A Quadratic Programming Formulation for the Design of Reduced Protein Models in Continuous Sequence Space

Sung K. Koh
Department of Mechanical Engineering
and Applied Mechanics,
University of Pennsylvania,
Philadelphia, PA 19104

G. K. Ananthasuresh
Department of Mechanical Engineering
and Applied Mechanics,
University of Pennsylvania,
Philadelphia, PA 19104
and Department of Mechanical Engineering,
Indian Institute of Science,
Bangalore-560 012, India

Christopher Croke
Department of Mathematics,
University of Pennsylvania,
Philadelphia, PA 19104

The notion of optimization is inherent in the design of a sequence of amino acid monomer types in a long heteropolymer chain of a protein that should fold to a desired conformation. Building upon our previous work wherein continuous parametrization and deterministic optimization approach were introduced for protein sequence design, in this paper we present an alternative formulation that leads to a quadratic programming problem in the first stage of a two-stage design procedure. The new quadratic formulation, which uses the linear interpolation of the states of the monomers in Stage I could be solved to identify the globally optimal sequence(s). Furthermore, the global minimum solution of the quadratic programming problem gives a lower bound on the energy for a given conformation in the sequence space. In practice, even a local optimization algorithm often gives sequences with global minimum, as demonstrated in the examples considered in this paper. The solutions of the first stage are then used to provide an appropriate initial guess for the second stage, where a rescaled Gaussian probability distribution function-based interpolation is used to refine the states to their original discrete states. The performance of this method is demonstrated with HP (hydrophobic and polar) lattice models of proteins. The results of this method are compared with the results of exhaustive enumeration as well as our earlier method that uses a graph-spectral method in Stage I. The computational efficiency of the new method is also demonstrated by designing HP models of real proteins. The method outlined in this paper is applicable to very large chains and can be extended to the case of multiple monomer types. [DOI: 10.1115/1.1901705]

1 Introduction and Background

In recent years, engineering research techniques, especially from kinematics and elastic mechanics viewpoints, have been applied to protein studies (e.g., [1–3]). In this paper we apply a design optimization technique to protein design. Proteins are heteropolymer linear chains comprised of 20 different types of amino acid molecules. Each amino acid molecule is composed of carbon (C), hydrogen (H), nitrogen (N) and oxygen (O) atoms, and the side chain molecule (R), as shown in Fig. 1(*a*). The atoms and the side chain are connected by covalent bonds shown by thin solid lines in the figure. The molecules are joined together with peptide bonds (thick solid lines) between C and N atoms and are also often referred to as *residues*. As indicated in Fig. 1(*b*), the residues can rotate relative to each other about the bonds C—N and $C—C^{\alpha}$. Since the 20 different types of side chain molecules (R) possess varied affinity to solvents and among themselves, the protein chain folds into a three-dimensional structure, called the *conformation*, such that the force equilibrium among the hydrophobic and other forces is achieved [4]. The conformation is often represented by the *backbone* that joins the carbon and nitrogen atoms with further detail given by the orientations of the side chain molecules, which are called *rotamer* configurations.

It is known that the functionality of a protein that provides mechanical support, enzymatic catalysis, regulation, signaling, etc., is determined by its three-dimensional conformation. Furthermore, the folded conformation of a protein chain is determined by its sequence of amino acid residues. Therefore, the protein design problem is equivalent to determining the amino acid sequence for a desired conformation that has the intended functionality. In this paper we treat protein sequence design as a quadratic programming problem and solve it using deterministic optimization methods as opposed to stochastic and statistical approaches reported in the literature.

The notion of optimization is inherent in the design of proteins because the predominantly used criteria for protein folding relate to a *minimum energy* or *maximum energy gap*. The energy of a protein chain in a folded conformation is the sum of energetic interactions acting between every pair of interacting neighboring residues. These are modeled with varying degrees of complexity from one extreme of considering atomistic details to the other extreme of highly simplified experimental or empirical data. For computational and design purposes, it is useful to have a simple model of the energy potential defined among different types of neighboring amino acid residues. An example of such a model for the pairwise inter-residue interactions is the MJ (Miyazawa and Jernigan) matrix of size 20×20, which is based on statistical analysis of known proteins [5]. Every entry in this matrix gives the level of energy between the corresponding pairs of types of amino acid monomer molecules.

In order to understand the principles of protein folding, it is useful to imagine a *conformation space* and a *sequence space*, which need to be searched to design a protein. The conformation space consists of all possible conformations (foldable states in three dimensions) for a protein chain of a given number of residues. The sequence space, on the other hand, contains all possible sequences for a chain of given length, i.e., the number of residues. According to Anfinsen's [6] thermodynamic hypothesis, a protein chain stably folds to its preferred conformation if the energy of the sequence is uniquely and globally the minimum among all its

Contributed by the Mechanisms and Robotics Committee for publication in the JOURNAL OF MECHANICAL DESIGN. Manuscript receieved: November 1, 2004; revised February 25, 2005. Associate Editor: C. Mavroidis.

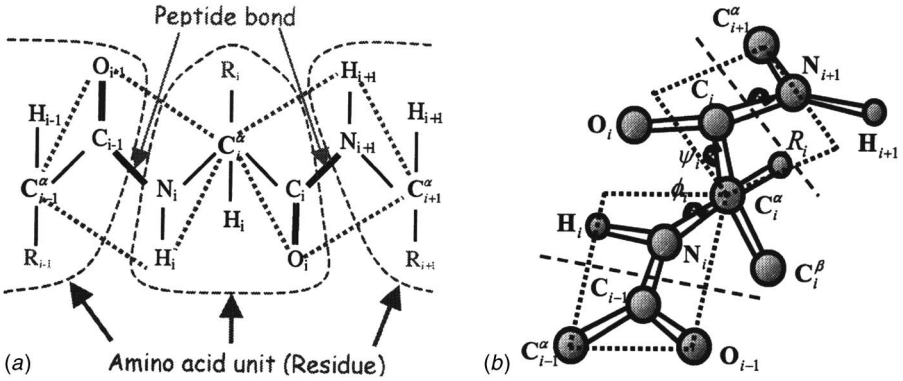

Fig. 1 Nomenclature of proteins. (*a*) Schematic of three residues in a protein chain, (*b*) ball and stick model showing the 3-D arrangement.

permissible conformations. Such a conformation is said to be the *native state* for the sequence.

Designing a protein requires a search both in the sequence and conformation space. Identifying the sequence, such that the designed protein chain folds to a desired conformation, requires a search in the conformation space that consists of an infinite number of conformations. In addition, the search of the conformation space needs to be carried out for every sequence in the sequence space to identify the sequences in the native state. A full search in both the sequence and the conformation space is computationally impractical. As a practical approach for a protein design, alternative approaches have been developed by many researchers to make the problem computationally tractable [13,22]. The predominant approach is to limit the search to the sequence space alone in order to identify subsets that satisfy certain criteria. A number of design criteria have been proposed such as energy, energy gap, the ratio of energy gap to the standard deviation of the nonfolded state energies, and the free energy separation between the target state and an unfolded ensemble [7,8]. In this paper, the sequence space is searched to identify a sequence that has minimum energy for the desired conformation among all sequences satisfying an additional constraint on the composition. This should be a useful component of the overall protein design problem.

The complexity of the sequence-design problem can easily be grasped by noting that there are 20^N sequences when there are N residues in a chain. For $N=100$, which is the order of the size for real proteins of medium size, the number of sequences is $1.27E130$—an inconceivably large number that makes solution by enumeration impossible and poses significant challenges for optimization methods.

Many techniques have been developed for protein sequence design. They include deterministic and stochastic methods as reviewed briefly in [7]. It is worth noting that all these techniques [9–13] approach it as a discrete problem with two exceptions: one that uses statistical design methods [7,14] and the other that uses mixed IP–LP (integer programming and linear programming) software [15]. As an alternative, we presented a continuous modeling of this problem [16] and reduced the problem size using a graph spectral method [17] so that sequence(s) with minimum energy could be found within minutes on a desktop computer using gradient-based continuous optimization methods. In particular, we showed that a minimum energy sequence for a 500-residue long chain of two types of monomers (i.e., $2^{500}\approx 10^{150}$ possible sequences) could be found within 10 min on a single-processor desktop computer. The global minimality of the solutions obtained with this method was confirmed for the case of 27 residues by exhaustive enumeration. However, since a local optimization algorithm is used, global optimality is not always assured and is dependent upon the given initial guess.

In this paper, we focus on the estimation of a good initial guess to our previous method by proposing and solving a new quadratic programming (QP) problem in the first stage of a two-stage design procedure. The global minimum solution of the QP problem can be found by the deterministic global optimization method presented by Lo et al. [18–20] on the basis of the algorithms developed by Tuy [21]. However, in this paper, we use a local optimization method to solve the QP problem since it performs fairly well in identifying a sequence with global minimum energy as observed with enumerated sequences for small chains. Thus, the new method consists of two stages: the QP problem is solved in Stage I, and the sequence obtained in Stage I is refined by the method proposed by Koh et al. [16] in Stage II. We compare the results of the new method with our old method [16] to show the improvement in the cases of large numbers of residues that are impractical to verify by enumeration.

In order to have a computationally tractable problem whose results can be verified independently by enumeration, simplified representations of proteins, called HP lattice models are often considered. This is based on grouping the 20 amino acid types into different categories based on their properties. The simplest, and the most important form of such a classification is based on hydrophobicity, i.e., the inclination to hide from the water molecules. Hydrophobic amino acids have greasy side chains and they like to stick together in order to minimize their contact with water, whereas polar amino acids have more affinity to water [22]. This leads to the so-called HP models of proteins where H groups are hydrophobic amino acid types and P the polar (or hydrophilic) types. This reduced representation enables, for simple cases, exhaustive enumeration and thus a way of confirming the validity of a folding criterion, minimal property of a sequence, or a new method of protein design.

To further simplify the problem with respect to conformation space, the locations of the residues are fixed in a lattice. Figure 2

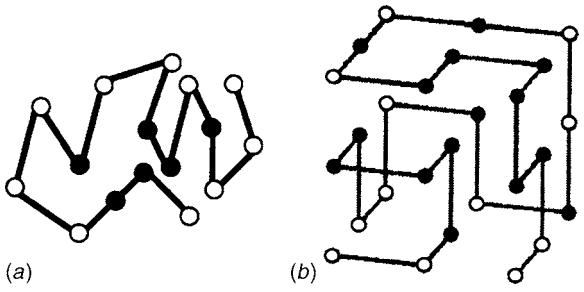

Fig. 2 (a) H-P models of proteins (*a*) a 2-D protein model on an irregular lattice, (*b*) a 3-D protein on a 3×3×3 regular lattice. Black dots represent hydrophobic (H) residues and white dot polar (P) residues.

shows an HP model of 2-D and 3-D proteins. In these, the positions of amino acids in the lattice (irregular as in Fig. 2(*a*) or regular as in Fig. 2(*b*)) are confined to particular locations in the 2-D or 3-D space. The allowed conformations are the self-avoiding compact chains in which the chain cannot visit a single site more than once [23]. The interacting pairs of residues in such an arrangement are identified as those that are not nearest neighbors in the chain, but are in space. It has been shown that HP lattice models possess principal energetic and thermodynamic properties of real proteins [24]. We use HP lattice models to validate our design algorithm in this paper.

The remainder of the paper is organized as follows. The new quadratic programming problem (QP) formulation is explained in Sec. 2 along with a discussion of its principal features and some results that demonstrate its computational efficiency. The QP problem is the first of a two-stage design procedure. The new combined method that uses the rescaled Gaussian distribution function based continuous interpolation is presented in Sec. 3. The results of the new method are presented and discussed in Sec. 4. The paper ends with concluding remarks in Sec. 5.

2 Quadratic Formulation For Protein Sequence Design

In HP modeling, for computational purposes, it is convenient to denote the H type of residue by one and the P type by zero. This 1–0 state representation naturally leads to a discrete sequence space. To make it amenable for continuous optimization methods, we interpolate the states 0 and 1 linearly to render the sequence design as a quadratic programming (QP) problem as shown below. To see how the total energy is written for this case, consider two interacting residue sites i and j, and two different types of monomers H and P. We assign two variables, whose range is [0,1], to each site so that they represent the states of the H and P monomer types and interpolate the states linearly between 0 and 1. As in topology optimization of structures and compliant mechanisms [25], a value of 0 indicates that the corresponding monomer type does not exist at that site and a value of 1 indicates that the site is completely occupied by that monomer type. In order to see how the energy can be continuously interpolated using this scheme, let x_1 and x_2 denote the H states at the sites i and j, and x_3 and x_4 the P states, respectively. Then the energy $E_{Q_{ij}}$ between them can be written in continuous form as

$$E_{Q_{ij}} = e(H,H)x_1x_2 + e(H,P)x_1x_4 + e(P,H)x_3x_2 + e(P,P)x_3x_4 \tag{1}$$

where $e(H,H)=-2.3$, $e(H,P)=e(P,H)=-1$, $e(P,P)=0$ as per the normalized values extracted from the eigenanalysis of the MJ matrix for HP models [26]. Since x_1 and x_3 indicate the H and P states of the same residue site i, a constraint is imposed to make sure that each monomer site is occupied by one and only one total monomer. That means the two variables quantifying the H and P states sum to unity. The same is true for x_2 and x_4. Thus, we have the following quadratic programming problem in four variables with two linear constraints if we want to minimize $E_{Q_{ij}}$ by choosing appropriate states for the two sites i and j.

$$\begin{aligned}
&\text{Minimize } E_{Q_{ij}}(\mathbf{x}) = e(H,H)x_1x_2 + e(H,P)x_1x_4 + e(P,H)x_3x_2 \\
&\text{w.r.t. } \mathbf{x} = \{x_1 \quad x_2 \quad x_3 \quad x_4\}^T \\
&\text{Subject to} \\
&\qquad x_1 + x_3 = 1 \\
&\qquad x_2 + x_4 = 1 \\
&\qquad 0 \leqslant x_k \leqslant 1 \text{ for } k = 1,2,3,4
\end{aligned} \tag{2}$$

The advantage of the above formulation is that any number of states (and hence many more types of monomers than the simplest two categories of H and P) can be introduced while retaining the form of the quadratic programming problem. This is shown below for the general case of m monomer types or states.

For any desired conformation for which the interacting pairs of residues are known, an adjacency matrix $\mathbf{A}$ can be constructed so that the entry A_{ij} in $\mathbf{A}$ is equal to one if residue sites i and j interact and zero otherwise. The matrix $\mathbf{A}$ is uniquely associated with a conformation in the lattice models. By letting α_i denote the ith monomer type, the energy of a sequence in the given conformation can be written as follows.

$$E_Q = \frac{1}{2}\left[\sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij}\left\{\sum_{k=1}^{m}\sum_{l=1}^{m} e(\alpha_k,\alpha_l)x_{(k-1)N+i}x_{(l-1)N+j}\right\}\right] \tag{3}$$

where $e(\alpha_k,\alpha_l)$ are the energy values in the MJ-type interaction matrix for m monomer types, and $x_{(k-1)N+i}$ is the state of type α_k monomer at site i. Thus, there will be Nm variables in the problem with each set of m variables at N residue sites interpolating m residue states linearly. There is a factor of half in Eq. (3) since the interaction between every pair of interacting monomers is counted twice in the summation. As there are m different types of monomers, each α_i indicates one of the elements in the set of amino acid types. The MJ matrix gives the energy values for every pair of amino acid types.

In some protein sequence design problems, monomer composition constraints may also be imposed. That is, if there should only be N_{α_k} residues of amino acid type α_k in the chain. It can be written as a linear constraint as follows.

$$\sum_{i=1}^{N} x_{(k-1)N+i} = N_{\alpha_k} \tag{4}$$

The above constraint is similar to the material-resource constraint in structural topology optimization [25]. By combining all of the above, the general quadratic programming (QP) problem can now be written as

$$\begin{aligned}
&\text{Minimize } E_Q = \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} \\
&\text{w.r.t. } \mathbf{x} = \{x_1, x_2, \ldots, x_{Nm}\} \\
&\text{Subject to} \\
&\qquad \sum_{k=1}^{m} x_{(k-1)N+l} = 1 \quad \text{for } l = 1,2,\ldots,N \\
&\qquad \sum_{l=1}^{N} x_{(k-1)N+l} = N_{\alpha_k} \quad \text{for } k = 1,2,\ldots,m \\
&\qquad 0 \leqslant x_j \leqslant 1 \quad \text{for } j = 1,2,\ldots,Nm
\end{aligned} \tag{5}$$

where $\mathbf{Q}$ is constructed as per Eq. (3). The above optimization problem can be expressed in the standard QP form as follows.

$$\begin{aligned}
&\text{Minimize } E_Q = \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} \\
&\text{w.r.t. } \mathbf{x} = \{x_1, x_2, \ldots, x_{Nm}\} \\
&\text{Subject to} \\
&\qquad \mathbf{B}\mathbf{x} = \mathbf{c} \text{ and } 0 \leqslant x_i \leqslant 1,\ i = 1,2,\ldots,N
\end{aligned} \tag{6}$$

where the constraints, which are all linear, are expressed as a single matrix equation. The above QP problem can be solved using efficient local gradient-based techniques available for this class of problems [27].

When the objective function is a polynomial, as E_Q in Eq. (6) is, the global minimum of the polynomial function can be identi-

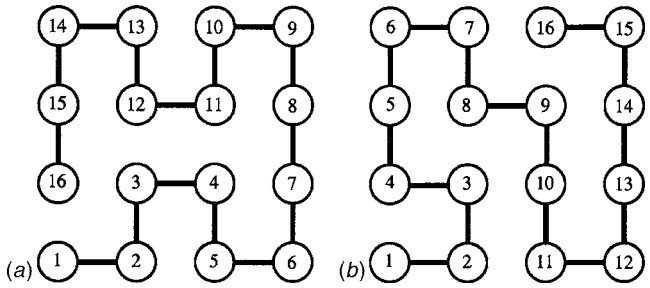

Fig. 3 Two 4×4 lattice conformations. (*a*) Conformation A: number of sequences in native state=1022; (b) Conformation B: Number of sequences in native state=459.

fied by a deterministic global optimization method [18–21]. Although this global optimization method is applicable for QP in Eq. (6), its implementation is cumbersome. Furthermore, this method becomes costly in both time and memory as N becomes large. Therefore, we tested the QP formulation in this work using a gradient-based local optimization method in order to examine its performance with local optimization methods, which are very efficient in terms of using computational resources. Despite using a local optimization method that may only identify a local minimum, the solutions presented later in the paper show excellent performance; the energy level of the sequences from QP corresponds to the global minimum of E_Q in most cases, as presented next.

2.1 Results of the QP Formulation. Two different 4×4 lattice conformations shown in Figs. 3(*a*) and 3(*b*) are considered to show that the local method used to solve the QP problem can give globally minimum results that are obtained by exhaustive enumeration. The small size of the lattice is considered because identifying the global minimum by exhaustive enumeration is not possible otherwise. The conformation A in Fig. 3(*a*) has 1022 sequences for which the energy in this conformation corresponds to the unique global minimum in the conformation space. The conformation with the largest number of sequences in the native state is called the *most designable* conformation [28]. By exhaustive enumeration, we found that the conformation in Fig. 3(*a*) is the most designable 4×4 lattice conformation. The confirmation B in Fig. 3(*b*), on the other hand, has 459 sequences in the native state.

For numerical optimization, we used the built-in optimization routine FMINCON or QUADPROG which are in the Optimization Toolbox of MATLAB software [29]. Since these are local optimization methods, we tried to identify the best solution with at least five random initial guesses. In order to examine the performance of the quadratic energy model, we compared the minimum energy $E_{Q\,\min}$ obtained from the numerical solution of the QP with the minimum energy $E_{\min}$ that is obtained with exhaustive enumeration in the discrete sequence space. Although $E_{Q\,\min}$ of the QP solution can possibly be greater than the $E_{\min}$ or the global minimum of quadratic energy function $E_{Qg\,\min}$, it turned out that, for most of the cases, $E_{Q\,\min}$ is lower than or equal to the minimum energy $E_{\min}$, as shown in Table 1. The results tabulated in Table 1 are for different numbers of H monomers imposed as a composition constraint. Table 1 shows that, in most of the cases, $E_{Q_{\min}}$ is lower than or equal to the minimum energy $E_{\min}$ in discrete sequence space as shown in Table 1. This observation indicates that the quadratic energy model solved with a local method performs very well to identify the sequence with $E_{\min}$, which is the global energy minimum in the discrete sequence space.

The success of the local optimization method in finding the global minimum of the QP problem can be attributed to the way the discrete problem is modeled in the continuous space, and the small size of the problem. The reason for $E_{Q\,\min}$ sometimes being lower than $E_{\min}$ is due to the fact that the minimizing sequence of the QP problem may not necessarily have 0 or 1 states at all residue sites; some may be in the intermediate state. But we need to convert intermediate states to discrete states in order to obtain the solution of the original discrete problem. This needs further examination, which is discussed next.

2.2 Analysis of the QP Results. In order to analyze the result of pushing the intermediate-state residue sites to 0 or 1, we denote the sequence for $E_{Q_{\min}}$ as $\mathbf{x}_0$, and define a set of sequences $\{\mathbf{x}_k\}$ in which the monomers with intermediate states (i.e., not 0 or 1) in $\mathbf{x}_0$ are replaced with either 1 (H) or 0 (P) monomers while satisfying the constraints in Eqs. (5):

$$\{\mathbf{x}_k\}=\{\{\beta_1,\ldots,\beta_{mN}\}|\beta_i=x_i\in\mathbf{x}_0 \text{ if } x_i\in\{0,1\};\quad \beta_i\in\{0,1\} \text{ if } 0<x_i<1,\tag{7}$$

satisfying $\Sigma_{i=1}^{m}\beta_{(i-1)N+l}=1$ for $l=1,\ldots,N$; $\Sigma_{i=1}^{N}\beta_{(j-1)N+i}=N_{\alpha_j}$ for $j=1,\ldots,m$. We now define the sequence space $\mathbf{X}_{\min}$ as the space convexly spanned by $\mathbf{x}_0$ and the $\mathbf{x}_k$:

$$\mathbf{X}_{\min}=\mathbf{x}_0+\sum_{k=0}^{n_{\mathbf{x}}}w_k(\mathbf{x}_k-\mathbf{x}_0)\tag{8}$$

where $w_k\geqslant0$, $\Sigma_{k=1}^{n_{\mathbf{x}}}w_k\leqslant1$. Note that the definition of the $\mathbf{x}_k$ guarantees that $\mathbf{x}_0$ is an interior point of $\mathbf{X}_{\min}$. Due to the quadratic nature of E_Q, a local minimum solution $\mathbf{x}_0$ is also the global minimum in $\mathbf{X}_{\min}$. One can see this by looking at any line $X(t)$ in $\mathbf{X}_{\min}$ with $X(0)=\mathbf{x}_0$ which will lie in the feasible domain for all small t (e.g. $-\varepsilon<t<\varepsilon$). The energy function $E_Q(X(t))=X^T(t)\mathbf{Q}X(t)$ is a quadratic polynomial in t with a local minimum at $t=0$. It means that the energy has a global minimum along $X(t)$ at $t=0$. Since every point in $\mathbf{X}_{\min}$ lies along such a line, $\mathbf{x}_0$ is a global energy minimum in $\mathbf{X}_{\min}$.

Next, note that there are six possible states of $E_{Q\min}$ relative to $E_{\min}$ and $E_{Qg\min}$:

$$E_{Q\min}=E_{\min}=E_{Qg\min}\tag{9a}$$

Table 1 Minimum energy of the conformations A and B of Fig. 3. N_H denotes the number of H monomers. $E_{\min}$ is the minimum energy in discrete sequence space. $E_{Q\,\min}$ is the minimum energy obtained from the local solution of the QP.

N_H	Conformation A		Conformation B		N_H	Conformation A		Conformation B	
	$E_{\min}$	$E_{Q_{\min}}$	$E_{\min}$	$E_{Q_{\min}}$		$E_{\min}$	$E_{Q_{\min}}$	$E_{\min}$	$E_{Q_{\min}}$
1	−2	−2.0750	−2	−2.0750	8	−14.8	−14.8000	−14.8	−14.8000
2	−4.3	−4.3	−4.3	−4.3000	9	−16.1	−16.1000	−16.1	−16.1000
3	−6.6	−6.6000	−6.6	−6.6000	10	−17.1	−17.1750	−17.1	−17.1750
4	−8.6	−8.6750	−8.6	−8.6750	11	−18.4	−18.4000	−18.4	−18.4000
5	−10.9	−10.9000	−10.9	−10.9000	12	−19.4	−19.4750	−19.4	−19.4750
6	−12.2	−12.2000	−12.2	−12.2000	13	−20.7	−20.7000	−20.7	−20.7000
7	−13.5	−13.5000	−13.5	−13.5000	14	−20.7	−20.7000	−20.7	−20.7000

Table 2 Energy minimizing sequences using exhaustive enumeration and solution of the QP problem for conformations A and B shown in Fig. 3

Method	Energy-minimizing sequence																Energy
	(i) Conformation A, $N_H=5$																
Enumerated	P	P	H	H	P	P	P	P	P	P	H	H	P	P	P	H	−10.9
QP solution																	
x_H	0	0	1	1	0	0	0	0	0	0	1	1	0	0	0	1	−10.9
x_P	1	1	0	0	1	1	1	1	1	1	0	0	1	1	1	0	
	(ii) Conformation B, $N_H=6$																
	P	P	H	P	**H**	P	P	H	H	H	P	P	P	P	P	H	−12.2
	P	P	H	P	P	P	**H**	H	H	H	P	P	P	P	P	H	−12.2
Enumerated	P	P	H	P	P	P	P	H	H	H	P	P	P	**H**	P	H	−12.2
	P	P	H	P	P	P	P	H	H	H	P	P	**H**	P	P	H	−12.2
QP solution																	
x_H	0	0	1	0	$\frac{1}{4}$	0	$\frac{1}{4}$	1	1	1	0	0	$\frac{1}{4}$	$\frac{1}{4}$	0	1	−12.2
x_P	1	1	0	1	$\frac{3}{4}$	1	$\frac{3}{4}$	0	0	0	1	1	$\frac{3}{4}$	$\frac{3}{4}$	1	0	
	(iii) Conformation A, $N_H=4$																
	P	P	H	H	P	P	P	P	P	P	H	P	P	P	P	H	−8.6
	P	P	H	H	P	P	P	P	P	P	H	H	P	P	P	P	−8.6
Enumerated	P	P	H	P	P	P	P	P	P	P	**H**	H	P	P	P	H	−8.6
	P	P	H	**H**	P	P	P	P	P	P	P	H	P	P	P	H	−8.6
QP solution																	
x_H	0	0	1	$\frac{1}{2}$	0	0	0	0	0	0	$\frac{1}{2}$	1	0	0	0	1	−8.75
x_P	1	1	0	$\frac{1}{2}$	1	1	1	1	1	1	$\frac{1}{2}$	0	1	1	1	0	
	(iv) Conformation B, $N_H=4$																
	P	P	H	P	P	P	P	P	H	H	P	P	P	P	P	H	
	P	P	H	P	P	P	P	H	H	P	P	P	P	P	P	H	
Enumerated	P	P	H	P	P	P	P	H	**H**	H	P	P	P	P	P	P	−8.6
	P	P	H	P	P	P	P	H	H	H	P	P	P	P	P	**H**	
QP solution																	
x_H	0	0	1	0	0	0	0	1	$\frac{1}{2}$	1	0	0	0	0	0	$\frac{1}{2}$	−8.75
x_P	1	1	0	1	1	1	1	0	$\frac{1}{2}$	0	1	1	1	1	1	$\frac{1}{2}$	

$$E_{Q\min} = E_{\min} > E_{Qg\min} \tag{9b}$$

$$E_{Q\min} > E_{\min} > E_{Qg\min} \tag{9c}$$

$$E_{\min} > E_{Q\min} = E_{Qg\min} \tag{9d}$$

$$E_{\min} = E_{Q\min} > E_{Qg\min} \tag{9e}$$

$$E_{Q\min} > E_{\min} > E_{Qg\min} \tag{9f}$$

Although there are six possible cases for $E_{Q\min}, E_{\min}$, and $E_{Qg\min}$, the example cases above show only two cases in which $E_{\min} = E_{Q\min}$ and $E_{\min} > E_{Q\min}$ as can be seen in Table 1. Using the notation introduced above, consider the case of $E_{Q_{\min}} = E_Q(\mathbf{x}_k)$ for some values of k, and the case of $E_{Q_{\min}} < E_Q(\mathbf{x}_k)$ for other values of k. If $E_{Q_{\min}} = E_Q(\mathbf{x}_k)$ for some $\mathbf{x}_k$ say for $k=1,\ldots,n_q$, and $n_q \leqslant n_{\mathbf{x}}$, then E_Q is constant in the subspace affinely spanned by $\mathbf{x}_0$ and $\mathbf{x}_1$ through $\mathbf{x}_{n_q}$. In order to prove this, let us consider a line $X_1(t)$ between the $\mathbf{x}_1$ and $\mathbf{x}_0$, $X_1(t)=(1-t)\mathbf{x}_0+t\mathbf{x}_1$, which lies in the feasible domain for $-\varepsilon < t \leqslant 1$. Then the energy function $E_Q(X_1(t))=X_1^T\mathbf{Q}X_1$ is a quadratic polynomial in t which can be expressed as $E_Q(t)=at^2+bt+c$ where $a,b,c \in \Re \subset (-\infty,\infty)$. Since $\mathbf{x}_0$ is a local minimum of E_Q, a must be greater than or equal to 0. If $a>0$ then $\mathbf{x}_0$ is a unique global minimum over all values of t. But, since $E_Q(\mathbf{x}_0)=E(\mathbf{x}_1)$, we can conclude that $a=0$. Noting that the only linear functions $E_Q(t)=bt+c$ that have a local minimum are the constant functions, we can also conclude that $E_Q(t)$ is constant on $X_1(t)$. As the energy E_Q along the line $X_1(t)$ between $\mathbf{x}_0$ and $\mathbf{x}_1$ is constant, and is the same as the minimum value $E_{Q_{\min}}$, the energy E_Q on the lines connecting any points on $X_1(t)$ and $\mathbf{x}_2$ is also constant, and by the same argument as above, the minimum value E_Q is constant in the plane affinley spanned by $\mathbf{x}_0, \mathbf{x}_1$, and $\mathbf{x}_2$. Therefore, continuing this process, the energy E_Q in the space affinely spanned by $\mathbf{x}_0$ and the $\mathbf{x}_k$ for $k=1,\ldots,n_q$ is constant.

If $E_{Q_{\min}} < E_Q(\mathbf{x}_k)$ for an $\mathbf{x}_k$, the sequence $\mathbf{x}_0$ corresponding to $E_{Q_{\min}}$ is the unique global energy minimum on the line connecting $\mathbf{x}_0$ and $\mathbf{x}_k$ in $\mathbf{X}_{\min}$. This can also be understood by considering a line $X_k(t)=(1-t)\mathbf{x}_0+t\mathbf{x}_k$ between $\mathbf{x}_0$ and $\mathbf{x}_k$ in $\mathbf{X}_{\min}$. Since E_Q is quadratic in t, and $\mathbf{x}_0$ is a local minimum of E_Q on $X_k(t), E_Q(t)$ is convex on the lines $X_k(t)$. Therefore, $E_{Q_{\min}}$ is the unique global minimum on $X_k(t)$. In the complement of $\{X_k(t)\}$ in $\mathbf{X}_{\min}$, the quadratic nature of E_Q guarantees that E_Q is greater than or equal to $E_{Q_{\min}}$. In the case $E_{Q_{\min}} < E_Q(\mathbf{x}_k)$ for all $\mathbf{x}_k, E_Q$ can still be the same as $E_{Q_{\min}}$ on a subspace that does not go through a discrete state.

The above analysis shows that the minimum found by the local optimization method using the continuous model will have a value that is equal to or lower than the complete discrete sequences (i.e., $\mathbf{x}_k$) constructed with it. This is corroborated by the results shown in Table 2.

Case (i) in Table 2 has only one energy minimum found by enumeration. The QP solution found the same without any intermediate densities. On the other hand, for case (ii), there are four energy-minimizing sequences and the state $\mathbf{x}_0$ of $E_{Q_{\min}}$ is located in the space spanned by the sequence states corresponding to those four sequences. In this case, $\mathbf{x}_0$ has four residue sites (numbered 5, 7, 13, and 14) in the intermediate state of valve equal to 0.25. It means that anyone of these four sites can be made to be 1 (H) so as to satisfy the constraint $N_H=6$ as shown in boldface letters in the sequences in the table. In these two cases, $E_{\min} = E_{Q\min}$.

In cases (iii) and (iv) shown in Table 2, $E_{\min} > E_{Q\min}$. In these examples, the sequence $\mathbf{x}_0$ corresponding to $E_{Q_{\min}}$ is a global energy minimum on $\mathbf{X}_{\min}$ and a unique global energy minimum on the lines connecting $\mathbf{x}_0$ and $\mathbf{x}_k$ in $\mathbf{X}_{\min}$ since $E_{Q_{\min}} < E_Q(\mathbf{x}_k)$ at all

$\mathbf{x}_k$. For the case of the conformations A and B in cases (iii) and (iv), the last two sequences in the table under the heading of enumerated solution consist of the set of $\mathbf{x}_k$. Since there are two sites in $\mathbf{x}_0$ with an intermediate density of $\frac{1}{2}$, two possibilities clearly exist for the fourth residue site. But as can be seen in Table 2, there are also other possibilities with the same minimum energy. This is due to the inability of the local optimization algorithm in identifying all minima.

In this simple case of 4×4 lattice, the analysis of QP solutions could be carried out in the space constructed by their corresponding discrete solutions, and the validity of the minimum could be established. In large problems, when there are many residue sites in an intermediate state, many combinations will once again arise in identifying the discrete energy-minimizing sequences. As stated earlier, this solution of the QP problem can then be used as an initial guess for the Stage II problem to provide a sequence with minimum energy in discrete states. This is described in the next section.

3 Peak Function-Based Energy Modeling for State II Optimization

In our earlier work, we have introduced a continuous modeling of the 1–0 binary states using an interpolating function based on rescaled Gaussian distribution functions for "material distribution" in the topology optimization of compliant mechanism design [30,31] and for HP models of proteins in [16] as shown below.

$$S_i(\rho) = \exp\left[-\left(\frac{\rho_i}{\sigma}\right)^2\right] \tag{10}$$

where ρ is the controlling variable for site i, $S_i(\rho)$ the state of the H monomer at site i, and $\{1-S_i(\rho)\}$ is the state of the P monomer at site i. The symbol σ denotes the parameter that governs the sharpness of $S_i(\rho)$. The variable ρ associated with a site continuously interpolates its state between 0 and 1, as can be imagined from the Gaussian probability distribution function, which is rescaled such that its maximum is 1 rather than its integral and termed as a peak function in [30]. For relatively large values of σ, the state is diffuse between 0 and 1 (i.e., between P and H) for a range of values of ρ. When σ is decreased, the definition of the two states becomes sharper, and eventually when $\sigma\to0$, the state function in Eq. (10) approaches the discrete 0–1 modeling of the state. It should also be noted that for any value of σ, $\rho=0$ precisely describes the H state, and a sufficiently large value of $|\rho|$ describes the P state. The state interpolation by this peak function has several advantages, which justify using it to determine a specific sequence. First of all, there are no bounds for the domain of the design space. The domain of independent variable ρ in this model is defined to be $-\infty<\rho<\infty$ so that one does not have to be concerned with the bounds on the variables. Second, the sharpness of peak function can be varied by the tuning parameter σ and thereby controlling the selection of the states.

In Ref. [16], we presented results that showed that HP models of very large proteins (i.e., number of residues as large as 513) could be handled efficiently. However, the obtained minimum is not guaranteed to be a global minimum if the unbiased initial guess to the optimization is taken to be the one in which all residues are uniformly in the same intermediate state to satisfy a composition constraint (i.e., the number of H monomers in the chain is equal to N_H) as is the common practice in structural topology optimization. On the other hand, if the initial guess were to be close to the global minimum, the method is most likely to converge to it. As noted earlier, our previous method used a result from a spectral graph theory based method [17]. It certainly improved the method's ability but did not have any guarantee of being close to a global minimum. This study is focused on the estimation of an initial guess located near the global minimum. In an effort to estimate a good initial guess for the optimization based on a peak function in Eq. (10), we propose to solve the QP problem in the first stage, which often provides a sequence near the global energy minimum. Furthermore, the new formulation is amenable for deterministic global optimization. The initial guess from QP improves the results for the optimization in the second stage using a peak function-based energy model.

3.1 Energy Minimization in the Second Stage. In this paper, the optimization problem in the second stage is limited to only two monomer types, H and P. Its mathematical form in terms of the Gaussian distribution function-based state interpolation is as follows.

$$\begin{aligned}&\text{Minimize } E=\frac{1}{2}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}A_{ij}e(S_i(\rho_i),S_j(\rho_j))\\ &\text{w.r.t. } \{\rho_1,\rho_2,\ldots,\rho_N\}\\ &\text{and subject to } \sum_{i=1}^{N}S_i-N_H=0\end{aligned} \tag{11}$$

where $e(S_i,S_j)=0.5\{e_{HP}S_i(1-S_j)+e_{HP}(1-S_i)S_j+e_{HH}S_iS_j\}$ and $S_i=e^{-(\rho_i/\sigma)^2}$. As noted earlier, the value of σ is gradually reduced during the iterative optimization process. A sufficiently small value of σ at the time of convergence ensures that all the monomers are in either the H or P state. Thus, the minimum of the discrete problem can usually be identified efficiently by solving problems in Eq. (5) and Eq. (11) sequentially. The details of the second-stage optimization to solve the problem in Eq. (11) are in Ref. [16]. Results of the new two-stage procedure are presented next.

4 Results and Discussion

The performance of the method and the validity of its results are demonstrated in two different ways in this section. First, we compare the sequence and its energy obtained from quadratic energy minimization with those obtained by sequence enumeration. Of necessity, only small proteins are considered for this purpose because larger sizes preclude verification by enumeration. Some lattice models are also examined and their results are compared with our previous method presented in [16]. Later, large real protein-based models are considered and improvement over the previous method is shown.

Figure 4(*a*) shows the ribbon representation of the folded conformation of Src tyrosine kinase transformation protein (PDB code [31], 1 SRL). An HP model was constructed with this protein with 55 H residues out of the total 64 residues that it has. Its sequence space consists of $2^{64}\cong1.8E19$ sequences. For the new two-stage method, it is only an optimization problem in 64 variables and hence it is easily manageable. But for verification by enumeration, it is impractical. However, for the assumed cases of small or large numbers of H residues, the number of sequences will be small. When there are N_H of H residues out of N, the number of sequences is given by

$$\text{Number of sequences with } N_H \text{ residues} = \frac{N!}{N_H!\,(N-N_H)!} \tag{12}$$

Thus, for very small or very large value of N_H, enumeration is possible. As shown with black dots in Fig. 4(*b*), the minimum energy sequences were found by exhaustive enumeration for N_H values of up to 6 and then from 51 to 55. As the circles in this figure show, the new two-stage method (noted as 2ST in the legend) is verified to find the sequences with globally minimum energy. The trend of the minimum energy in between the verified N_H compositions indicates that it is very likely that it is a global minimum even for the cases in between.

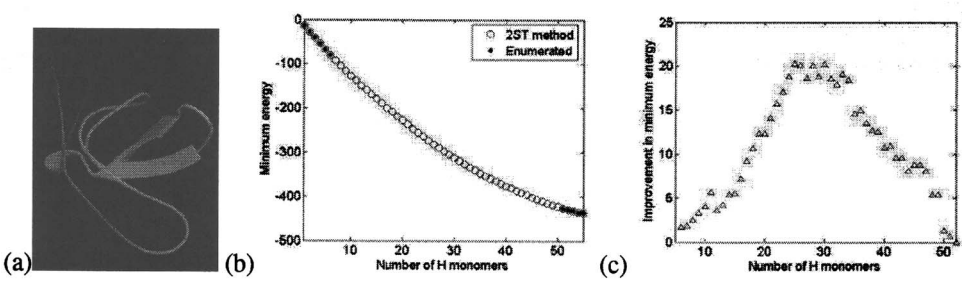

Fig. 4 (*a*) Ribbon representation of 1SRL protein. (*b*) Minimum energy as found by the new method (circles) and exhaustive enumeration (black dots) for a different number of H residues. (*c*) Improvement energy in the minima found by the new method over the previous method.

Unfortunately, it becomes increasingly prohibitive to verify it by enumeration. For $N_H=7$, there will be $6.2E8$ sequences to be enumerated and it gets larger for larger values of N_H. Noting this inability to find global minima for comparison, In Fig. 4(*c*), the improvement in the minimum energy found by the new method when compared to the old method is shown. In the old method, the first stage was a result of a graph spectral method. It is worth noting that in some cases there is an improvement of over 20 units in the minimum energy with the new method of this paper.

In Figs. 5(*a*) and 5(*b*), the energy improvement over the old method of [16] is shown for 3-D lattice models of sizes $4\times4\times4$ and $5\times5\times5$, respectively. For the first-stage optimization, the initial guess is given uniformly as 0.5 for all design variables. For the second-stage optimization, the monomer states obtained from the first-stage optimization are used as an initial guess and the peak function tuning parameter σ is gradually decreased from 0.5 until discrete sequences are obtained. The improvement in the minimum energy is clear from Figs. 5(*a*) and 5(*b*). There is only one case in both where the new two-stage synthesis method gives a result that is worse than the result of the earlier method. For those cases, the energy level from the first-stage optimization was probably near an inferior local minimum. By trying another initial guess for the first-stage optimization, a sequence showing an improved energy level could be obtained, and the final sequence at the end of the second-stage optimization showed the same energy level as the one obtained by the graph theory based method. If a global minimization method were to be used, even such cases will not arise. This is the strength of the QP and the new method.

Next, HP models of several real proteins from the Protein Data Bank (PDB [32]) were considered. The number of H residues in each case was determined by the usual classification based on the type of amino acids. Figures 6–9 shows the results in which the considered proteins, respectively, were Ubiquitin (PDB code: 1UBQ) with 76 residues out of which 71 were taken as of the H type, Csk homologous kinase (1JWO) with 97 residues and 82 H type, triosephosphate isomerase with sulfate with 250 residues, and tobacco ringspot virus capsid protein with 500 residues. In all cases, significant improvement in the minimum energy over the old method can be observed. However, each case has a few instances where the improvement was not achieved.

Based on the above results, the strength and the weakness of the new two-stage method can be summarized as follows. In the new method, the QP problem in the first stage uses linear interpolation of the residue states but with twice the number of variables in the case of HP models. But this being a QP problem, a deterministic global optimization method could be used. It is shown through results in this paper that the QP problem, when solved even with a local optimization method, often gives a globally minimum result. A weakness of the QP formulation is that its solution may consist of the intermediate states between H and P. It means that

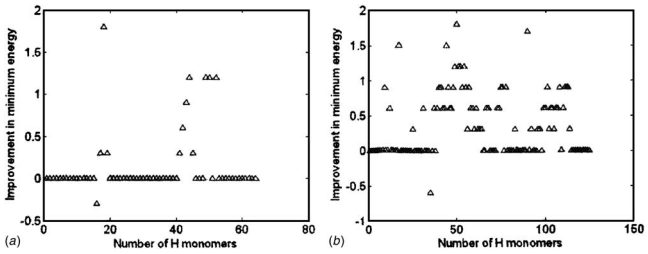

Fig. 5 The improvement in the energy beyond what the graph spectral-based old method gave plotted against the number of H residues for a (*a*) $4\times4\times4$ lattice; (*b*) $5\times5\times5$ lattice

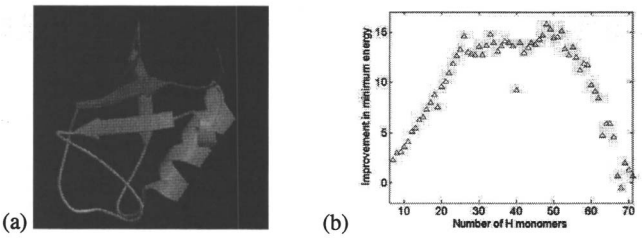

Fig. 6 (*a*) Ribbon representation of Ubiquitin (1UBQ). (*b*) Energy improvement over the old method.

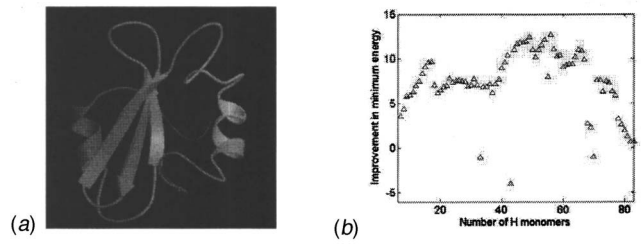

Fig. 7 (*a*) Ribbon representation of Csk Homologous Kinase (1JWO). (*b*) Energy improvement over the old method.

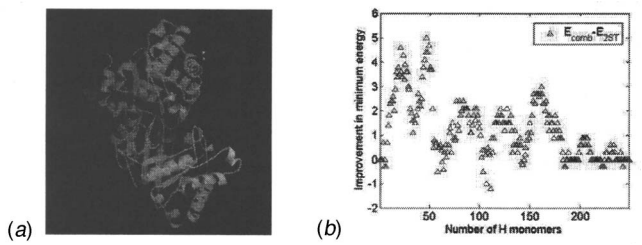

Fig. 8 (*a*) Ribbon representation of triophosphate isomerase complex with sulfate (5TIM). (*b*) Energy improvement over the old method.

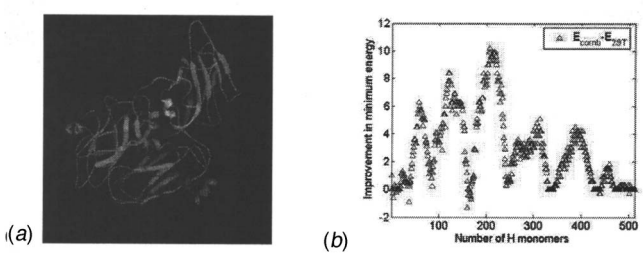

Fig. 9 (*a*) Ribbon representation of tobacco ringspot virus capsid protein (1A6C). (*b*) Energy improvement over the old method.

once again a combinatorial problem arises. This is efficiently tackled by the peak-function-based state-interpolation method in the second stage, which gives discrete states of H or P in the final solution. The results presented in this paper show that the solution of the QP method provides an excellent initial guess. The implementation of the methods of this paper was done in MATLAB [29]. The minimum was found within a few (<10) minutes on a single-processor desktop computer for each case considered in this paper. Considering the extremely large number of sequences, this computation time shows that the method is computationally very efficient.

5 Conclusions

Protein sequence synthesis is a discrete design problem because each residue in the heteropolymer chain of a protein may be any one of the 20 types of amino acids. If there are N residues in a chain, the number of possible sequences ($=20^N$) is very large due to combinatorial explosion. In this paper, we briefly reviewed our earlier method that proposed continuous modeling of this discrete problem. The earlier method consisted of two stages wherein the first stage used a spectral graph theory-based method to reduce the problem size and provide a good initial guess for the second stage. In this paper, we presented a new two-stage method wherein a quadratic programming (QP) problem was used in the first stage. This QP problem is amenable to applying deterministic global minimization algorithms. However, in this paper, we used local optimization algorithms and showed that the global minimum can still be obtained often as was verified using exhaustive enumeration for small protein models. The examples with large proteins show that the new method shows improvement over the earlier method, and often provides sequences with global minimum energy.

The main achievement of the current work is the drastic reduction in the computational cost and the improvement in minimum energy of the obtained solution in synthesizing the sequence of the HP lattice model of a protein. More importantly, the QP formulation can be solved to find the global minimum. Furthermore, the QP problem is not limited to only two types of amino acid residues. Our ongoing work is aimed towards developing a scalable deterministic global optimization method as well as extending the second-stage optimization to multiple types of amino acid residues.

Acknowledgments

The support of the National Science Foundation Grant No. DMI02-00362 is gratefully acknowledged. The third author thanks the National Science Foundation Grant No. DMS02-02536 for support. The authors thank Professor Jeffrey Saven (Chemistry, University of Pennsylvania) for many useful discussions.

References

[1] Kim, M. K., Li, W., Shapiro, B. A., and Chirikjian, G. S., 2003, "A Comparison Between Elastic Network Interpolation and MD Simulation of 16S Ribosomal RNA," J. Biomol. Struct. Dyn., **21**, pp. 395–405.
[2] Kazerounian, K., 2004, "From Mechanisms and Robotics to Protein Confrmation and Drug Design," ASME J. Mech. Des., **126**, pp. 40–45.
[3] Dubey, A., Sharma, G., Mavroidis, C., Tomassone, M. S., Nikitczuk, K., and Yarmush, M. L., 2004, "Computational Studies of Viral Protein Nano-Actuator," J. Comput. Theor. Nanosci., **1**, pp. 18–28.
[4] Lesk, A. M., 2001, *Introduction to Protein Architecture*, 1st ed., Oxford University Press, Oxford, NY.
[5] Miyazawa, S., and Jernigan, R., 1985, "Estimation of Effective Inter-Residue Contact Energies From Protein Crystal Structures," Macromolecules, **18**, pp. 534–552.
[6] Anfinsen, C., 1973, "Principles That Govern the Folding of Protein Chains," Science, **181**, pp. 223–230.
[7] Zou, J., and Saven, J. G., 2000, "Statistical Theory of Combinatorial Libraries of Folding Proteins," J. Mol. Biol., **296**, pp. 281–294.
[8] Sun, S., Brem, R., Chan, H. S., and Dill, K. A., 1995 "Designing Amino Acid Sequences to Fold With Good Hydrophobic cores," Protein Eng., **8**, pp. 1205–1213.
[9] Jones, D. T., 1994, "De Novo Protein Design Using Pairwise Potentials and Genetic Algorithm," Protein Sci., **3**, pp. 567–574.
[10] Pande, V. S., Grosberg, A. Y., and Tanaka, T., 1994, "Protein Superfamilies and Domain Superfolds," Nature (London), **372**, pp. 631–634.
[11] Hellinga, H. W., and Richards, F. M., 1994, "Optimal Selection of Sequences of Proteins of Known Structure by Simulated Evolution," Proc. Natl. Acad. Sci. U.S.A., **91**, pp. 5803–5807.
[12] Saven, J. G., and Wolynes, P. G., 1997, "Statistical Mechanics of the Combinatorial Synthesis and Analysis of Folding Macromolecules," J. Phys. Chem. B, **101**, pp. 8375–8389.
[13] Shaknovich, E. I., and Gutin, A. M., 1993, "Engineering of Stable and Fast-Folding Sequences of Model Proteins," Proc. Natl. Acad. Sci. U.S.A., **90**, pp. 7195–7199.
[14] Park, S., Yang, X., and Saven, J., "Advances in Computational Protein Design," Curr. Opin. Struct. Biol. (submitted).
[15] Singh, M., "Computational Methods Towards Predicting Aspects of Protein Structure and Interactions," *Special Session on Geometry of Protein Modeling in 248th Regional Meeting of the American Mathematical Society*, Lawrenceville, NJ, 17–19 April 2004.
[16] Koh, S. K., Ananthasuresh, G. K., and Vishveshwara, S., 2005, "A Deterministic Optimization Approach to Protein Sequence Design Using Continuous Models," Int. J. Robot. Res., **24**, pp. 109–130.
[17] Sanjeev, B. S., Patra, S. M., and Vishveshwara, S., 2001, "Sequence Design in Lattice Models by Graph Theoretical Methods," J. Chem. Phys., **114**, pp. 1906–1914.
[18] Lo, C., and Papalambros, P. Y., 1995, "On Global Feasible Search for Global Design Optimization with Application to Generalized Polynomial Models," ASME J. Mech. Des. **117**, pp. 402–408.
[19] Lo, C., and Papalambros, P. Y., 1996a, "A Deterministic Global Design Optimization Method for Nonconvex Generalized Polynomial Problems," ASME J. Mech. Des., **118**, pp. 75–81.
[20] Lo, C., and Papalambros, P. Y., 1996b "A Convex Cutting Plane Algorithm for Global Solution of Generalized Polynomial Optimal Design Models," ASME J. Mech. Des., **118**, pp. 82–88.
[21] Tuy, H., and Thuong, N. V., 1988, "On the Global Minimization of a Convex Function Under General Nonconvex Constraints," Appl. Math. Optim., **18**, pp. 13–20.
[22] Yue, K., and Dill, K. A., 1992, "Inverse Protein Folding Problem: Designing Polymer Sequences," Proc. Natl. Acad. Sci. U.S.A., **89**, pp. 4163–4167.
[23] Lau, K. F., and Dill, K. A., 1989, "A Lattice Statistical Mechanics Model of the Conformational and Sequence Spaces of Proteins," Macromolecules, **22**, pp. 3986–3997.
[24] Dill, K. A., Bromberg, S., Yue, K., Fiebig, K. M., Yee, D. P., Thomas, P. D., and Chan, H. S., 1995, "Principles of Protein Folding: A Perspective from Simple, Exact Models," Protein Sci., **4**, pp. 561–602.
[25] Bendsøe, M. P., and Sigmund, O., 1999, "Material Interpolation Scheme in Topology Optimization," Arch. Appl. Mech., **69**, pp. 635–654.
[26] Li, H., Tang, C., and Wingreen, N. S., 1997, "Nature of Driving Force for Protein Folding: A Result from Analyzing the Statistical Potential," Phys. Rev. Lett., **79**, pp. 765–768.
[27] Rao, S. S., *Engineering Optimization*, 3rd ed, Wiley, New York, 1996.
[28] Li, H., Helling, R., Tang, C., and Wingreen, N., 1996, "Emergence of Preferred Structures in a Simple Model of Protein Folding," Science, **273**, pp. 666–669.
[29] *Matlab*, 2004, Numerical Analysis Software from Mathworks, Inc., Woburn, MA, www.mathworks.com.
[30] Yin, L., and Ananthasuresh, G. K., 2001, "Topology Optimization of Compliant Mechanisms with Multiple Materials Using a Peak Function Material Interpolation Scheme," Struct. Multidiscip. Optim., **23**, pp. 49–62.
[31] Yin, L., and Ananthasuresh, G. K., 2002, "Novel Design Technique for Electro-Thermally Actuated Compliant Micromechanisms," Sens. Actuators, A, **97–98**, pp. 599–609.
[32] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E., 2001, "The Protein Data Bank." Nucleic Acids Res. , **28**, pp. 235–242. Also see: http://www.pdb.org