

# Maximum Feasibility Guideline in the Design and Analysis of Protein Folding Potentials

JAROSLAW MELLER,<sup>1,2</sup> MICHAEL WAGNER,<sup>3</sup> RON ELBER<sup>1</sup>

<sup>1</sup>Department of Computer Science, Upson Hall 4130, Cornell University, Ithaca, New York 14853

<sup>2</sup>Department of Computer Methods, Nicholas Copernicus University, 87-100 Torun, Poland

<sup>3</sup>Department of Mathematics and Statistics, Old Dominion University, Norfolk, Virginia 23529-0011

Received 23 February 2001; Accepted 2 August 2001

**Abstract:** Protein folding potentials are expected to have the lowest energy for the native shape. The Linear Programming (LP) approach achieves exactly that goal for a training set, or indicates that this goal is impossible to obtain. If a solution cannot be found (i.e., the problem is infeasible) two possible routes are possible: (a) choosing a new functional form for the potential, (b) finding the best potential with a feasible subset of the data, and (or) detecting inconsistent subset of the data in the training set. Here, we explore option (b). A simple heuristic for finding an approximate solution to an infeasible set of linear inequalities is outlined. An approximately feasible solution is obtained iteratively, starting from a certain initial guess, by computing a series of analytic centers of the polyhedra defined by all the inequalities satisfied at the subsequent iterations. Standard interior point algorithms for Linear Programming can be used to compute efficiently the analytic center of a polyhedron. We demonstrate how this procedure can be used for the design of folding potentials that are linear in their parameters. The procedure shows an improvement in the quality of the potentials and sometimes points to flaws in the original data.

© 2002 John Wiley & Sons, Inc. J Comput Chem 23: 1–8, 2002

**Key words:** linear programming; interior-point methods; folding potentials

## Introduction

The basic requirement for protein folding potentials is their ability to distinguish native-like from nonnative shapes. This can be achieved by an appropriate choice of the potential (or energy) function, such that for each pair of native and misfolded structures the following constraints are satisfied:

$$\Delta E_{\text{mis, nat}} = E_{\text{misfolded}} - E_{\text{native}} \geq \varepsilon. \quad (1)$$

Here,  $E_{\text{native}} \equiv E(\mathbf{X}_{\text{nat}}; \mathbf{z})$  is the energy of the native structure  $\mathbf{X}_{\text{nat}}$ ,  $\mathbf{z}$  is the vector of parameters,  $E_{\text{misfolded}} \equiv E(\mathbf{X}_{\text{mis}}; \mathbf{z})$  represents the energies of the misfolded (nonnative) structures  $\mathbf{X}_{\text{mis}}$  and  $\varepsilon$  is a positive constant. In other words, we require that the energies of native structures are lower than the energies of misfolded structures.

For energy models linear in their parameters, the set of inequalities in eq. (1) can be solved for the parameters  $\mathbf{z}$  by standard Linear Programming (LP) tools. Note that the inequalities of eq. (1) define a set of cuts (hyperplanes) in the parametric space. The intersection of the corresponding feasible (closed) half spaces defines a convex polyhedron (see Fig. 1). LP solvers provide a feasible solution  $\mathbf{z}^*$  that belongs to the feasible polyhedron [i.e.,  $\mathbf{z}^*$  satisfies all the constraints in (1)] and optimizes certain linear objective function.

The LP approach for the design of protein folding potentials, which was pioneered by Maiorov and Crippen,<sup>1</sup> usually involves solving very large sets of inequalities, and the efficiency of LP algorithms is an important issue.

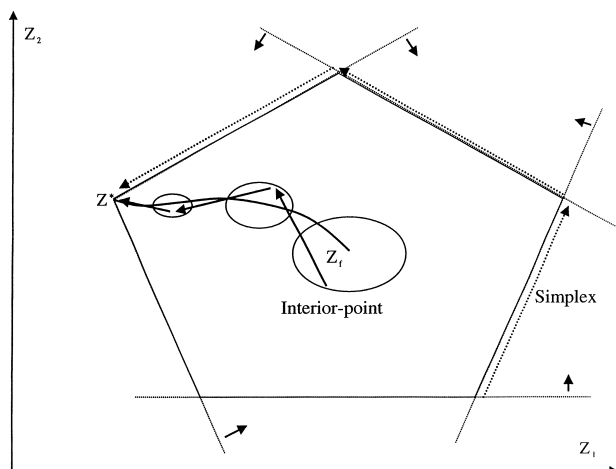
Recently, the LP approach has been applied to the design of various folding and threading potentials.<sup>2–7</sup> It has been found, for example, that simple contact pairwise potentials are not sufficient for recognition of all types of protein shapes.<sup>2,3</sup> The set of inequalities in eq. (1), which we attempt to solve, proves infeasible for a sufficiently large sample of native and misfolded shapes. Infeasibility of large training sets with different functional models was also used as a guideline to design optimal threading potentials. We seek functional forms for the potentials that preserve the exact recognition of all the proteins in the training set and minimize the number of required potential parameters.<sup>5,6</sup>

Here, we present a heuristic approach to find an approximate solution, which satisfies a possibly large subset of an infeasible set

**Correspondence to:** R. Elber; e-mail: ron@cs.cornell.edu

Contract/grant sponsors: NIH NCRR grant to the Cornell Theory Center and DARPA (to R.E.)

Contract/grant sponsor: Polish State Committee for Scientific Research; contract/grant number: 6 P04A 066 14



**Figure 1.** A schematic plot of a polyhedron representing the feasible volume defined by cuts in the parametric space. Given certain feasible set of linear inequalities with bounded variables, the intersection of feasible half spaces (indicated by arrows) takes a form of a polytope. Central path starts at the analytic center of this polytope ( $\mathbf{z}_c$ ) and terminates at the optimal solution ( $\mathbf{z}^*$ ) of the LP problem with an objective function to optimize. The interior point methods proceed through a series of interior points (usually obtained by the subsequent steps of the Newton method) that are located near the central path (arrows in the figure). In practical implementations steps out of the feasible polytope may be allowed, and only the converged solution is guaranteed to be feasible. If there is no function to optimize the interior point algorithms converge to the analytic center.

of inequalities. We call it the “maximum feasibility” (MaxF) guideline. The MaxF procedure is based on a special property of interior point algorithms for LP. Namely, the interior point methods provide the so-called analytic center of the feasible polyhedron (defined in terms of logarithmic barriers “repelling” the solution from the constraints) when the objective function is not used to “force” the convergence to an optimal solution on a facet of the polyhedron. For a bounded polyhedron (which is called the *polytope*) the analytic center is unique. We consider here only bounded problems.

Starting from a set of constraints that are satisfied by a certain initial guess of the solution, a series of “maximally feasible” approximations is computed. The subset of all the inequalities satisfied by the previous approximation, which defines a feasible polytope, is solved using an interior point method. The analytic center of the feasible polytope, obtained as a solution, becomes our next “maximally feasible” approximation. The new approximation satisfies at least as many constraints as the previous partial solution. If no further constraints can be satisfied the procedure stops. The idea behind this heuristic is that the analytic center, which is usually located close to the center (in the topological sense) of the feasible polytope, is likely to satisfy more constraints than an off-centered guess.

Using the MaxF guideline allows us to go beyond a simple feasibility test when assessing the quality of a given model, and may provide a better insight for improving the functional models of folding potentials. It also provides a simple way to improve potentials that are not optimized to satisfy inequality constraints of the type of

eq. (1), for example, the commonly used statistical potentials.<sup>8–10</sup> The method is outlined in the Methods section, whereas the results of numerical examples (for a series of medium size problems) are demonstrated in the Results section.

## Methods

### Interior Point Algorithms

Interior point methods, due to their polynomial time complexity<sup>11, 12</sup> and practical efficiency are nowadays a method of choice for large-scale linear optimization problems.<sup>13–16</sup> An interior point algorithm generates a series of points away from the boundary of the polyhedron (unlike the simplex algorithm, which proceeds along the edges of the feasible region;<sup>14</sup> see also Fig. 1). These points are near a smooth curve, called the *central path*, which is contained within the interior of the feasible polyhedron and terminates at an optimal and complementary solution on a facet or at the vertex (if the optimal solution is unique) of the polyhedron.<sup>17</sup>

Let us consider a linear programming problem [which will be referred to as (LP)] of the form:

$$\{\min f_0(\mathbf{z}); \mathbf{z} \in \mathbf{R}^n; f_i(\mathbf{z}) \leq b_i, i = 1, \dots, m\}, \quad (2)$$

where  $\mathbf{z}$  is a vector of  $n$  variables and the objective function to be minimized,  $f_0$ , as well as the constraints functions,  $f_i$ , are linear. One can define the logarithmic barrier function associated with (LP) as:

$$\phi_B(\mathbf{z}, \mu) = \frac{f_0(\mathbf{z})}{\mu} - \sum_{i=1}^m \ln(b_i - f_i(\mathbf{z})), \quad (3)$$

where  $\mu > 0$  is the barrier parameter. If the feasible region of (LP) is bounded (i.e., all variables,  $z_j$ ;  $j = 1, \dots, n$ , are bounded from below and from above by finite numbers) and nonempty [otherwise (LP) is called *infeasible*], then for each value of  $\mu$  the barrier function,  $\phi_B(\mathbf{z}, \mu)$ , achieves the minimal value at a unique (feasible) point,  $\mathbf{z}(\mu)$ , which is called the  $\mu$ -center.<sup>13, 16</sup>

The *central path* is defined as the set of  $\mu$ -centers, where  $\mu$  changes from  $\infty$  to 0. In the limit of  $\mu \rightarrow 0$ , when minimizing the barrier function of eq. (3), one obtains the desired optimal and feasible solution of (LP)—see Figure 1. Barrier functions of the form specified in eq. (3) are commonly used in the interior point methods for inequality constraints.<sup>13–17</sup> The advantage of reformulating the constrained optimization problem of (2) into unconstrained, nonlinear optimization problem of (3) is that the nonlinear minimization techniques (e.g., gradient or Newton methods) can be applied.

The unique minimum of the barrier function in the limit of  $\mu \rightarrow \infty$  is called the *analytic center* of the feasible region.<sup>13, 16</sup> The central path always starts at the analytic center and, in the absence of an objective function to optimize, the interior point algorithms converge to the analytic center. We emphasize that, in practice (as in the popular infeasible primal-dual implementations, for example<sup>17</sup>) the functional constraints,  $f_i$ , are often initially relaxed and the method proceeds through points away from the central path that may not belong to the feasible polytope. Therefore, the analytic center is reached only upon convergence of the Newton

procedure. There are many parameterizations of the central path. In particular, different barrier functions (as, e.g., weighted logarithmic barriers) can be applied.<sup>16, 18</sup> Therefore, the actual position of the analytic center may vary between different implementations.

Note that solving a set of linear inequalities is equivalent to solving a special case of (LP), obtained by setting the objective function in (2) to zero,  $f_0(\mathbf{z}) = \mathbf{0} \cdot \mathbf{z}$ . Therefore, when solving a set of inequalities by an interior point algorithm we obtain the analytic center of the feasible polyhedron as a solution. We comment that just solving a set of inequalities (which is by duality theorem equivalent to solving an LP problem<sup>14</sup>) is of the same complexity as the original (LP) problem, with an objective function to optimize.

It should also be pointed out that the analytic center does not correspond (in general) to the center of the feasible polytope in the topological sense. Redundant constraints that do not define the boundaries of the polytope contribute to the barrier function in eq. (3) as well, “repulsing” the analytic center. However, the analytic center is always located away from any individual cutting hyperplane, due to singularity of the logarithm function at zero.

#### “Maximum Feasibility” Guideline

So far, we were assuming that the problem was feasible, for instance, that there exists a solution to (LP). If the problem proves infeasible it is useful to understand the source of the infeasibility and to assess the “hardness” of the problem. In other words, we would like to know what is the largest subset of constraints that can be satisfied simultaneously, which will be referred to as the Largest Feasible Subset (LFS). The LFS (in the mathematical literature often referred to as the maximum cardinality satisfiable subset) can be used to generate an approximate solution to our problem. Moreover, the analysis of the constraints that cannot be satisfied may help in suggesting a new functional form, new parameters for the potential, or perhaps points to problems in the database.

Unfortunately, finding the LFS is an NP-hard problem.<sup>19</sup> Several heuristic approaches that provide approximate solutions at a low computational cost have been proposed in the past.<sup>20, 21</sup> Such heuristics are of theoretical interest as well, because the problem of finding the LFS is closely related to the so-called satisfiability problem, which is at the origin of complexity theory.<sup>22</sup> Below, we define a simple, iterative procedure, referred to as the “maximum feasibility” guideline. The MaxF heuristic can be advantageous when a reasonable partial solution to an infeasible problem is available, which is usually the case in the design of folding potentials.

Let  $\mathbf{z}_0 \in \mathbf{R}^n$  be our initial guess of the solution, which satisfies certain a subset of inequalities in (LP). We will denote this set as  $\mathbf{P}(\mathbf{z}_0) = \{f_i; f_i(\mathbf{z}_0) \leq b_i\}$ . Because we assumed that (LP) is infeasible, there are some inequalities in (LP) that are not satisfied by  $\mathbf{z}_0$ . Let  $\mathbf{z}_1$  be the analytic center of the set of inequalities satisfied by the initial guess,  $\mathbf{P}(\mathbf{z}_0)$ . As described in the previous section,  $\mathbf{z}_1$  can be obtained by an interior point method when solving the following set of inequalities:

$$\{f_i(\mathbf{z}) \leq b_i; f_i \in \mathbf{P}(\mathbf{z}_0)\}. \quad (4)$$

In other words, we solve a feasible LP problem with the inequalities satisfied by the initial guess and without a function to optimize (the objective function is set to zero).

The analytic center of the initial polytope becomes our new guess of the solution. Let  $\mathbf{P}(\mathbf{z}_1) = \{f_i; f_i(\mathbf{z}_1) \leq b_i\}$  be the set of inequalities satisfied by  $\mathbf{z}_1$ . The new solution satisfies all the constraints of the initial problem, and therefore,  $\mathbf{P}(\mathbf{z}_0) \subseteq \mathbf{P}(\mathbf{z}_1)$ . In general, let  $\mathbf{z}_{k+1}$  be the analytic center of the polytope defined by  $\mathbf{P}(\mathbf{z}_k)$ , i.e.,  $\mathbf{z}_{k+1}$ , is obtained as the solution of the following set of inequalities:

$$\{f_i(\mathbf{z}) \leq b_i; f_i \in \mathbf{P}(\mathbf{z}_k)\}. \quad (5)$$

Obviously,  $\mathbf{P}(\mathbf{z}_k) \subseteq \mathbf{P}(\mathbf{z}_{k+1})$ , that is at each iteration we solve at least all the constraints included in the previous iteration. If no improvement is observed, i.e., when  $\mathbf{P}(\mathbf{z}_k) = \mathbf{P}(\mathbf{z}_{k+1})$ , the procedure stops.

The analytic center of the final polytope, which we denote as  $\mathbf{z}_f$ , defines our best approximate (partial) solution to an infeasible set of inequalities, which is our goal here. Alternatively, if our original problem involves a function to optimize, it can be now optimized over the final feasible polytope. The set of inequalities defining the final polytope,  $\mathbf{P}(\mathbf{z}_f)$ , becomes our approximation to the LFS in (LP). Note that this is not an approximation in the topological sense because just one inequality may dramatically change the shape of the polytope. The number of inequalities in the problem that do not belong to  $\mathbf{P}(\mathbf{z}_f)$  can be used to measure the quality of the approximation.

The level of success of the MaxF procedure is critically dependent on the choice of the initial guess and the structure of the problem (in practice a reasonable guess can be obtained from a statistical potential). Imagine, for example, a feasible problem with its feasible polytope  $\mathbf{P}$ . Let us now define a new problem by adding one more constraint, such that the intersection of the polytope and the feasible half space of the new cut is empty, for instance, the new problem is infeasible. The LFS of the new problem is defined by the initial polytope  $\mathbf{P}$ . If we start from a point in the feasible half space of the new cut our procedure will fail to provide a reasonable approximation to  $\mathbf{P}$ . On the other hand, however, if we start from a point in the infeasible half space, then we observe an improvement of the initial guess (see the MaxF “trajectories” in Fig. 2).

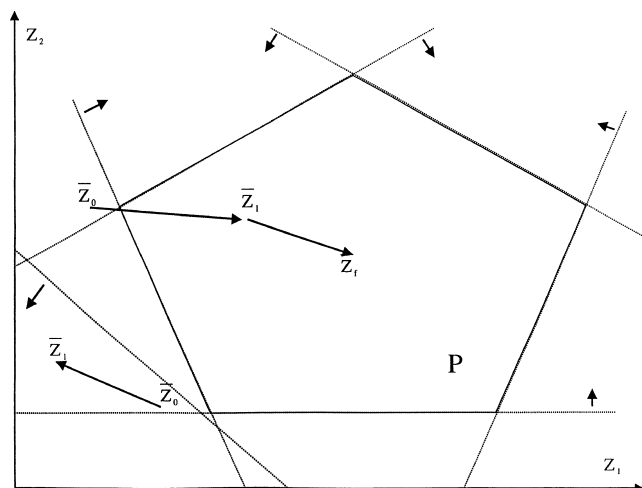
#### Linear Programming Protocol for the Optimization of Folding Potentials

In the next section we demonstrate the numerical performance of the MaxF procedure using realistic examples, relevant for the design of folding potentials. The LP problem in the design of folding potentials is concerned with the exact recognition of the native structures with respect to misfolded shapes in a training set.

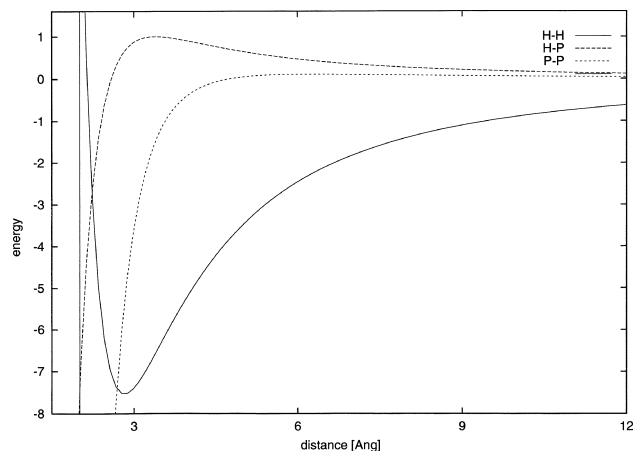
Any potential energy function  $E(\mathbf{X}; \mathbf{z})$  can be expanded in terms of a basis set (say  $\{n_\gamma(\mathbf{X})\}_{\gamma=1}^\infty$ ), in which the coefficients are unknown parameters:

$$E(\mathbf{X}; \mathbf{z}) = \sum_{\gamma=1}^{\infty} z_\gamma n_\gamma(\mathbf{X}). \quad (5')$$

The information on the protein structure  $\mathbf{X}$  (and implicitly on its sequence  $S$ ) is “buried” in  $n_\gamma(\mathbf{X})$ . A good choice of the basis set will converge the sum to the right solution with only a few terms. Of course, such a choice is not trivial to find and one of the advantages



**Figure 2.** A pictorial representation of a series of analytic centers obtained by applying the “maximum feasibility” guideline to an infeasible LP problem (feasible half spaces are indicated by arrows pointing out from the cutting hyperplanes). Bad and good scenarios are illustrated by two trajectories starting from different initial solutions. In practice, a reasonable guess that provides a good starting point for MaxF can be obtained from statistical potentials.



**Figure 3.** A Lennard–Jones-like potential for two types of amino acids obtained using the MaxF procedure. The functional form is  $A_{\alpha\beta}/r_{ij}^6 + B_{\alpha\beta}/r_{ij}^2$ , where the indices  $\alpha$  and  $\beta$  denote the amino acid types, the indices  $i$  and  $j$  are the positions along the chain and the coefficients  $A_{\alpha\beta}$ ,  $B_{\alpha\beta}$  are optimized using the LP approach coupled with the MaxF guideline. Interactions of different types are denoted as HH, HP, and PP, where H stands for hydrophobic and P for polar residues, respectively. The coefficients  $A$  and  $B$  are given in Table 3. Note that, similar to the contact HP potentials, the HH interactions are highly favorable, whereas the HP and PP interactions contribute little to the energy because there are only few contacts corresponding to very short distances (251 with distances shorter than 3 Å and 9 with distances shorter than 2.5 Å, respectively, out of the 291,651 native contacts used in the training). There are no contacts in the training set with distances shorter than 2 Å, which corresponds to an infinite wall at that distance (represented as a vertical line in the figure).

of the LP approach to the design of folding potentials is that it allows exploring different possibilities and assess them using the infeasibility test.

Let us consider the widely used pairwise folding potentials.<sup>8–10</sup> The energy of the protein of a sequence  $S$  and a structure  $\mathbf{X}$  is a sum of all pairs of interacting amino acids,

$$E_{\text{pairs}} = \sum_{i < j} \phi'_{ij}(\alpha_i, \beta_j, r_{ij}). \quad (6)$$

The pair interaction model— $\phi_{ij}$  depends on the distance between sites  $i$  and  $j$ , and on the types of the amino acids,  $\alpha_i$  and  $\beta_j$  at sites  $i$  and  $j$ , respectively. We consider both: a simple contact potential and a continuous pairwise potential.

In the case of the contact potential, two amino acids are considered in contact if the geometric centers of the side chains are closer than 6.4 Å. The interaction model reads:

$$\phi_{ij}(\alpha_i, \beta_j, r_{ij}) = \begin{cases} \varepsilon_{\alpha\beta} & 1.0 < r_{ij} < 6.4 \text{ \AA} \\ 0 & \text{otherwise} \end{cases}, \quad (7)$$

where  $\varepsilon_{\alpha\beta}$  is a matrix of all the possible contact types (we drop the subscripts  $i$  and  $j$  for convenience). For example, it can be a  $20 \times 20$  matrix for the 20 amino acids. Alternatively, it can be a smaller matrix if the amino acids are grouped together to fewer classes. The entries of  $\varepsilon_{\alpha\beta}$  are the target of parameter optimization.

An example of a more realistic interaction model is the “distance power” potential:

$$\phi_{ij}(\alpha_i, \beta_j, r_{ij}) = \frac{A_{\alpha\beta}}{r_{ij}^m} + \frac{B_{\alpha\beta}}{r_{ij}^n}. \quad (8)$$

Two matrices of parameters are determined:  $A_{\alpha\beta}$  and  $B_{\alpha\beta}$ . The indices  $m$  and  $n$  are predetermined in advance. We consider here the ( $m = 6$ ,  $n = 2$ ) model, which we found more accurate for the reduced representation of protein structure than the atomic Lennard–Jones (LJ) potential.<sup>6</sup> Hence, the index of the vector in eq. (5'),  $\gamma \equiv \alpha\beta$ , runs in our case over the types of contacts, whereas  $n_\gamma$  is the number of contacts of a specific type found in  $\mathbf{X}$ . In case of the LJ model, the “number” includes an additional geometric weight hidden in a continuous “number” function— $n_\gamma \propto 1/r^m$ .

The set of eq. (1) can be rewritten now as follows:

$$\begin{aligned} E(\mathbf{X}_j; \mathbf{z}) - E(\mathbf{X}_n; \mathbf{z}) &= \sum_{\gamma} z_{\gamma} (n_{\gamma}(\mathbf{X}_j) - n_{\gamma}(\mathbf{X}_n)) \\ &= \mathbf{z} \cdot \Delta \mathbf{n}_{j,n} \geq \varepsilon \quad \forall (j, n), \end{aligned} \quad (9)$$

where index  $j$  runs over the misfolded structures of a given protein, and index  $n$  runs over the native structures in the training set. The difference in contacts vector,  $\Delta \mathbf{n}_{j,n}$ , is a result of counting contacts of specific types in both native and misfolded structures. We solve the set of eq. (9) for  $\mathbf{z}$ , without optimizing an objective function. We use the BPMPD program of Cs. Mészáros,<sup>23</sup> which is based on the primal-dual interior point algorithm and allows us to compute a series of analytic centers according to the MaxF procedure. In practice, the right hand sides of the inequalities in (9) are set to be equal to a small positive number,  $\varepsilon = 10^{-6}$ . We also bound the variables,  $-10 \leq z_{\gamma} \leq 10$ , for each  $\gamma$ .

A convenient way to generate a set of misfolded structures is the so-called gapless threading. Consider a set of proteins  $\{(\mathbf{X}_{n_k}, S_{n_k}); k = 1, \dots, N\}$ . Each native sequence  $S_{n_i}$  is fitted without deletions and insertions into the other (longer) structures in the training set,  $\mathbf{X}_{n_j}, n_j \neq n_i$ , which provide alternative (misfolded) packing of the protein chain. Thus, each gapless alignment of a native sequence into an alternative structure provides one misfolded (decoy) structure and the corresponding inequality, as defined in eq. (9).

We use the Hinds and Levitt (HL) set of 246 proteins.<sup>24</sup> Gapless threading of all sequences into all structures generated the set of 4,003,727 inequalities (note that there are many ways a shorter sequence can be aligned to a longer structure), which we will refer to as the HL problem. We also use a subset of 627,567 constraints (referred to as the HLs problem) that result from aligning all the sequences into structures that are less than 33% longer. Thus, many alignments of very short sequences into long structures are excluded from the training set, reducing the size of the problem. Tobi and Elber’s (TE) set of 594 proteins is used as a control set.<sup>4</sup> Gapless threading of all sequences into all structures in the TE set generates about 30 million of inequalities. We use the program LOOPP<sup>25</sup> to generate the inequalities for the LP training.

## Results

In a previous work<sup>6</sup> we addressed the question of the minimal number of parameters that is required to obtain an exact solution for the HL problem. We found that the HL problem proves infeasible when using pairwise potentials with less than 10 types of amino acids (i.e., with less than 55 types of contacts between amino acids). Here, we revisit this problem using the “maximum feasibility” guideline.

We consider two reduced alphabets of amino acids: first of two letters only, namely H and P (for hydrophobic and polar residues, respectively), and the second of four letters, namely H, P, C<sub>+</sub>, and C<sub>-</sub> (C<sub>+</sub> standing for positively charged and C<sub>-</sub> for negatively charged residues, respectively). The assignment of the different amino acids to the letters of the reduced alphabets is in Table 1. The HL and HLs problems are infeasible when formulated in terms of the four-letter alphabet. In other words, even the smaller (HLs) set of inequalities cannot be solved exactly with four types of amino acids, corresponding to 10 types of amino acid contacts.

## Contact Model with Four Types of Amino Acids

We first apply the MaxF rule to the HLs problem in terms of a contact pairwise model, as defined in eq. (7). Four types of amino acids are employed. Results for a number of different starting points are discussed. The first initial guess is the statistical potential derived from the HL set of native structures. Notice that such a potential can always be generated for the problem at hands. Statistical potentials employ contact energies defined as logarithm of the properly normalized probabilities of observing a given type of contact.<sup>8</sup> With a proper choice of the sample of native shapes, the statistical potentials proved to be quite successful in distinguishing native from misfolded structures.<sup>9, 10, 15, 16</sup>

In our case, the statistical potential derived from the HL set of proteins for the four-letter alphabet (see Table 2) performs poorly. It does not satisfy 57,211 inequalities, and fails to recognize 144 proteins (that is for 144 proteins there are decoy structures with energies lower than the native energy). However, the dominant (stabilizing) contributions to the native energies come from the HH interactions. Therefore, one may expect that our initial guess still captures important characteristics of a good solution, with a significant room for improvement. Indeed, as we can see from Table 4, just the first iteration of MaxF procedure dramatically improves the initial solution. The analytic center of the first polytope, defined by all the inequalities satisfied by the initial guess, misses only 6,800 constraints and 22 proteins.

To characterize the shape of the distribution of energy differences in eq. (1),  $\Delta E_{\text{mis, nat}}$ , we compute the so-called Z-score, which is defined as the ratio of the average over the standard deviation of the distribution,  $Z = \langle \Delta E_{\text{mis, nat}} \rangle / \sigma$ . The Z-score of the distribution obtained with the statistical potential is equal to 1.22, and increases to 1.98 after the first iteration of MaxF. Hence, not only the tail of the distribution is corrected, but also the whole distribution is shifted away from the native energies. This is expected because the analytic center provides in general a more uniform distribution of energy differences (distances to the cutting hyperplanes), as compared to an off-centered guess.

We observe further improvement in the subsequent iterations. The converged solution, which we will refer to as 4HLs potential (short for the four-letter potential, trained on the HLs problem), misses only 1928 inequalities and 11 proteins. The inspection of the constraints that are not satisfied reveals that 1922 of them are due to six membrane proteins included in the HL set (1prcC, 1prcL, 1prcM, 4rcrL, 4rcrM, 2por). The remaining six constraints refer to five other proteins that are not recognized (1pp2R, 2bbkB, 2ltnA,

**Table 1.** Definitions of Different Groups of Amino Acids That Are Used in the Present Study.

|   |   |
|---|---|
| Hydrophobic (H, HYD)                      | ALA CYS HIS ILE LEU MET PHE PRO TRP TYR VAL |
| Polar (P, POL)                            | ARG ASN ASP GLN GLY LYS SER THR             |
| Positively Charged (C <sub>+</sub> , CHG) | ARG LYS                                     |
| Negatively Charged (C <sub>-</sub> , CHN) | ASP GLU                                     |

When the charged residues are included explicitly the group of polar residues is reduced correspondingly. In a previous study we found that 10 types of amino acids were necessary to solve exactly the Hinds–Levitt set of proteins by pairwise interaction models.<sup>6</sup> Using the MaxF procedure we find that four types of amino acids are essentially sufficient to recognize all but membrane proteins in the HL set.

Table 2. Parameters for Contact Pairwise Potentials with Four Types of Amino Acids.

|       |       |       |       |       | MaxF |       |       |       |       |
|-------|-------|-------|-------|-------|------|-------|-------|-------|-------|
|       | HYD   | POL   | CHG   | CHN   | HYD  | POL   | CHG   | CHN   |       |
| Init1 |       |       |       |       |      |       |       |       |       |
| HYD   | -0.57 | -0.55 | -0.28 | -0.17 | HYD  | -0.34 | 0.11  | 0.17  | 0.29  |
| POL   | -0.55 | 0.16  | -0.22 | -0.23 | POL  | 0.11  | -0.07 | 0.24  | 0.36  |
| CHG   | -0.28 | -0.22 | 1.01  | -0.97 | CHG  | 0.17  | 0.24  | 1.00  | -0.40 |
| CHN   | -0.17 | -0.23 | -0.97 | 0.82  | CHN  | 0.29  | 0.36  | -0.40 | 0.12  |
| Init2 |       |       |       |       |      |       |       |       |       |
| HYD   | -0.46 | 0.04  | 0.41  | 0.27  | HYD  | -0.45 | 0.08  | 0.37  | 0.35  |
| POL   | 0.04  | 0.03  | 0.13  | 0.09  | POL  | 0.08  | 0.08  | 0.32  | 0.14  |
| CHG   | 0.41  | 0.13  | 0.60  | -0.41 | CHG  | 0.37  | 0.32  | 0.98  | -0.65 |
| CHN   | 0.27  | 0.09  | -0.41 | 0.39  | CHN  | 0.35  | 0.14  | -0.65 | 0.12  |

A statistical potential resulting from the Hinds–Levitt set of proteins (denoted as Init1) and the converged MaxF potential obtained when using Init1 as a starting guess are given in the two upper blocks. A projection of 10-letter potential trained previously (denoted as Init2) and the converged MaxF potential obtained when using Init2 as an initial guess are included in the lower blocks.

2mev3, 3sdpA). Removing the membrane proteins as well as two other proteins (which were not recognized due to the presence of structural relatives) from the training set results in a feasible problem.

The quality of the 4HLs potential is comparable to the previously trained 10-letter potential,<sup>6</sup> despite the fivefold decrease in the size of the parametric space. When 4HLs potential is applied to the full HL problem, 23 proteins and 3652 inequalities are missed (3465 of them due to the membrane proteins). However, when applied to the larger TE set, both potentials recognize correctly the same number of proteins (504 out of 594). Hence, we were forced to use as many as 55 parameters just to solve the full HL problem exactly, without significant improvement in the performance on the TE set. The MaxF procedure effectively reduces the number of parameters by filtering out “hard” constraints due to inherently different protein environments.

Our second initial guess is adopted from the 10-letter potential that solves the HL problem exactly. Because the 10-letter alphabet contains our reduced model, we simply take the relevant  $4 \times 4$  block from the table of energy parameters (see Table 2). Such a “guess” (which is a projection of the exact solution) is expected to perform well and indeed it only misses 3508 inequalities and 18 proteins. The converged solution is very similar to the previously obtained 4HLs potential. The new approximation misses the same set of 11 proteins. A slightly smaller number of constraints is now violated—1865, including 1858 due to the membrane proteins. When applied to the TE set the same 504 proteins are recognized.

We tried several different perturbations of the original statistical potential to further test the convergence of MaxF with different starting points. Physically motivated initial solutions converge to potentials resembling (numerically and in terms of performance) the 4HLs potential. On the other hand, MaxF procedure fails when starting from nonphysical potentials. Inverting the signs of the diagonal elements in Table 2, for example, results in a nonphysical potential that penalizes the HH contacts and misses 625,444 inequalities and 138 proteins. MaxF procedure yields in this case

a potential that is still trapped in the subspace of the parametric space, in which the HH interactions are penalized. The final solution misses 624,466 inequalities and 138 proteins. The Z-score of the initial distribution is negative and remains essentially unchanged during iterations.

We remark that as many as 15 iterations may be needed until the procedure stops, and no further improvement is obtained. However, a nearly converged solution is found already after four to six iterations. We would also like to point out that we do not use a “warm” start at the subsequent iterations, for instance, we do not use the current solution to restart the LP solver in the next iteration. With the promise of better warm start strategies for interior point methods<sup>28</sup> we expect that our problem, with the subsequent iterations being only a perturbation over the previous problem, would be solved in a much smaller number of iterations. Each iteration of MaxF procedure for the HLs problem with 10 parameters (including formulating and solving the problem) takes several minutes on a SUN Ultra Sparc2 machine.

#### Contact Model with Two Types of Amino Acids

Can we further reduce the number of parameters, without deteriorating the quality of the potentials? Motivated by the relative success of the HP model advocated by Dill,<sup>29</sup> we consider only two types of amino acids. In the original Dill potential the interactions of pairs of amino acids other than HH are set to zero,  $\varepsilon_{HP} = \varepsilon_{PP} = 0$ , whereas  $\varepsilon_{HH} = -\lambda$ . The positive parameter  $\lambda$  determines the scale of the energy. For the HL problem, the Dill potential fails to predict the correct fold of 46 out of 246 proteins, violating only 29,129 inequalities. For the larger TE set, the Dill potential recognizes 456 of the 594 proteins. This result is remarkable considering the simplicity of the model.

We applied our procedure to the full HL problem, which contains significantly more constraints than the HLs problem. When starting with the Dill potential as the initial guess, we obtain only a modest improvement. The converged MaxF solution ( $\varepsilon_{HH} = -0.57$ ,  $\varepsilon_{HP} = 0.02$ ,  $\varepsilon_{PP} = 0.05$ ) misses 41 proteins and 22,220

inequalities, including 20,336 inequalities due to the membrane proteins. The Z-score improves only slightly: from 1.91 to 1.94. A minor improvement is also observed for the TE set—472 proteins are recognized. When trying different perturbations of the Dill potential or an HP statistical potential, as the initial guess, we converge to very similar solutions (although the quality of the starting point may be much worse).

Thus, the physically motivated, effective projection of the problem into one-dimensional subspace is close to the best solution in the three-dimensional parametric space (for the sampling of misfolded structures by the gapless threading). Significantly better results are only obtained when the polar residues are further differentiated. This is additionally confirmed by the fact that the reduced HLs problem (that was solved exactly using four types of amino acids) proves infeasible with two types of amino acids.

### Continuous Model with Two Types of Amino Acids

The last example we consider is the continuous model of the LJ(6,2) type, defined in eq. (8). We apply it to the smaller HLs problem, using two types of amino acids only, which corresponds to six energy parameters to be optimized. Despite the fact that we are using the same training set, this is a very different problem now, with real coefficients of the constraint matrix,  $n_\gamma$  [see eq. (9)].

To obtain an initial guess we take advantage of the LJ(6,2) potential in terms of 10-letter alphabet that solved the full HL problem.<sup>6</sup> This potential, with just 110 parameters, was shown to be comparable in performance to the best contact potential with 210 parameters and significantly better than 10-letter contact potential trained on the HL set.<sup>6</sup> Therefore, one might expect that, similar to the contact model, using the projection of such a potential into two-letter alphabet would provide a very good starting point.

The projected potential (see Table 3) performs poorly, however, missing 55,894 inequalities and 59 proteins. The MaxF procedure results in only a minute improvement—the converged MaxF potential is numerically very similar to the initial guess, and it misses

**Table 3.** Parameters for LJ(6,2) Potentials with Two Types of Amino Acids.

| $A_{ij}$ | Init1 |       | $A_{ij}$ | Init2 |      | $A_{ij}$ | MaxF  |       |
|----------|-------|-------|----------|-------|------|----------|-------|-------|
|          | HYD   | POL   |          | HYD   | POL  |          | HYD   | POL   |
| HYD      | 9.32  | 1.45  | HYD      | 1.00  | 0.00 | HYD      | 2.61  | -1.06 |
| POL      | 1.45  | -1.19 | POL      | 0.00  | 0.00 | POL      | -1.06 | -4.26 |
| $B_{ij}$ | HYD   |       | $B_{ij}$ | HYD   |      | $B_{ij}$ | HYD   |       |
|          | HYD   | POL   |          | HYD   | POL  |          | HYD   | POL   |
| HYD      | -2.34 | 0.47  | HYD      | -2.34 | 0.00 | HYD      | -9.99 | 1.94  |
| POL      | 0.47  | 0.01  | POL      | 0.00  | 0.00 | POL      | 1.94  | 0.69  |

A projection of 10-letter LJ(6,2) potential from ref. 6 (denoted as Init1) and its modification with a smoother HH repulsion term and HP, PP interactions set to zero (denoted as Init2), as well as the converged MaxF potential obtained when starting from Init2 are presented. Init1 provides a much worse initial guess, which is not improved significantly by MaxF. Note that the “repulsive” coefficients  $A$  are given first, followed by the “attractive” coefficients  $B$ . The coefficients are expressed in terms of the unit distance of 3 Å.

55 proteins and 50,405 inequalities. Setting parameters for HP and PP interactions to zero, while keeping  $A_{HH}$  and  $B_{HH}$  the same as previously, provides even worse guess that misses 84 proteins and 61,150 constraints. The MaxF procedure again fails to improve it significantly, resulting in a potential that violates 54,067 constraints and does not recognize 81 proteins. MaxF solutions are trapped in the neighborhood of the starting point.

Motivated by the relatively better performance of the contact HP model, we next start from a potential with a much softer repulsion term (denoted as Init2 in Table 3). As can be seen from Table 4, the new guess indeed performs much better. Only 18,985 inequalities and 49 proteins are missed, which is further reduced (after applying MaxF) to 12,362 inequalities (11,822 of them due to the membrane

**Table 4.** Results of the MaxF Procedure for the Design of Reduced Folding Potentials.

| Iteration       | (N_ineq/N_prot/Z-score) |                        |                        |
|-----------------|-------------------------|------------------------|------------------------|
|                 | Contact, 4-lett, Init1  | Contact, 4-lett, Init2 | LJ(6,2), 2-lett, Init2 |
| Initial guess   | 57,211/144/1.22         | 3508/18/1.99           | 18,985/49/1.74         |
| First iteration | 6800/22/1.98            | 3125/14/1.99           | 18,022/43/1.76         |
| Converged MaxF  | 1928/11/2.01            | 1865/11/2.01           | 12,362/27/1.92         |

Gapless threading on the Hinds–Levitt set of 246 proteins<sup>23</sup> is used to generate inequalities for training. An infeasible (in terms of reduced alphabets) set of 627,567 inequalities is used (HLs problem—see the second section). Two types of folding potentials are considered to illustrate how the “maximum feasibility” guideline improves the initial solution, which satisfies certain subset of the constraints. The results for the contact pairwise model and four types of amino acids are presented in the second and third columns, using as the initial guess the statistical potential of Table 2 (Init1) and the projected 10-letter potential of Table 3 (Init2), respectively. The results for the continuous pairwise model of a Lennard–Jones 6-2 type, using as a starting guess a “soft repulsion” potential denoted as Init2 in Table 3, are included in the last column. For each potential the number of inequalities that are not satisfied (N\_ineq), the number of proteins that are not recognized (N\_prot) and the Z score at a given iteration are reported. Note that in each of the cases reported here a significant improvement with respect to the initial guess is achieved.

proteins) and 27 proteins only. The initial Z-score of 1.74 reaches 1.92 upon convergence.

However, when compared to the simple contact potential with two types of amino acids, the continuous pairwise model is not advantageous. Applied to the full HL problem, the LJ(6-2) potential violates 26,583 inequalities, and does not recognize 40 proteins. Applied to the larger TE problem, the new LJ(6,2) potential recognizes 476 out of 594 proteins, that is only four proteins more than the best contact HP potential we obtained and 20 proteins more than the simple Dill potential.

The two types of amino acids enforce a common distance law for side-chain centers of amino acids of very different volume. Pairs of the type Gly–Ala and Arg–Leu, for example, have the same interaction law. The difficulty with obtaining significant improvement using MaxF procedure and two types of amino acids, together with the success of the 10-letter LJ(6,2) potential (which treats explicitly small residues), may suggest the importance of differentiating amino acids of a different size.

## Discussion

The problem of identifying the sources of infeasibility in LP problems (that often come simply from errors in the formulation of the problem) is of significant practical importance, and promotes development of heuristic methods for finding approximations to LFS.<sup>19,20</sup> Probably the most popular method, implemented in some LP packages, is based on the idea of “elastic programming.”<sup>20,30</sup> Instead of solving the original (infeasible) problem one solves a modified problem, with “elastic” variables added first to ensure that the “elastic” problem has a solution and then iteratively removed until infeasibility is reached again.<sup>20</sup>

In principle, such an “elastic filter” could be used as well to remove the “hard” constraints. The MaxF guideline, applied to the resulting feasible subset of inequalities, would provide a partial solution of the problem. Unfortunately, numerical tests suggest that the elastic filter heuristic is not very effective for problems that require a removal of a large number of constraints to obtain a feasible subset.<sup>31</sup> Moreover, using the elastic filter approach implies that an elastic variable is added for each constraint, increasing dramatically the size of the LP problem when solving millions of inequalities. Therefore, such an approach is rather impractical.

Finally, we comment that the examples considered here are much smaller than those required to train folding potentials of sufficient accuracy. Our experience shows that a much more complete sampling of native and misfolded structures, resulting in a huge number of inequalities, is necessary. Such problems are very likely to be infeasible with simple functional models of folding potentials and a limited number of parameters.

However, due to the underlying physical principles, most of the constraints should be satisfied by commonly used statistical potentials. The MaxF procedure provides a simple way to improve further such potentials, both in terms of the number of inequalities that are not satisfied and in terms of the overall shape of the distribution of energy gaps, as defined in eq. (1).

## References

1. Maiorov, V. N.; Crippen, G. M. *J Mol Biol* 1992, 227, 876.
2. Vendruscolo, M.; Domany, E. *J Chem Phys* 1998, 109, 11101.
3. Tobi, D.; Elber, R. *Proteins Struct Funct Genet* 2000, 41, 40.
4. Tobi, D.; Shafran, G.; Linial, N.; Elber, R. *Proteins Struct Funct Genet* 2000, 39, 71.
5. Meller, J.; Elber, R., submitted.
6. Meller, J.; Elber, R., submitted.
7. Akutsu, T.; Tashimo, H. *Proc. Pacific Symposium on Bio-computing*, 1998, p. 413.
8. Sippl, M. J.; Weitckus, S. *Proteins* 1992, 13, 258.
9. Miyazawa, S.; Jernigan, R. L. *J Mol Biol* 1996, 256, 623.
10. Godzik, A.; Kolinski, A.; Skolnick, J. *Proteins Struct Funct Genet* 1996, 4, 363.
11. Khachiyan, L. G. *Doklady Akad Nauk USSR* 1979, 244, 1093.
12. Karmakar, N. K. *Combinatorica* 1984, 4, 373.
13. Ye, Y. *Interior Point Algorithms: Theory and Analysis*; Wiley: New York, 1997.
14. Vanderbei, R. J. *Linear Programming: Foundations and Extensions*; Kluwer Academic Publishers: New York, 1996.
15. Wright, S. J. *Primal-Dual Interior-Point Methods*; SIAM Publications: 1997.
16. den Hertog, D. *Interior Point Approach to Linear, Quadratic, and Convex Programming*; Kluwer Academic Publishers: New York, 1994.
17. Adler, I.; Monteiro, R. D. C. *Math Program* 1991, 50, 29.
18. Monteiro, R. D. C.; Adler, I. *Math Program* 1989, 44, 43.
19. Chakravarti, N. *Eur J Oper Res* 1994, 73, 139.
20. Parker, M.; Ryan, J. *Ann Math Artif Intell* 1996, 17, 107.
21. Chinneck, J. W. *INFORMS J Comput* 1997, 9, 164.
22. Garey, M. R.; Johnson, D. S. *Computers and Intractability: A Guide to the Theory of NP-Completeness*; W.H. Freeman and Company: New York, 1979.
23. Meszaros, C. S. *Comput Math Appl* 1996, 31, 49.
24. Hinds, D. A.; Levitt, M. *J Mol Biol* 1994, 243, 668.
25. Meller, J.; Elber, R. <http://www.tc.cornell.edu/CBIO/loopp>.
26. Liwo, A.; Oldziej, S.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *J Comput Chem* 1997, 18, 849.
27. Xia, Y.; Huang, E. S.; Levitt, M.; Samudrala, R. *J Mol Biol* 2000, 300, 171.
28. Yildirim, E. A.; Wright, S. J. Technical Report 1258, School of Operations Res. and Industrial Eng., Cornell University.
29. Chan, H. S.; Dill, K. A. *Proteins Struct Funct Genet* 1998, 30, 2.
30. Brown, G.; Graves, G. Presented at ORSA/TIMS conference, Las Vegas, 1975.
31. Chinneck, J. W. *Ann Math Artif Intell* 1996, 17, 127.