
asn	1.01	0.14	−0.45	−0.53	0.18	0.57	−0.19	0.12	−0.06	0.22	−0.19	0.38	0.66	0.50	−0.25	−0.17	1.23	−0.04	0.22	0.54
asp	0.62	−0.57	−0.53	0.89	−0.50	0.78	0.83	0.37	−0.07	−0.18	1.12	−0.55	0.42	0.35	0.83	0.29	−0.35	−1.41	−0.27	0.20
cys	−0.96	−0.25	0.18	−0.50	−2.11	−0.54	0.65	−0.74	−0.66	−1.02	−0.24	−0.38	−1.30	−0.96	−0.45	−0.41	0.00	−0.32	−0.23	−0.58
gln	0.16	1.24	0.57	0.78	−0.54	0.41	0.28	−0.36	−0.60	0.66	0.10	−0.04	0.41	−0.10	−0.07	0.27	−0.28	0.00	−0.79	−0.06
glu	−0.02	−0.70	−0.19	0.83	0.65	0.28	1.59	0.48	1.27	0.31	0.99	−0.36	0.00	−0.56	0.24	1.14	−0.11	1.13	−0.84	0.47
gly	0.23	0.84	0.12	0.37	−0.74	−0.36	0.48	0.01	0.42	−0.04	−0.20	0.95	−0.63	−0.10	0.16	−0.09	0.73	0.87	−0.62	−0.43
his	0.29	0.36	−0.06	−0.07	−0.66	−0.60	1.27	0.42	−2.26	−0.31	−0.31	−0.01	0.18	0.52	0.93	−0.28	−0.27	0.91	−1.34	0.51
ile	−0.85	−0.22	0.22	−0.18	−1.02	0.66	0.31	−0.04	−0.31	−1.10	−1.15	−0.01	0.07	−0.76	0.43	−0.14	−0.18	−1.90	−1.42	−1.46
leu	−0.72	−0.04	−0.19	1.12	−0.24	0.10	0.99	−0.20	−0.31	−1.15	−1.60	1.02	−1.39	−1.24	−0.27	1.13	−0.18	−0.85	−0.53	−0.77
lys	0.96	1.24	0.38	−0.55	−0.38	−0.04	−0.36	0.95	−0.01	−0.01	1.02	2.28	1.66	−0.01	0.45	0.13	−0.16	0.33	−1.06	0.75
met	−0.85	0.67	0.66	0.42	−1.30	0.41	0.00	−0.63	0.18	0.07	−1.39	1.66	−1.89	−1.04	−0.22	−0.59	−0.62	−0.30	0.02	−1.22
phe	−0.79	0.43	0.50	0.35	−0.96	−0.10	−0.56	−0.10	0.52	−0.76	−1.24	−0.01	−1.04	−1.39	−0.09	−0.01	−0.29	0.58	−1.43	−0.89
pro	0.08	−0.31	−0.25	0.83	−0.45	−0.07	0.24	0.16	0.93	0.43	−0.27	0.45	−0.22	−0.09	0.41	0.48	0.34	−2.51	−0.17	0.75
ser	0.24	0.91	−0.17	0.29	−0.41	0.27	1.14	−0.09	−0.28	−0.14	1.13	0.13	−0.59	−0.01	0.48	1.31	0.03	−0.77	−0.86	0.14
thr	0.33	−0.44	1.23	−0.35	0.00	−0.28	−0.11	0.73	−0.27	−0.18	−0.18	−0.16	−0.62	−0.29	0.34	0.03	−0.41	0.32	−0.19	0.29
trp	0.24	−0.14	−0.04	−1.41	−0.32	0.00	1.13	0.87	0.91	−1.90	−0.85	0.33	−0.30	0.58	−2.51	−0.77	0.32	−1.35	−1.96	−0.18
tyr	−0.50	0.54	0.22	−0.27	−0.23	−0.79	−0.84	−0.62	−1.34	−1.42	−0.53	−1.06	0.02	−1.43	−0.17	−0.86	−0.19	−1.96	−1.35	−0.30
val	−0.36	0.53	0.54	0.20	−0.58	−0.06	0.47	−0.43	0.51	−1.46	−0.77	0.75	−1.22	−0.89	0.75	0.14	0.29	−0.18	−0.30	−1.32

[†]The same as in Table IV. This time the matrices are derived for the set of 572 proteins (Table II). The matrices are available electronically from the authors.

TABLE VB. (Continued)

	ala	arg	asn	asp	cys	gln	glu	gly	his	ile	leu	lys	met	phe	pro	ser	thr	trp	tyr	val
ala	−0.30	−0.17	0.65	0.43	−1.19	0.27	−0.06	0.46	0.69	−0.80	−0.73	0.71	−0.86	−0.31	−0.27	−0.05	0.41	0.09	−0.26	−0.27
arg	−0.17	0.33	−0.37	−0.51	1.65	0.13	−0.81	0.62	0.60	−0.54	0.19	0.71	1.49	0.24	0.15	0.83	−1.04	−0.88	0.42	0.07
asn	0.65	−0.37	−0.43	−0.79	0.38	1.31	−0.39	0.88	1.24	0.17	0.06	0.55	0.38	0.59	−0.78	0.24	0.47	−0.67	−0.18	0.27
asp	0.43	−0.51	−0.79	0.18	−0.38	0.33	0.65	−0.08	1.11	0.32	0.72	−0.63	0.44	0.10	0.90	0.27	−0.06	0.10	0.21	0.12
cys	−1.19	1.65	0.38	−0.38	−3.74	−1.22	1.37	−1.25	−0.70	−1.56	−0.35	−0.99	−2.16	−1.51	−0.96	−0.21	−0.20	−0.73	−0.63	−0.79
gln	0.27	0.13	1.31	0.33	−1.22	0.44	0.74	−0.16	0.86	0.12	−0.04	−0.55	0.22	0.06	0.62	−0.07	−0.33	−1.02	−1.01	−0.13
glu	−0.06	−0.81	−0.39	0.65	1.37	0.74	2.19	0.47	0.71	0.02	0.09	−0.53	−0.66	−0.43	0.90	0.43	0.39	0.07	−0.86	0.29
gly	0.46	0.62	0.88	−0.08	−1.25	−0.16	0.47	0.10	0.96	−0.63	−0.02	0.61	0.50	−0.17	−0.06	0.06	0.33	0.50	−0.99	−0.41
his	0.69	0.60	1.24	1.11	−0.70	0.86	0.71	0.96	−2.03	0.30	−0.30	0.12	−1.21	0.17	−0.07	0.17	1.13	0.78	−1.69	−0.33
ile	−0.80	−0.54	0.17	0.32	−1.56	0.12	0.02	−0.63	0.30	−0.93	−1.11	0.17	−0.15	−1.05	1.22	−0.03	−0.52	−1.24	−1.53	−1.07
leu	−0.73	0.19	0.06	0.72	−0.35	−0.04	0.09	−0.02	−0.30	−1.11	−1.02	0.33	−0.49	−1.06	−0.69	1.79	0.19	−1.45	−0.28	−1.08
lys	0.71	0.71	0.55	−0.63	−0.99	−0.55	−0.53	0.61	0.12	0.17	0.33	2.42	0.74	−0.47	0.41	0.18	−0.21	−0.19	−0.71	0.31
met	−0.86	1.49	0.38	0.44	−2.16	0.22	−0.66	0.50	−1.21	−0.15	−0.49	0.74	−0.84	−0.41	−0.29	−0.23	0.76	−0.56	−0.84	−0.98
phe	−0.31	0.24	0.59	0.10	−1.51	0.06	−0.43	−0.17	0.17	−1.05	−1.06	−0.47	−0.41	−1.45	0.59	0.39	−0.32	0.61	−1.91	−0.89
pro	−0.27	0.15	−0.78	0.90	−0.96	0.62	0.90	−0.06	−0.07	1.22	−0.69	0.41	−0.29	0.59	1.00	0.65	0.46	−2.16	−0.77	0.81
ser	−0.05	0.83	0.24	0.27	−0.21	−0.07	0.43	0.06	0.17	−0.03	1.79	0.18	−0.23	0.39	0.65	0.56	−0.41	−1.38	−0.86	0.16
thr	0.41	−1.04	0.47	−0.06	−0.20	−0.33	0.39	0.33	1.13	−0.52	0.19	−0.21	0.76	−0.32	0.46	−0.41	−0.28	0.39	−0.16	0.20
trp	0.09	−0.88	−0.67	0.10	−0.73	−1.02	0.07	0.50	0.78	−1.24	−1.45	−0.19	−0.56	0.61	−2.16	−1.38	0.39	−1.83	−1.91	−0.52
tyr	−0.26	0.42	−0.18	0.21	−0.63	−1.01	−0.86	−0.99	−1.69	−1.53	−0.28	−0.71	−0.84	−1.91	−0.77	−0.86	−0.16	−1.91	−1.87	−0.62
val	−0.27	0.07	0.27	0.12	−0.79	−0.13	0.29	−0.41	−0.33	−1.07	−1.08	0.31	−0.98	−0.89	0.81	0.16	0.20	−0.52	−0.62	−1.46

evaluate all the inequalities. If the result was unsatisfactory (i.e., some negative values were detected), the data were sorted and the new lowest 170,000 inequalities were solved again. This process was iterated until convergence, or until the solution was proven impossible. In practice, we found that only a few iterations were required to obtain a positive answer. More iterations were required to prove (when relevant) that the set had no solution.

Generating Decoy Structures and Inequalities

The simple contact potential outlined above is limited in scope and does not contain geometric constraints on the structure of the protein, such as specific restricted domains for the torsional angles, or the description of short-range interactions as hydrogen bonds. Such restrictions go far beyond the usual constraints of a self-avoiding chain. It is therefore important to select structures that do not violate these requirements. At present, these requirements are not built into the pairwise potentials.

Our selection of structures enforced "typical" backbone shapes by using (a) structures from the PDB, or (b) structures generated by the MONSSTER program[1] written by Skolnick, Kolinski, and Ortiz. MONSSTER is a Monte Carlo program to fold proteins on a lattice that uses extensive information on the local structure of the protein chain.

Inequalities based on structures from the PDB were generated as follows:

(a) Gapless threading through (i) 229 of the 246 struc-

TABLE VI. Potential with 211 Parameters†

	ala	arg	asn	asp	cys	gln	glu	gly	his	ile	leu	lys	met	phe	pro	ser	thr	trp	tyr	val	cyx
ala	−0.23	0.32	0.71	0.23	−0.51	0.03	−0.08	−0.10	0.35	−0.66	−0.24	0.58	−0.18	−0.54	−0.02	0.01	0.01	0.95	−0.57	−0.45	−0.51
arg	0.32	1.26	−0.04	−0.87	−0.22	−0.09	−0.73	0.49	1.02	0.02	−0.13	0.94	−0.11	0.31	0.36	0.83	0.21	−0.08	0.29	−0.01	−0.22
asn	0.71	−0.04	−0.41	−0.60	−0.13	0.16	0.00	0.28	−0.06	−0.02	0.13	0.74	0.72	0.55	−0.18	−0.65	0.28	−0.53	0.09	0.16	−0.13
asp	0.23	−0.87	−0.60	0.54	−0.16	0.47	0.40	−0.07	−0.07	0.27	0.64	0.01	0.29	0.36	0.12	0.04	−0.20	0.14	0.19	0.16	−0.16
cys	−0.51	−0.22	−0.13	−0.16	−0.74	−0.51	0.74	−0.05	−1.14	−0.70	−0.32	0.40	−1.47	−0.99	1.34	−0.08	0.27	−0.90	−0.62	−0.57	−0.74
gln	0.03	−0.09	0.16	0.47	−0.51	1.26	0.32	−0.08	0.30	−0.22	−0.37	−0.40	1.50	0.19	−0.01	0.22	0.21	0.50	0.09	−0.30	−0.51
glu	−0.08	−0.73	0.00	0.40	0.74	0.32	0.56	0.80	0.34	0.34	0.27	−0.35	0.46	−0.27	0.27	0.32	0.06	0.44	−0.80	0.23	0.74
gly	−0.10	0.49	0.28	−0.07	−0.05	−0.08	0.80	−0.16	0.32	0.10	−0.10	0.49	−0.30	0.09	0.23	0.01	0.88	0.25	−0.01	−0.24	−0.05
his	0.35	1.02	−0.06	−0.07	−1.14	0.30	0.34	0.32	−1.49	0.64	−0.13	0.01	−0.40	0.12	0.34	−0.37	0.36	−0.42	−0.41	0.35	−1.14
ile	−0.66	0.02	−0.02	0.27	−0.70	−0.22	0.34	0.10	0.64	−0.58	−0.86	0.16	−0.39	−0.28	0.39	0.28	−0.17	−1.16	−0.70	−1.07	−0.70
leu	−0.24	−0.13	0.13	0.64	−0.32	−0.37	0.27	−0.10	−0.13	−0.86	−1.13	0.41	−0.96	−0.84	−0.34	0.22	−0.13	−0.45	−0.43	−0.56	−0.32
lys	0.58	0.94	0.74	0.01	0.40	−0.40	−0.35	0.49	0.01	0.16	0.41	1.14	0.86	−0.52	0.53	0.21	−0.01	−0.23	−1.04	0.46	0.40
met	−0.18	−0.11	0.72	0.29	−1.47	1.50	0.46	−0.30	−0.40	−0.39	−0.96	0.86	−0.81	−1.09	−0.69	0.57	−0.02	−0.87	−0.55	−1.36	−1.47
phe	−0.54	0.31	0.55	0.36	−0.99	0.19	−0.27	0.09	0.12	−0.28	−0.84	−0.52	−1.09	−1.00	0.29	−0.16	−0.38	−0.06	−1.04	−0.59	−0.99
pro	−0.02	0.36	−0.18	0.12	1.34	−0.01	0.27	0.23	0.34	0.39	−0.34	0.53	−0.69	0.29	0.50	0.75	0.34	−1.04	−0.69	0.22	1.34
ser	0.01	0.83	−0.65	0.04	−0.08	0.22	0.32	0.01	−0.37	0.28	0.22	0.21	0.57	−0.16	0.75	0.19	0.02	−0.69	−0.20	−0.14	−0.08
thr	0.01	0.21	0.28	−0.20	0.27	0.21	0.06	0.88	0.36	−0.17	−0.13	−0.01	−0.02	−0.38	0.34	0.02	0.55	−0.23	−0.28	0.29	0.27
trp	0.95	−0.08	−0.53	0.14	−0.90	0.50	0.44	0.25	−0.42	−1.16	−0.45	−0.23	−0.87	−0.06	−1.04	−0.69	−0.23	−0.80	−0.90	0.30	−0.90
tyr	−0.57	0.29	0.09	0.19	−0.62	0.09	−0.80	−0.01	−0.41	−0.70	−0.43	−1.04	−0.55	−1.04	−0.69	−0.20	−0.28	−0.90	−0.72	−0.32	−0.62
val	−0.45	−0.01	0.16	0.16	−0.57	−0.30	0.23	−0.24	0.35	−1.07	−0.56	0.46	−1.36	−0.59	0.22	−0.14	0.29	0.30	−0.32	−0.80	−0.57
cyx	−0.51	−0.22	−0.13	−0.16	−0.74	−0.51	0.74	−0.05	−1.14	−0.70	−0.32	0.40	−1.47	−0.99	1.34	−0.08	0.27	−0.90	−0.62	−0.57	−6.03

†An optimized potential on the set of 572 proteins that includes 211 parameters. The extra parameter describes the possibility of a short-range (covalent) contact of cysteine residues. The "new" residue cyx has identical properties to cysteine except when it interacts with itself. In the last case, the binding energy (a result of the calculation) is much larger.

tures listed in the Hinds-Levitt set (Table I),[7] (ii) 572 proteins, which is a merge of the HL,[7] FSSP,[23] and SCOP[24] databases. In contrast to the HL set in which explicit protein names are given, the FSSP and the SCOP are conceptual divisions of the PDB into families. Only the first five classes of the SCOP database were considered: 1. All alpha, 2. All beta, 3. Alpha and beta (a/b), 4. Alpha + beta (a + b), 5. Multi-domain. Further screening was used to ensure reasonable quality of structures. For example, structures with less than one (average) contact per amino acid, and structures that include only C_α's were removed. The result was Table II. The set of 572 structures and sequences has homologous proteins (no more than 65% identity) as well as short chains (37 proteins with less than 65 residues). Considering homologous sets and small proteins broaden the scope of the usual 20 × 20 matrices and present a significant challenge to the optimization protocol.

(b) Application of the MONSSTER program to 75 proteins (Table III). We attempted to fold each of 75 proteins using 100 trajectories. Each folding trajectory was sampled during the run 200 times to generate "fresh" decoys. The "fresh" decoys were filtered to avoid structures with bad contacts. Then, the new structures together with the 572 sequences and structures were "fed" into the Linear Programming optimization protocol to obtain parameters consistent with the new enlarged set. A new run of the MONSSTER program with the last optimized potential was initiated to generate yet another set of fresh structures. The simulation with a self-consistent potential was repeated three times to yield 4,299,167 additional decoys. The last set has no feasible solution for (of course) the square well potential and also for the seven-step potentials provided in equation 6. (Nevertheless, some multistep potentials can be extracted and analysis of their properties will be published elsewhere.)

In the (a.i) to (a.ii) options, we thread the sequences through the known shapes from the PDB using all possible sequential arrangements of the sequence without gaps (gapless threading). We pre-compute the contact vectors associated with each residue according to the structure of the native proteins. A contact was defined between two amino acids if the distance between the geometric centers of the side chains (for the native sequence) was less than 6.4 and greater than 2 angstroms. The definition of contact types in option (b) follows the definition of the segments. The generalization is that for each pair of amino acid there are seven contact types, depending on the distance.

If the total number of residues in a native protein is r, we have r contact vectors, $\{\nu_j\}_{j=1}^{r}$ that list the contacts of each amino acid in the native protein. The same contact vectors are used for all other sequences and are not recomputed for each new sequence threading. This is an additional approximation that is difficult to avoid using a residue-based interaction potential.

For example, a sequence of length s is threaded into the above structure by computing the energy of all possible $r - s + 1$ arrangements (as mentioned earlier, only gapless threading is considered). The threading through the subset of the Hinds-Levitt set provided 3,415,191 inequalities, whereas the threading through the set of 572 proteins, resulted in 28,213,009 inequalities.

We also experimented with a potential with 211 parameters in which the Cys-Cys interactions were described by two parameters at two ranges of distances to account for the possibility of a covalent cross-link: (distances $2.0 \le r^{(1)}_{\text{cys-cys}} \le 3.1$ and $3.1 \le r^{(2)}_{\text{cys-cys}} \le 6.4$; parameters $p^{(1)}_{\text{cys-cys}}$ and $p^{(2)}_{\text{cys-cys}}$).

The feasible sets that were mentioned above were solved using the BPMPD program,[19] and the different potentials are summarized in the following tables: Table IV presents the contact potentials designed on the HL set, Table V lists the

TABLE VII. Testing Designed Potentials[†]

1aac	1ecrA	1mla	1thw
1ad2	1eczA	1mldA	1tlcA
1ad3A	1efrA	1mngA	1tml
1adt	1efuA	1molA	1tns
1agrA	1efuB	1mtyG	1ttqB
1ah6	1enh	1mut	1ulb
1ail	1fc2C	1mxa	1vaoA
1air	1fid	1nbaA	1vid
1ako	1finB	1neu	1vii
1akz	1fkd	1nfn	1vjs
1aln	1fps	1nkl	1vmoA
1alo	1froA	1nre	1vnc
1aly	1fruA	1nulA	1whi
1amm	1fua	1oacA	1wpoA
1an2A	1fuiA	1occH	1xaa
1aonA	1fxd	1octC	1yge
1aonO	1gadO	1olgA	1ytfB
1arb	1garA	1oneA	1zdhA
1axn	1gdoA	1ospO	2abd
1ayl	1gds	1parA	2ace
1bco	1geo	1pauA	2bct
1bdo	1gln	1pax	2bgu
1bglA	1gof	1pdo	2chsA
1bgw	1grj	1pdr	2dri
1ble	1gsa	1pfkA	2dtr
1bme	1gtrA	1pgs	2eng
1bmfA	1hbq	1phc	2erl
1bmtA	1hcb	1phr	2gstA
1bncA	1hcl	1pkm	2kauC
1bp1	1hleA	1pkn	2masA
1bta	1hme	1pne	2phlA
1bucA	1hpm	1poc	2polA
1bvp1	1hrdA	1pprM	2sil
1cdg	1httA	1pydA	2spcA
1cdwA	1hxn	1qasA	2tct
1chd	1igd	1rgeA	2tmdA
1cnsA	1ihfA	1rgp	2ts1
1crkA	1jbc	1rhgA	2vik
1cseE	1jetA	1rlaA	3btoA
1cseI	1jswA	1rlr	3cox
1ctj	1jud	1ropA	3ecaA
1cuk	1kapP	1rtm1	3fisA
1cwdL	1kfd	1rvvA	3grs
1dar	1kuh	1ryt	3iI8
1dcoA	1lbu	1sceA	3pte
1ddt	1lci	1seiA	3sdhA
1deaA	1lcpA	1sig	4aahA
1difA	1lfaA	1sphA	4at1A
1dik	1lit	1sriA	4icb
1dkgA	1lla	1sryA	4rhn
1dkzA	1llp	1sso	7aatA
1dnpA	1lxa	1stu	8catA
1dsbA	1mai	1tadA	8icoA
1ecl	1mbb	1tbd	
1ecmA	1mioA	1tfr	

[†]The list of proteins that was used to test the potential derived from the set of Table I and listed in Table IV. Gapless threading through all the structures generated the required decoys.

potential parameters obtained from training with the large set of 572 proteins, and Table VI has the potential with 211 parameters that was computed for the set of 572 proteins.

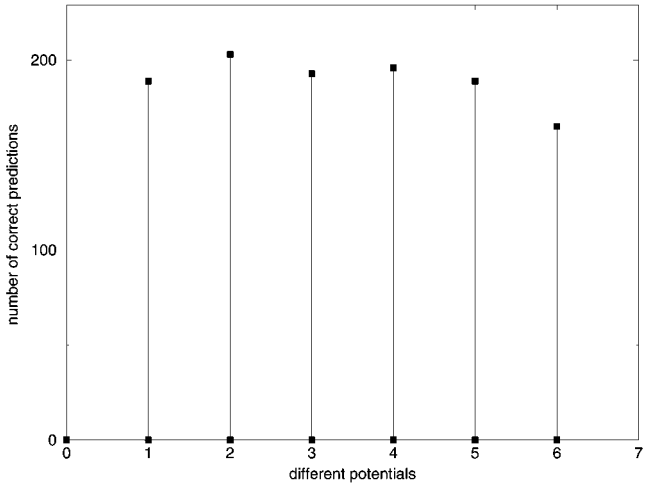

Fig. 1. The prediction success rate of different potentials on the SCOP database. The SCOP database is a division of the whole PDB into protein families. In Table VII, we list the representatives of each family that we used to construct a database of 218 proteins. The different potentials that were tested are: 1—Miyazawa-Jernigan Potential, 2—Betancourt-Thirumalai, 3—Hinds-Levitt, 4—Godzik et al., 5—Tobi et al. (derived from the Levitt set), 6—Tobi et al. (derived from the Levitt set and optimized). The present results are circled. See text for more details.

ANALYSIS OF FOLDING POTENTIALS

At the time in which this manuscript was written, there are already numerous published contact matrices, which are used in a variety of studies. It is likely that a few more will appear before this manuscript will appear in print. How different are these potentials from each other? Do they share critical features? Is there enough room to improve the 20 × 20 matrices, or is it just more of the same? These questions are clearly of considerable interest and we attempt below to provide some answers. Godzik et al.[20] addressed these questions by dividing the folding potential into two groups and suggesting alternative thermodynamic reference states. Here we are using a different analysis and consider a few parameter sets. We demonstrate what we considered to be a remarkable common feature.

The potentials we investigated are those of Hinds and Levitt,[7] Godzik et al.,[5] Miyazawa and Jernigan,[21] and Betancourt and Thirumalai.[9] The potential of Miyazawa and Jernigan was derived from the table in Ref. 21. We subtract the average hydrophobicity ($\varepsilon_{rr} = -2.55$), as suggested by the authors, p. 633).

The obvious way to compare potentials is by performance. Here we emphasize recognition: Is the energy of the native structure lower than the energies of all other structures? From thermodynamic and mechanistic point of view, a dimensionless measure, $\Sigma = \langle\Delta_i\rangle/(\langle\Delta_i^2\rangle - \langle\Delta_i\rangle^2)^{1/2}$ is also of interest in assessing the quality of the results.

We examine first the potential derived from the subset of the Hinds-Levitt database of protein structures. To make an independent test, structures were selected from the SCOP definition of protein families.[22,23] The selected structures were further analyzed to avoid repetition with the

TABLE VIII. Dimensionless Energy Measure†

Hinds-Levitt[7]	1.366014
Miyazawa-Jernigan[21]	1.477343
Godzik et al.[5]	1.544759
Betancourt-Thirumalai[9]	1.481314
Tobi et al. (no optimization)	1.521162
Tobi et al. (optimization)	1.754896

†The value of the function, $\langle\Delta\rangle/[\langle\Delta^2\rangle-\langle\Delta\rangle^2]^{1/2}$ for different potentials calculated by gapless threading using the set of 572 proteins.

Levitt set. The list of remaining proteins is in Table VII. Each of the sequences was threaded (without gaps) through the known structures of all the other proteins (including self) with the expectation that the true structure will provide the lowest energy. The results of the computer experiments are summarized in a histogram plot (Fig. 1).

The quality of most of the potentials is about the same, with our potential doing somewhat worse. Our potential with or without the additional optimization of the target function can be systematically improved, as the database is made larger. An improvement that is harder to get in statistical potentials (if the distribution functions converged numerically), or from optimization of average quantities (that may converge rapidly as well).

On the set of 572 proteins, we obviously do better than the other approaches. Also our dimensionless score, Σ, is significantly better (see Table VIII). Of course, this was also our training set so the test is not independent. Nevertheless, the set is extensive and covers essentially all families of folds in the PDB. It is not easy to come up with another independent (nonhomologous) set of structures for a check. The systematic and monotonic improvement that we see in the quality of the potential suggests the present approach as the method of choice for dealing with a very large number of decoys.

We constructed a limited set of 31 structures to test the potential derived from 572 proteins. It consists of the set: 1ptx, 1thx, 1pou, 1vin, 2pcy, 1kaz, 1hcp, 1erw, 1ghr, 2hvm, 1csn, 5p21, 1rgs, 1xel, 1djz, 1dyr, 1esl, 1fkj, 1gdy, 1gen, 1iae, 1icn, 1jcv, 1mls, 1onc, 1put, 1rci, 1tiv, 1bbt:1, 2wrp:R, 3sdp:A. The 31 sequences were threaded (with gapless threading) through their 31 corresponding structures and the 572 structures of the training set.

The potential we obtained from the 572 proteins (without optimization) missed the protein 1tiv, with optimization it misses also 2wrp:R. The Hinds-Levitt potential misses 5 proteins, The recent Miyazawa-Jernigan potential 4, Betancourt-Thirumalai 3, and Godzik at al., misses 3 as well. Hence, the potential we derived from 572 proteins is somewhat better than the potential we constructed from the subset of Hinds-Levitt structure database.

Eigenvalue Decomposition of Energy Matrices

It is of interest to analyze and compare the different energies (scoring functions) that were introduced in the past and in this manuscript. In what way are the matrices different? What aspects of the computations can be improved?

We seek a partitioning of the matrices to highlight those parts that are dissimilar and to those parts shared by the different parameterizations. Godzik et al.[20] proposed an interesting comparison based on differences in the reference state. Betancourt and Thirumalai[9] adjusted the reference state to obtain a new matrix with "zero" interaction approximately set on a contact with a water molecule. This choice is expected to work better when a difference between the folded and unfolded conformations is considered.

The correlation coefficient of the alternative matrix elements once adjusted to the same reference state is usually quite high (of order of 0.8, Godzik et al.[20]). The high correlation observed in earlier studies is somewhat discouraging. It suggests that further improvement of the energy function might be difficult because different methodologies converge to similar results. It is therefore interesting to note that the correlation between the parameters derived with the LP method and other approaches is quite low (Table IX).

We use eigenvalue decomposition to analyze the parameter set $\{p_\alpha\}_{\alpha=1}^{210}$. The parameters are written as a 20×20 symmetric matrix, Φ. An element of the matrix, Φ_{ij} is the interaction strength between amino acid type i and amino acid type j. The parameter matrix can be diagonalized and written in terms of its eigenvectors and eigenvalues: $\Phi = \sum_{l=1}^{20} e_l\rangle\,\lambda_l\,\langle e_l$, where the summation is over the 20 different

TABLE IX. Correlation Coefficients of Different Potentials†

	HL	MJ	BT	SK	T_HL	T_HLQ	T572	T572Q
HL	1	0.7	0.72	0.78	0.45	0.38	0.52	0.41
MJ	0.7	1	0.66	0.73	0.46	0.42	0.57	0.53
BT	0.72	0.66	1	0.76	0.57	0.47	0.61	0.62
SK	0.78	0.73	0.76	1	0.58	0.49	0.6	0.59
T_HL	0.45	0.46	0.57	0.58	1	0.81	0.57	0.62
T_HLQ	0.38	0.42	0.47	0.49	0.81	1	0.48	0.55
T572	0.52	0.57	0.61	0.6	0.57	0.48	1	0.79
T572Q	0.41	0.53	0.62	0.59	0.62	0.55	0.79	1

†Linear correlation is computed for the matrix elements Φ_{ij}. HL—the Hinds Levitt potential, MJ—Miyazawa Jernigan, BT—Betancourt Thirumalai, SK—Skolnick Kolinski, T_HL—Tobi et al. trained on the HL set, T_HLQ—Tobi et al. trained and optimized on the HL set, T572—Tobi et al. trained on the set of 572 proteins, T572Q—Tobi et al. trained and optimized on the set with 572 proteins.

TABLE X. Scalar Product of Different Eigenvectors of Contact Matrices[†]

Betancourt-Thirumalai	Hinds-Levitt	1	2	0.756
Betancourt-Thirumalai	Hinds-Levitt	2	1	0.799
Betancourt-Thirumalai	Hinds-Levitt	13	12	−0.737
Betancourt-Thirumalai	Hinds-Levitt	20	20	0.870
Betancourt-Thirumalai	Tobi et al.	1	2	0.731
Betancourt-Thirumalai	Tobi et al.	20	20	−0.956
Betancourt-Thirumalai	Godzik et al.	1	2	0.911
Betancourt-Thirumalai	Godzik et al.	2	1	−0.738
Betancourt-Thirumalai	Godzik et al.	11	9	−0.705
Betancourt-Thirumalai	Godzik et al.	20	20	0.837
Betancourt-Thirumalai	Miyazawa-Jernigan	1	2	−0.974
Betancourt-Thirumalai	Miyazawa-Jernigan	4	4	0.905
Betancourt-Thirumalai	Miyazawa-Jernigan	5	6	0.937
Betancourt-Thirumalai	Miyazawa-Jernigan	6	7	−0.884
Betancourt-Thirumalai	Miyazawa-Jernigan	7	8	0.993
Betancourt-Thirumalai	Miyazawa-Jernigan	9	9	0.852
Betancourt-Thirumalai	Miyazawa-Jernigan	12	11	0.868
Betancourt-Thirumalai	Miyazawa-Jernigan	13	12	−0.768
Betancourt-Thirumalai	Miyazawa-Jernigan	14	13	−0.788
Betancourt-Thirumalai	Miyazawa-Jernigan	15	15	−0.881
Betancourt-Thirumalai	Miyazawa-Jernigan	16	16	−0.890
Betancourt-Thirumalai	Miyazawa-Jernigan	18	18	0.803
Betancourt-Thirumalai	Miyazawa-Jernigan	19	19	−0.727
Betancourt-Thirumalai	Miyazawa-Jernigan	20	20	0.901
Betancourt-Thirumalai	Tobi et al. (opt.)	20	20	0.871
Hinds-Levitt	Tobi et al.	10	14	−0.766
Hinds-Levitt	Tobi et al.	20	20	−0.898
Hinds-Levitt	Godzik et al.	1	1	−0.859
Hinds-Levitt	Godzik et al.	2	2	0.868
Hinds-Levitt	Godzik et al.	19	19	0.902
Hinds-Levitt	Godzik et al.	20	20	0.957
Hinds-Levitt	Miyazawa-Jernigan	1	1	−0.721
Hinds-Levitt	Miyazawa-Jernigan	2	2	−0.813
Hinds-Levitt	Miyazawa-Jernigan	20	20	0.960
Hinds-Levitt	Tobi et al. (opt.)	1	1	0.733
Hinds-Levitt	Tobi et al. (opt.)	11	8	−0.740
Hinds-Levitt	Tobi et al. (opt.)	20	20	0.823
Tobi et al.	Godzik et al.	1	1	−0.705
Tobi et al.	Godzik et al.	20	20	−0.858
Tobi et al.	Miyazawa-Jernigan	1	1	−0.717
Tobi et al.	Miyazawa-Jernigan	20	20	−0.942
Tobi et al.	Tobi et al. (opt.)	12	15	0.784
Tobi et al.	Tobi et al. (opt.)	20	20	−0.914
Godzik et al.	Miyazawa-Jernigan	1	1	0.773
Godzik et al.	Miyazawa-Jernigan	2	2	−0.954
Godzik et al.	Miyazawa-Jernigan	18	19	−0.879
Godzik et al.	Miyazawa-Jernigan	20	20	0.927
Godzik et al.	Tobi et al. (opt.)	1	1	−0.760
Godzik et al.	Tobi et al. (opt.)	2	2	−0.827
Godzik et al.	Tobi et al. (opt.)	20	20	0.872
Miyazawa-Jernigan	Tobi et al. (opt.)	2	2	0.767
Miyazawa-Jernigan	Tobi et al. (opt.)	20	20	0.855

[†]The different matrices are diagonalized and sorted according to descending order of their eigenvalues. Scalar products between all the eigenvectors is calculated and products larger than 0.75 are reported in Table XI. Note the significant similarity between all the eigenvector with the lowest eigenvalue (eigenvector 20). Note also the significant similarity between the Betancourt-Thirumalai eigenvectors and those derived from the Miyazawa-Jernigan potential. See text for more details.

eigenstates. λ_l is the eigenvalue and $e_l\rangle$ is the corresponding eigenvector. The Dirac notation is used in which "$\langle e_l$" means a transposed vector of "$e_l\rangle$".

The first comparison we made is of the eigenvectors. The vectors are sorted in a descending order of the corresponding eigenvalues. In Table X we provide a list of pairs of eigenvectors from different matrices whose scalar product exceeds 0.7. In most matrices, only a few (extreme) eigen-

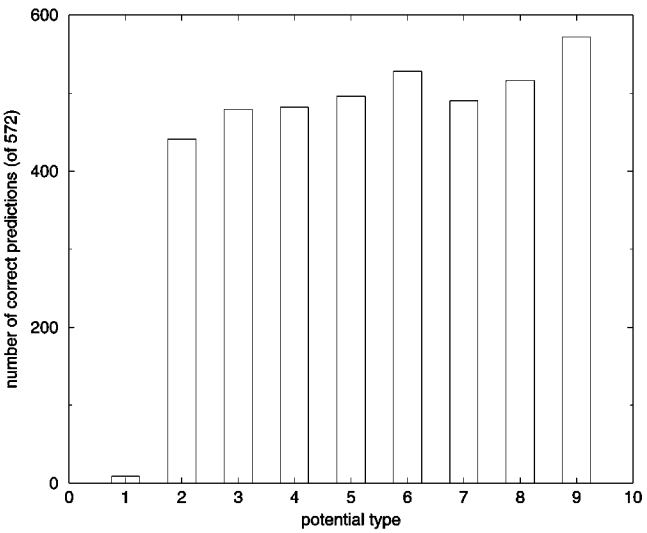

Fig. 2. Testing potential performances on the set of 572 protein structures (see Table II). 1—a random matrix with global attraction, 2—the H/P model, 3—A potential constructed from a single eigenvector with the lowest eigenvalue of the Miyazawa-Jernigan matrix, 4—A potential constructed from two eigenvectors of the Miyazawa-Jernigan matrix (the largest and the smallest eigenvalues), 5—The Miyazawa-Jernigan potential, 6—the Betancourt-Thirumalai potential, 7—the Hinds-Levitt potential, 8—the Godzik et al. potential, 9—Tobi et al. potential(s). Note that all the Tobi et al.'s potentials solve (by design) the set of 572 proteins exactly. The present results are circled. A better test of the last potential is described in the text.

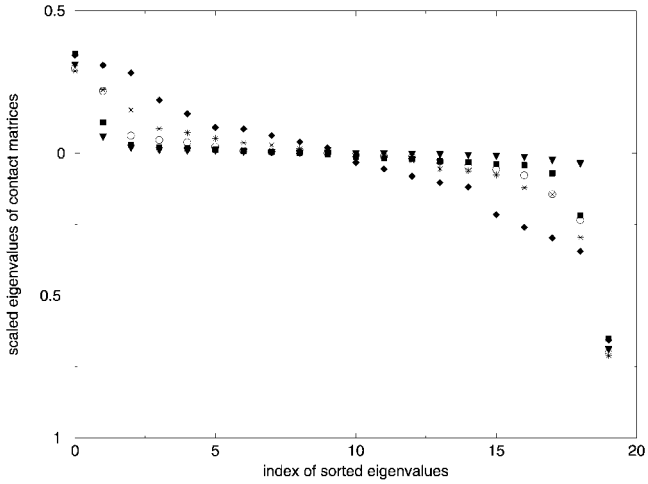

Fig. 3. Sorted and scaled eigenvalue of different potential matrices. The eigenvalues are sorted from large to small and are scaled so that $\lambda_{max} - \lambda_{min} = 1$. The dark squares are the eigenvalues of the Godzik et al matrix, the empty circles of Betancourt and Thirumalai, the dark diamonds of Tobi et al., the stars of Hinds and Levitt, and the dark triangles of Miyazawa and Jernigan.

vectors are similar. The matrix of Betancourt and Thirumalai (BT) and of Miyazawa and Jernigan (MJ) is an exception. The BT matrix was derived from the MJ matrix using a different reference state (the definition of the zero of the potential). It is therefore not surprising that significant eigenvector similarities remain.

Perhaps the most striking comparison is the persistent similarity of 20th eigenvectors throughout all the matrices we compared. Eigenvector 20 has the lowest eigenvalue. The lowest overlap of any of the pairs of 20th eigenvectors was 82%, and it went up to 96% for the comparison of the Betancourt-Thirumalai matrix and the matrix of Tobi at al. The two matrices were derived by very different means.

In addition to the obvious similarity of the eigenvectors with the lowest eigenvalue, some similarity is observed at the other end of the spectrum for the largest eigenvalues. The scalar product of either the first or the second eigenvectors with the corresponding eigenvectors of other matrices is typically above 0.70.

Scalar products of the rest of the eigenvectors do not show significant overlap. Hence, the matrices are different as far as 90% of their parameter space. Naively this observation would suggest that the individual matrix elements and the overall performance in predicting protein structures should be quite different as well. This is however not the case. When we tested the different matrices on the 572 proteins, we found comparable prediction performance for matrices that were not derived with linear programming (Fig. 2). Moreover, as we mentioned earlier, the correlation between the different matrix elements is quite high.[20]

The "mystery" of why the matrices are so similar in overall properties but so different in 18 eigenvectors is solved when we consider the eigenvalues of the matrices. In Figure 3 we showed the spread of all the eigenvalues from all matrices. To place all the data on the same scale, the eigenvalues are normalized so $\lambda_{max} - \lambda_{min} = 1$, which determines a unique energy scale. Because in our calculations the energy scale is arbitrary, such a scaling is convenient. Note that the position of the zero is about the same for all matrices. The location of the zero is not determined from the adjustment of the scale we performed. It therefore suggests that the different matrices have about the same relative balance of negative and positive eigenvalues. The negative (attractive) values are "stronger". Moreover, with the exception of the matrices derived with the LP protocol, the eigenvalues are dominated by the last and first few elements.

It is tempting to use a matrix with a few dominant eigenvectors and to reconstruct it using a smaller number of eigenvectors. The MJ matrix was "re-designed" twice. In the first design, a single eigenvector (the last one) was used, and in the second design, two eigenvectors (the first and last) were employed in the reconstruction. The performance of the reduced model, tested on the set of 572 proteins is reported in Figure 2. The reduction in prediction capacity is remarkably small and an indicator that the MJ matrix is essentially a matrix of rank two.

The success of the reduced matrix encourages us to try another simple model—H/P. Twenty amino acids are grouped into H (hydrophobic) or P (Polar) residues. The residues that are assigned to the hydrophobic (or polar) are listed in Table XI. A contact of type H-H scores "-1" and the contacts H-P and P-P scores 0. Hence, only the contacts of H-H type contribute to the energy. If the

TABLE XI. H/P Assignment: Division of the Different Amino Acids Into Hydrophobic (H) and Polar (P) Groups

Hydrophobic	Polar
ala	arg
val	asn
leu	asp
ile	gln
phe	glu
pro	gly
trp	lys
his	ser
cys	thr
met	
tyr	

number of H-H contacts is N_{HH}, the energy is $-N_{HH}$ and is parameter free.

The prediction capabilities of this potential are reported on Figure 2, and are quite good considering the simplicity of the model. A possible interpretation of the H/P model is of a single-eigenvector reduction of the contact matrix (a matrix of rank one) in which the contributions of individual amino acids to the single eigenvector are predetermined (and equal). Drawing on the similar performance of the H/P model and other approaches we expect that the single-eigenvector representation of the energy matrices will be based on a hydrophobic vector. This is indeed the case will be demonstrated.

In Figure 4, we show the components of vector 20 (the hydrophobic vector) with the dominant negative contribution to the energy. The components of all matrices are shown together with the component of a corresponding H/P model. It is clear that all the vectors are quite similar and indeed hydrophobic. The H/P model gives all the hydrophobic residues the same weight. The variable weighting of the different hydrophobic residues in the calculated hydrophobic vectors result in somewhat better prediction ability. It is nevertheless remarkable that significant recognition ability was already recovered in the simple model.

Another simple energy model is of global attraction with noise. Hence, all interaction elements, p_a-s are given by $p_a = \bar{p} + r$, where $\bar{p}$ is the average attraction between the residues and r is a random number. Hence, there are an infinite number of amino acid types. The average (attractive) energy is dressed by the addition of "noise". The above rule can be used to create (by sampling) a 20×20 matrix, and a matrix of this kind is reported in Table XII ($\bar{p}$ was set to -1 and r was sampled from a uniform distribution between -0.5 and 0.5). As is clear from Figure 2, the performance of this example is significantly lower compared to other models and it is unable to repeat the success of the other simple model, H/P.

It is also of interest to analyze the most repulsive vectors (with positive eigenvalues) that are quite similar. The degree of similarity is not as high as the "hydrophobic" vectors. Nevertheless, some interesting common features can be observed. For example, surface residues dominate the repulsive vectors and the residue that hates its neighbors the most is lysine (Fig. 5).

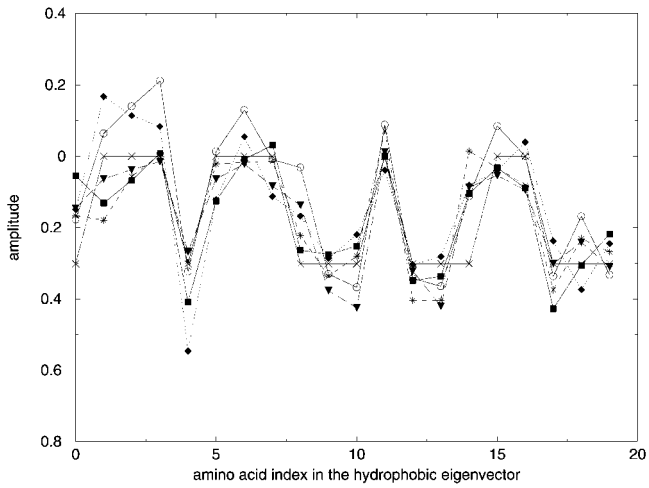

Fig. 4. The amplitudes of different amino acids in the eigenvector with the lowest eigenvalue (the hydrophobic eigenvector). The amino acids are sorted alphabetically, similarly to Tables IV and V. Different force fields are considered: The dark squares are from the Godzik et al. matrix, the empty circles from Betancourt and Thirumalai, the dark diamonds from Tobi et al. (optimized on the set of 572 proteins), the stars from Hinds and Levitt, and the dark triangles from Miyazawa and Jernigan. Also shown are "X"-s—the amino acid amplitudes used in the H/P model.

DISCUSSION AND CONCLUSIONS

At this point, it is useful to consider the operational differences between the present approach and other techniques to derive potential parameters. One common approach employs statistical analysis of contacts.[3,5–9,12] The use of statistical potentials goes back to the work of Tanaka and Scheraga[12] and has gathered significant momentum in the last decade. Other studies further explored the properties of statistical potentials, and suggested improvements and different protocols to derive these interactions (Hendlich et al.,[3] Miyazawa and Jernigan,[4] Godzik et al.,[5] Bryant and Lawrence,[6] Hinds and Levitt,[7] Goldstein et al.,[8] Betancourt and Thirumalai,[9] and Hao and Scheraga[14]).

Calculating statistical potentials does not require significant computational resources. However a number of conceptual difficulties remain. The probability of two residues A and B, $P_{\mathrm{contact}}(A,B)$ to be within a distance r_{cut} is computed from a carefully selected set of structures. The probability is compared with the reference distribution, $P(A)P(B)$, the probabilities of observing both A and B if they were independent. A "mean force" pair potential is then defined, which is also called a statistical potential:

$$V_{AB} = -k_B T \ln[P(A, B)/(P(A)P(B))] \quad (9)$$

Clearly, to correctly estimate these probabilities, the structures that are used to compute the frequencies of the contacts must be carefully chosen and a sufficient number of contacts must be enumerated. For example, they should not include any homologous proteins. The presence of

TABLE XII. A Random Matrix Potential†

	ala	arg	asn	asp	cys	gln	glu	gly	his	ile	leu	lys	met	phe	pro	ser	thr	trp	tyr	val
ala	−1.02	−1.31	−0.93	−1.45	−1.45	−1.42	−0.53	−1.48	−0.72	−0.59	−1.48	−1.01	−0.50	−0.97	−1.48	−0.81	−1.38	−1.16	−1.44	−1.32
arg	−0.74	−1.20	−1.13	−1.35	−0.63	−1.10	−0.92	−0.73	−0.92	−0.50	−1.08	−0.68	−1.24	−1.11	−0.60	−0.79	−0.72	−1.48	−1.17	−0.53
asn	−0.98	−1.00	−1.18	−1.21	−1.12	−0.69	−1.01	−0.91	−1.24	−0.69	−0.87	−0.94	−0.79	−1.28	−1.08	−1.21	−1.48	−1.39	−1.26	−0.56
asp	−1.40	−1.29	−1.22	−0.54	−0.96	−0.59	−1.16	−0.77	−0.70	−0.74	−1.01	−0.57	−1.02	−0.68	−0.57	−0.55	−1.45	−1.18	−0.78	−0.86
cys	−1.04	−1.02	−0.56	−1.43	−0.60	−0.77	−1.19	−1.16	−1.23	−1.08	−1.45	−1.17	−1.36	−1.45	−1.49	−0.99	−0.52	−0.93	−1.10	−0.91
gln	−0.60	−0.64	−1.06	−1.45	−0.51	−0.77	−0.59	−1.27	−1.34	−1.10	−0.80	−0.99	−0.58	−0.72	−1.06	−0.68	−0.89	−1.50	−1.34	−0.91
glu	−1.00	−0.76	−0.65	−1.08	−0.61	−1.23	−0.60	−0.80	−1.01	−1.32	−1.16	−0.79	−1.27	−1.03	−0.65	−0.79	−1.44	−0.70	−0.93	−0.57
gly	−1.47	−1.00	−0.92	−0.67	−1.41	−1.14	−1.06	−0.89	−1.04	−0.98	−0.72	−0.63	−1.08	−0.56	−0.76	−0.63	−0.76	−1.11	−0.99	−0.77
his	−1.22	−0.91	−1.30	−1.16	−0.91	−0.56	−1.28	−1.29	−0.56	−0.82	−0.58	−0.60	−1.10	−1.15	−0.65	−0.50	−1.06	−0.81	−0.68	−0.80
ile	−1.08	−0.85	−1.01	−1.16	−0.55	−1.11	−0.90	−1.30	−0.98	−1.04	−1.16	−1.30	−0.96	−1.31	−0.84	−0.71	−0.61	−1.14	−1.20	−1.48
leu	−0.96	−0.82	−0.79	−0.61	−0.62	−0.51	−0.57	−1.04	−0.97	−1.22	−0.92	−1.09	−0.83	−0.62	−1.27	−1.43	−1.05	−0.83	−0.57	−1.50
lys	−0.77	−0.51	−1.36	−0.67	−1.34	−1.46	−0.74	−0.82	−0.76	−1.13	−1.33	−0.58	−0.62	−1.17	−1.35	−0.53	−0.60	−1.34	−0.87	−0.57
met	−1.30	−0.57	−1.08	−1.02	−1.09	−0.76	−0.89	−1.01	−0.83	−1.43	−1.47	−1.39	−0.80	−1.21	−0.74	−0.78	−1.26	−1.11	−0.53	−0.66
phe	−1.08	−0.58	−0.64	−0.98	−0.76	−0.51	−1.47	−1.45	−0.83	−1.46	−0.63	−0.72	−0.66	−0.89	−0.65	−0.73	−1.40	−0.56	−1.24	−1.39
pro	−1.34	−0.82	−0.81	−1.05	−1.16	−1.10	−1.18	−1.28	−1.35	−1.45	−1.00	−1.29	−0.92	−1.29	−0.73	−1.06	−1.04	−1.32	−0.64	−0.84
ser	−0.88	−0.95	−0.59	−0.95	−0.88	−1.37	−1.33	−0.50	−1.21	−0.68	−1.12	−1.38	−0.87	−1.30	−1.09	−1.43	−1.14	−0.78	−0.65	−0.98
thr	−0.86	−0.60	−0.89	−0.92	−0.83	−1.25	−1.40	−0.97	−1.42	−1.25	−1.40	−1.37	−1.34	−0.52	−0.77	−0.75	−0.93	−0.71	−0.98	−1.01
trp	−1.08	−0.67	−0.87	−1.14	−1.14	−0.65	−0.55	−0.69	−1.48	−0.96	−0.90	−0.58	−0.67	−0.91	−1.09	−1.37	−0.76	−1.22	−1.48	−0.81
tyr	−0.99	−0.86	−0.58	−0.55	−0.81	−0.63	−1.23	−0.74	−0.55	−0.87	−1.28	−0.90	−1.08	−0.95	−0.89	−0.52	−1.12	−1.21	−1.02	−0.98
val	−1.00	−1.20	−1.19	−0.74	−1.15	−1.47	−1.29	−1.07	−1.17	−1.33	−1.29	−0.59	−1.21	−0.64	−0.99	−1.50	−1.03	−0.60	−0.99	−1.49

†The random matrix we attempted to use for structure predictions.

highly similar proteins may over count some types of contacts. Moreover, misfolded structures cannot be used in the extraction of the correct frequencies because their weight in the computations of distances is not known. These restrictions on the database limit considerably the set of structures that can be studied and the information content of the final potential.

In contrast, the solution of the inequalities can "digest" any structure that we "feed" in. It is possible that the new structure will not add to the quality of the potential but as a decoy it is unlikely to make it worse.

Finally, we note that the computations of the reference distributions $P(A)$ or $P(B)$ are not trivial. For example, Betancourt and Thirumalai (BT)[9] suggested a physically based refinement of the reference state and an adjustment to the Miyazawa-Jernigan (MJ) contact matrix. The existence of numerous reference states, and the difficulties in choosing appropriate training coordinates, make the "statistical" approach difficult to apply to an arbitrary set of structures. The present approach is influenced less by a choice of a reference state and is therefore more convenient conceptually. Numerically, however, it is considerably more demanding.

Other common approaches to design folding potentials include the optimization of the ratio T_f/T_g ,[8] the Z-score,[15] the σ folding parameter,[18] or the energy gap.[16] Detailed discussion of the alternatives is beyond the scope of the present work and is not included here. The optimization of the dimensionless gap is similar in spirit to the optimization of these functions. Of course, we also add the solution of the inequalities on top of the optimization.

An optimization with constraints is usually more difficult to perform compared to an optimization without constraints. So, we are definitely doing extra work here. However, a feature in our favor is that within the training set we are guaranteed to get the native state to be the lowest in energy. No such guarantee is provided for the unconstrained optimization.

Hence, our protocol examines also the local environment of the native state that (in principle) is where we expect most of the action to be. However, if the focus is only on the few lowest energy structures, important entropic effects are ignored. The number of different folds with energy close to the native can be high, resulting in an almost degeneracy of the native state. To avoid entropic bottlenecks, optimizing one of the above average functions has a significant merit, pushing the distribution away from the lowest values. Of course, using both viewpoints of optimizing quantities related to local and average properties is likely to be better than optimizing only one of them. The combination of the two is the approach taken here.

We derived a number of contact potentials using Linear Programming approach in conjunction with the optimization of average quantities. The present technique solves exactly the training set (provided that the solution is feasible). At the same time, it employs an optimization of a dimensionless measure, which is similar to the optimization of the thermodynamic measures such as T_f/T_g [8] σ,[18] or the Z-score.[15] The combination of both local and global approaches is likely to produce results that are better than the results of each of the methods separately. In addition, the new potential can be systematically improved as the number of decoys enlarged. There are no conceptual problems in adding more decoys except that of computer power.

The difficulty in optimizing averages over many decoys (without enforcing the inequalities constraints) is in the tail of the distribution of the energy gaps. If the number of decoys is very large (we already solved 10^7 structures), the tail can be significant. For example, having 99.99% of the inequalities correct (which will still optimize the averages nicely) will provide for the above set 1,000 structures with energies lower than the energy of the native state. Optimizing the dimensionless measure Σ and energy only, will not necessarily be enough to eliminate the misbehaving decoys. Using the inequality constraints as a "wall" that

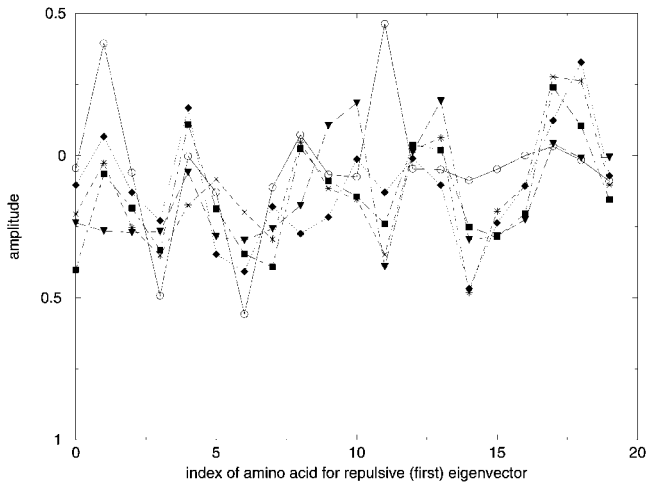

Fig. 5. The same as Figure 4, this time for the eigenvector with the largest eigenvalue. The H/P model is not shown. The amino acids are sorted alphabetically, similarly to Tables IV and V.

prevents the leak of the gap distribution to negative eigenvalues is therefore a useful feature of the present approach that cannot be found in the other methods.

We also analyzed different energy matrices and concluded that the representation of the hydrophobic interaction has reached a consensus. Further refinement of the interaction will need to focus on the weaker polar interactions that contribute little to the total energy but are able to make further subtle differences.

In another line of research employing linear programming techniques, we analyze the feasibility of the solution if the structures are sampled from a sophisticated computer program generating plausible proteins shapes. The set of inequalities was infeasible even for an extended set of 7×210 parameters. This is perhaps the most striking observation of the present investigation and leads to the conjecture that an exact folding potential for low-resolution models, which is based on a sum of pairwise interaction, does not exist.

ACKNOWLEDGMENTS

This research was supported by a grant from the Israel Science Foundation to R.E., and by the NIH Resource at the Cornell Theory Center. This research was also supported and by a FIRST award from the Israel Science Foundation to N.L. R.E. thanks Eytan Domany for discussions and for early sharing of his results, and Jon Kleinberg and Paul Chew for discussions, reading the manuscript, and making useful suggestions.

REFERENCES

1. Skolnick J, Kolinski A, Ortiz AR. MONSSTER: a method for folding globular proteins with a small number of distance restraints. J Mol Biol 1997;265:217–241.
2. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasi chemical approximations. Macromolecules 1985;18:534–552.
3. Hendlich M, Lackner P, Weitckus S, Floeckner H, Froschauer R, Gottsbacher K, Casari G, Sippl MJ. Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. J Mol Biol 1990;216:167–180.
4. Maiorov VN, Crippen GM. Contact potential that recognizes the correct folding of globular proteins. J Mol Biol 1992;227:876–888.
5. Godzik A, Kolinski A, Skolnick J. Knowledge-based potentials for protein folding: what can we learn from protein structures? Proteins 1996;4:363–366.
6. Bryant SH, Lawrence CE. An empirical energy function for threading protein sequence through folding motif. Proteins 1993; 16:92–112.
7. We used the structure database reported in: Hinds DA, Levitt M. Exploring conformational space with a simple lattice model for protein structure. J Mol Biol 1994;243:668–682; the potential for the comparison purposes was taken from Hinds DA, Levitt M. A lattice model for protein structure prediction at low resolution. Proc Natl Acad Sci USA 1992;89:2536–2540.
8. Goldstein RA, Luthey-Schulten ZA, Wolynes PG. Protein tertiary structure recognition using optimized Hamiltonians with local interactions. Proc Natl Acad Sci USA 1992;89:9029–9033.
9. 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. Prot Sci 1999;2:361–369.
10. Vendruscolo M, Najamanovich R, Domany E. A novel method to derive contact energy parameters from large databases obtained by threading (preprint).
11. Akutsu T, Tashima H. Linear programming based approach to the derivation of a contact potential for potential threading. Proceeding of the Pacific Symposium on Biocomputing, 1998, p. 413–424.
12. Tanaka S, Scheraga HA. Medium and long range interaction parameters between amino acids for predicting three dimensional structures of proteins. Macromolecules 1976;9:945–950.
13. Vendruscolo M, Domany E. Pairwise contact potentials are unsuitable for protein folding. J Chem Phys 1998;109:11101–11108.
14. Hao HA, Scheraga HA. Optimizing potential functions for protein folding. J Phys Chem 1996;100:14540–14548.
15. 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.
16. Mirny AL, Shakhnovich EI. How to drive a protein folding potential? A new approach to an old problem. J Mol Biol 1996;264: 1164–1179.
17. Park BH, Huang ES, Levitt M. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. J Mol Biol 1997;266:831–846.
18. Klimov DK, Thirumalai D. Criterion that determines the foldability of proteins. Phys Rev Lett 1996;76:4070–4073.
19. Meszaros CS. Fast Cholesky factorization for interior point methods of linear programming. Comp Math Appl 1996;31:49–51.
20. Godzik A, Kolinski A, Skolnick J. Are proteins ideal mixtures of amino acid? Analysis of energy parameter sets. Prot Sci 1995;4: 2107–2117.
21. 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.
22. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
23. Holm L, Sander C. The FSSP database of structurally aligned protein fold families. Nucl Acids Res 1994;22:3600–3609.