Sung K. Koh

Mechanical Engineering and Applied Mechanics
University of Pennsylvania
Philadelphia, 19104-6315, USA

G. K. Ananthasuresh

Mechanical Engineering and Applied Mechanics
University of Pennsylvania
Philadelphia, 19104-6315, USA
and
Mechanical Engineering, Indian Institute of Science
Bangalore 560 012, India
suresh@mecheng.iisc.ernet.in

Saraswathi Vishveshwara

Molecular Biophysics Unit
Indian Institute of Science
Bangalore 560 012, India

A Deterministic Optimization Approach to Protein Sequence Design Using Continuous Models

Abstract

Determining the sequence of amino acid residues in a heteropolymer chain of a protein with a given conformation is a discrete combinatorial problem that is not generally amenable for gradient-based continuous optimization algorithms. In this paper we present a new approach to this problem using continuous models. In this modeling, continuous "state functions" are proposed to designate the type of each residue in the chain. Such a continuous model helps define a continuous sequence space in which a chosen criterion is optimized to find the most appropriate sequence. Searching a continuous sequence space using a deterministic optimization algorithm makes it possible to find the optimal sequences with much less computation than many other approaches. The computational efficiency of this method is further improved by combining it with a graph spectral method, which explicitly takes into account the topology of the desired conformation and also helps make the combined method more robust. The continuous modeling used here appears to have additional advantages in mimicking the folding pathways and in creating the energy landscapes that help find sequences with high stability and kinetic accessibility. To illustrate the new approach, a widely used simplifying assumption is made by considering only two types of residues: hydrophobic (H) and polar (P). Self-avoiding compact lattice models are used to validate the method with known results in the literature and data that can be practically obtained by exhaustive enumeration on a desktop computer. We also present examples of sequence design for the HP models of some real proteins, which are solved in less than five minutes on a single-processor desktop computer. Some open issues and future extensions are noted.

KEY WORDS—protein sequence design, deterministic optimization, inverse folding, lattice models, and graph spectral method

1. Introduction

Proteins are heteropolymer chains of 20 types of amino acid residues strung together with peptide bonds and folded into intricate three-dimensional (3D) structures. It is generally agreed that the sequence of residues in the polypeptide chain determines its folded structure, also called a "conformation". The study of sequence-structure relationships in proteins involves two distinct problems: determining the conformation for a given amino acid sequence in the chain (sequence-to-structure problem) and determining the sequences for a desired conformation (structure-to-sequence problem). The second problem is also known as the "inverse folding problem" (Pabo 1983; Ponder and Richards 1987). The two problems are in some sense "what is" and "what is to be" problems, respectively. The latter is naturally a design problem and is the focus of this paper. In view of the observation that proteins can fit into only a limited number of folds (Chothia 1992; Banavar and Maritan 2003), efficient and robust algorithms to identify all possible sequences to a chosen conformation

The International Journal of Robotics Research
Vol. 24, No. 2–3, February–March 2005, pp. 109-130,
DOI: 10.1177/0278364905050354

can help in assigning structures to a large, rapidly increasing number of sequences in the databases such as SWISS-PROT (an annotated protein sequence database established in 1986; see http://www.ebi.ac.uk/swissprot/).

In the simplest manifestation of the sequence design problem, a desired folded conformation of a protein's backbone consisting of only C^{α} atoms is used to determine the side chains (and thus the sequence) to optimize a suitable criterion with the help of some interaction energies among the residues within an environment. In addition to the design of new proteins that fold to a desired conformation, this problem can shed light on understanding the principles underlying protein folding and the variability in the sequences of naturally occurring proteins (Zou and Saven 2000).

The computational complexity of the protein sequence design problem can be understood by considering a chain consisting of N residues. Since there are 20 amino acid types, there will be a total of 20^N possible sequences. If we also consider different orientations of the side chains (called "rotamer configurations"), there will be much more than 20 possibilities for each residue site in the chain. This makes the number of possibilities even larger. Out of these, one or more which satisfy a criterion that discriminates in favor of a given folded conformation are to be found. Therefore, exhaustive enumeration of all possible sequences and thereby finding the best for real proteins ($N >\sim 50$ and reaching a few thousands) is beyond the scope of the computational power even today. Hence, search methods are developed to identify sequences that are likely to fold to a desired conformation. These include stochastic and deterministic methods. Since there are papers that review various methods (e.g., Zou and Saven 2000), only a small sample of works, one in each category, is noted here. For example, Hellinga and Richards (1994) used Monte Carlo methods, Desjarlais and Handel (1995) used genetic algorithms, and Deutsche and Kurosky (1996) used simulated annealing. Desmet et al. (1992) proposed a dead-end elimination algorithm to screen out improbable sequences efficiently. Saven and Wolynes (1997) used statistical mean field theory based methods to determine site-specific probabilities for most probable amino acid types using deterministic optimization algorithms. Sanjeev, Patra, and Vishveshwara (2001) used a graph spectral method, which ranks the sites for amino acid types with very little computation and thereby designs a sequence. There have also been attempts based on the deterministic global optimization methods but mainly for structure prediction rather than sequence design (see an overview by Phillips, Rosen, and Dill 2001).

All the aforesaid methods make simplifying assumptions for several reasons. First, there is no universally accepted criterion to characterize the folded conformation of a protein based on its amino acid sequence. Several criteria are proposed based on theoretical analyses and experimental observations. The researchers who use computational methods to find the best sequences have adopted one or more of these criteria, which include minimum energy, maximum gap in energy from the average energy of unfolded conformations, maximum entropy, etc. Secondly, in computationally evaluating these criteria, some use very simple interaction potentials between neighboring pairs of residues, while others model forces even up to the atomistic detail. Some consider all the 20 amino acid types including permissible rotamer configurations while others consider a reduced set. Thirdly, validation of an obtained sequence as one of the best is difficult because there are too many possibilities to enumerate and there are uncertainties in modeling the interactions as well as identifying the reduced set of amino acid types. Experimental results provide valuable insight but not conclusive evidence that a sequence is the best. Hence, simple exact lattice models have been proposed (Go 1983; Lau and Dill 1989), which allow the identification of best sequences for a given conformation by complete enumeration.

In lattice models, the positions of the residue sites are fixed as an orderly grid in either two or three dimensions. Compact, self-avoiding chains, such as those shown in Figures 1(a) and (b), are used to describe desired conformations. Furthermore, only two types of residues are considered: hydrophobic (H) and polar (P). This is supported by a widely accepted belief that the hydrophobicity of some amino acid types is one of the principal driving forces for protein folding. In HP models, a very simple normalized interaction energy, e, between neighboring sites that interact with each other is used:

$$e_{HH} = -2.3; \qquad e_{HP} = -1.0; \qquad e_{PP} = \quad 0.0. \quad (1)$$

The above energetic information is deduced from the widely adopted MJ (Miyazawa and Jernigan 1985) interaction matrix using a reduced-order eigenanalysis (Li, Tang, and Wingreen 1997). Dill et al. (1995) note that the lattice models possess many of the observed kinetic and thermodynamic properties of real proteins. Some new properties (e.g., designability by Li et al. 1996) have also been proposed based on the analysis of lattice models. Hence, lattice model based studies are instructive while being computationally tractable. With lattice models, the sequence design problem reduces to identifying the type of the residue (H or P) at each site for a given conformation to satisfy a chosen criterion. An advantage of this approach is that all possibilities can be enumerated for small grid sizes. This serves as a way for the validation of results based on only computation.

Many studies on lattice models use enumerative, pattern matching or other methods that require explicit realization of the sequence fully or partially. Perhaps hindered by excessive computation, these studies have been limited to 6×6 in two dimensions and $4\times3\times3$ in three dimensions because each of these two cases has $2^{36} \approx 68E9$ possible sequences and demands large computation time. Larger grid sizes become impractical with desktop computers and warrant parallel processors and supercomputers. Graph spectral theory based techniques offer an attractive alternative and have been

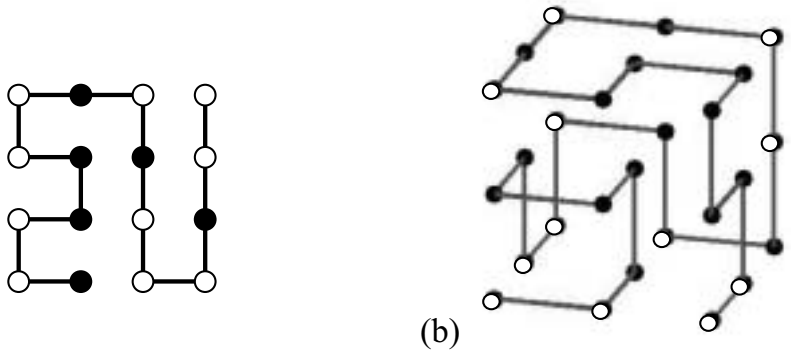

Fig. 1. HP lattice models: (a) 2D 4×4 lattice; (b) 3D 3×3×3 lattice. Black dots represent hydrophobic (H) residues, and white dots polar (P) residues.

successfully explored by Vishveshwara, Brinda, and Kannan (2002). An appealing feature of this approach is that it gives explicit consideration to the topology of the protein. This approach is described in Section 5, as it is used later in this paper.

A notable feature of almost all the protein sequence design methods developed until now[1] is that they tackle the discrete and combinatorial nature of this problem directly. In general, algorithms for solving discrete combinatorial problems are not as efficient as those for searching for a minimum in a continuous space using deterministic algorithms. Stochastic algorithms, such as simulated annealing, and genetic algorithms are other options and they have been applied to this problem (Desjarlais and Handel 1995; Deutsche and Kurosky 1996). An alternative approach is proposed in this paper by developing a continuous model of the discrete problem and thereby circumventing the combinatorial explosion and making way for smooth and deterministic (i.e., gradient-based) optimization algorithms to find the optimum sequence. This approach is explained in Section 3. A computationally efficient method of solution, called the optimality criteria method, is presented in Section 4 along with examples solved using this and other gradient-based algorithms. By combining the continuous model approach with a graph theory based ranking method, the sequence design problem can be solved in a few minutes (often less than 1 min and only occasionally exceeding 10 min) on a single-processor desktop computer for the HP models consisting of hundreds of residues. This is presented in Section 6. The results and a discussion including limitations and future extensions are given in Section 7. Concluding remarks are in Section 8. Next, the scope of the sequence design problem as addressed in this paper is laid out in Section 2.

1. The method using the site-specific probabilities by Zou and Saven (2000) is an exception and is commented upon later in Section 7.1 of this paper.

2. Scope of the Paper

According to the Anfinsen (1973) thermodynamic hypothesis and other subsequent studies, the designed sequences need to satisfy three criteria. First, a designed sequence should have the desired conformation as its native state. That is, this sequence should have the minimum energy for this native conformation among all possible conformations. Secondly, for this sequence there should not be any other conformations with the same minimum energy. In other words, there should not be degeneracy in the native states. A chain of such a sequence is likely to fold uniquely to this structure (i.e., conformation). Thirdly, the sequence should be stable in that its energy in the desired conformation should be widely separated from the average of energies of unfolded conformations. Such a large energy gap will enable the protein to stably adhere to that conformation. Alternatively, a different reasoning as explained below can identify best sequences.

Imagine a sequence space **S** (the set of all possible sequences) and a conformation *space* **C** (the set of all possible conformations) for a chain of Nsites. Assume that there exists a subset $\mathbf{S}^{ns}_{c^*} \subset \mathbf{S}$, the elements of which have desired conformation c^* as the native state (ns). As noted before, according to the thermodynamic hypothesis (Anfinsen 1973), such sequences are likely to fold to that conformation. The identification of members of $\mathbf{S}^{ns}_{c^*}$ requires searching in the **C** space for every considered sequence $s \in \mathbf{S}$ to see if s has a unique global minimum energy in conformation c^*. Given the difficulty of finding a global minimum and ensuring that it is unique in the conformation space, alternative approaches are reported in the literature to make the problem computationally tractable (Yue and Dill 1992; Shaknovich and Gutin 1993). One practical approach followed by many researchers is to avoid the simultaneous search in both **S** and **C**, and to limit it to a search in **S** alone to identify the subset $\mathbf{S}^{e}_{c^*} \subset \mathbf{S}$ with minimum energy. Although this approach does not consider the search in the conformation space explicitly to confirm $s \in \mathbf{S}^{ns}_{c^*}$,

the search to identify $\mathbf{S}^{e}_{c^*}$ includes the energy-competition of unfolded state approximately when the amino acid composition is constrained (Park, Yang, and Saven 2004). Therefore, searching in $\mathbf{S}$ alone to identify $s \in \mathbf{S}^{e}_{c^*}$ has been used as an indirect way of identifying $s \in \mathbf{S}^{ns}_{c^*}$. It has also been observed that most sequences that are best in the sequence space are also well behaved in the conformation space (Sanjeev, Patra, and Vishveshwara 2001). This paper focuses on computationally efficient methods for determining the sequence (or more if there is degeneracy) with minimum energy in the sequence space for any desired conformation using continuous modeling.

If only H and P types of residues are considered, the sequence space will have 2^N sequences. If all 20 amino acids are considered, there will be 20^N sequences. With the variation in rotamer states included, the base in the possible number of sequences will be even larger, as stated earlier. The structure space for a real protein is unlimited. However, for a lattice model, it can be restricted. For compact self-avoiding (CSA) lattice models, the number of possible structures for a given grid size is finite. In general, the conformation space is much smaller than the sequence space. For example, a $3\times3\times3$ grid has 103,346 CSA conformations (Sali, Shakhnovich, and Karplus 1994) and $2^{27} \approx 0.13E9$ sequences. Hence, benchmarking a new sequence design technique is possible with such lattice models since all possible conformations can be practically enumerated and the native conformation for a designed sequence can be verified. Thus, the new technique of this paper is first illustrated with HP lattice models. Extension to irregular lattice models is not precluded as demonstrated with the examples of real proteins. Extension to a larger number of residue types is commented upon at the end of the paper. The technique combines two methods: one based on optimization with continuous models set forth in this paper, and the other based on the graph theory, which has already been explored (Sanjeev, Patra, and Vishveshwara 2001) and is briefly described with additional insight later in the paper.

3. Continuous Modeling of the Discrete Problem of Sequence Design

When a target conformation of a protein chain is given, sequence design entails choosing the type of the residue at each site in the chain. First, consider the simple HP lattice models such as those shown in Figure 1. In this, either an H residue or a P residue can occupy each site. That is, the state of a site can be H or P. Let the H state be denoted by 1 and P state by 0. The discrete nature of the state of a site leads to a combinatorial explosion and warrants appropriate methods to deal with it. Instead, consider the following state function S that describes the state of the site continuously between 1 and 0:

$$S = e^{-\left(\frac{\rho}{\sigma}\right)^2}. \tag{2}$$

The variable ρ associated with a site continuously interpolates its state between 0 and 1 by the Gaussian distribution function in eq. (2), and as shown in Figure 2. 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 $\sigma \to 0$, the state function in eq. (2) approaches the Dirac delta function. 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. Next, the construction of the continuous energy landscape due to the interactions between residues using this continuous modeling is described.

First, consider a simple situation involving only three residues where the middle residue has interactions with its neighbors. As shown in Figure 3(a), the middle residue is fixed to be in the H state while the states of the two neighboring sites are to be determined such that the total energy of the system is a minimum. An enumerative approach will consider $2^2 = 4$ possibilities to conclude that both residues should be of the type H. Using the continuous model, the total energy of the system can be written as

$$E = \{e_{HP}(1 - S_1) + e_{HH}S_1\} + \{e_{HP}(1 - S_2) + e_{HH}S_2\} \tag{3}$$

where

$$S_1 = e^{-\left(\frac{\rho_1}{\sigma}\right)^2} \text{ and } S_2 = e^{-\left(\frac{\rho_2}{\sigma}\right)^2}.$$

$e_{HP}$and e_{HH} are as given in eq. (1), and ρ_1 and ρ_2 are the variables that determine the states of left and right residues in Figure 3(a), respectively. The function in eq. (3) takes into account the possibility that left and right lattice sites in Figure 3(a) can assume values in between 1 and 0. When both sites are precisely in one or the other state, the energy is computed correctly. The plot of the energy as calculated by eq. (3) is shown in Figure 3(b). Clearly, there is a minimum at ($\rho_1 = 0$, $\rho_2 = 0$) with $E = -4.6$. Thus, by finding the minimum of the function in eq. (3) using a continuous, deterministic optimization algorithm, the set of residue types (i.e., the sequence) can be determined without enumerating all the possibilities.

If the state of the middle residue is also unknown, by adding a third variable ρ_3 to interpolate the state of the middle residue, the energy function can be written as

$$\begin{aligned} E = {} & \{e_{HP}S_3(1 - S_1) + e_{HP}(1 - S_3)S_1 + e_{HH}S_3S_1\} \\ & + \{e_{HP}S_3(1 - S_2) + e_{HP}(1 - S_3)S_2 + e_{HH}S_3S_2\} \end{aligned} \tag{4}$$

where $S_3 = e^{-(\rho_3/\sigma)^2}$. The minimization of the function of three variables in eq. (4) is equivalent to finding the best of the eight possibilities involving three residues. When the size of the protein chain is long and the number of interactions is large, the continuous optimization method is computationally more efficient than enumeration and other methods, as will be seen in the examples later.

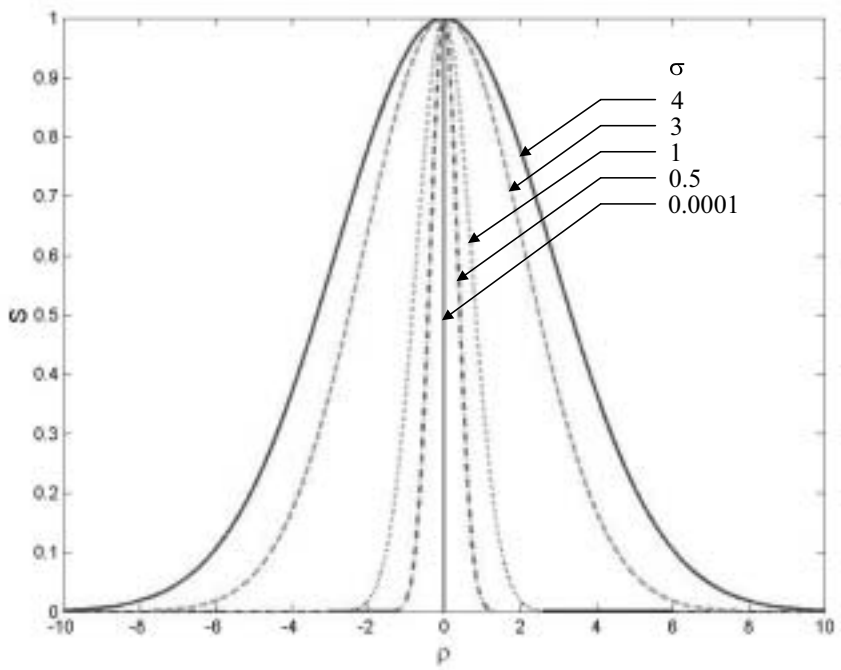

Fig. 2. Continuous interpolation of the state of a lattice site between 1 and 0 as determined by a single variable $-\infty < \rho < \infty$.

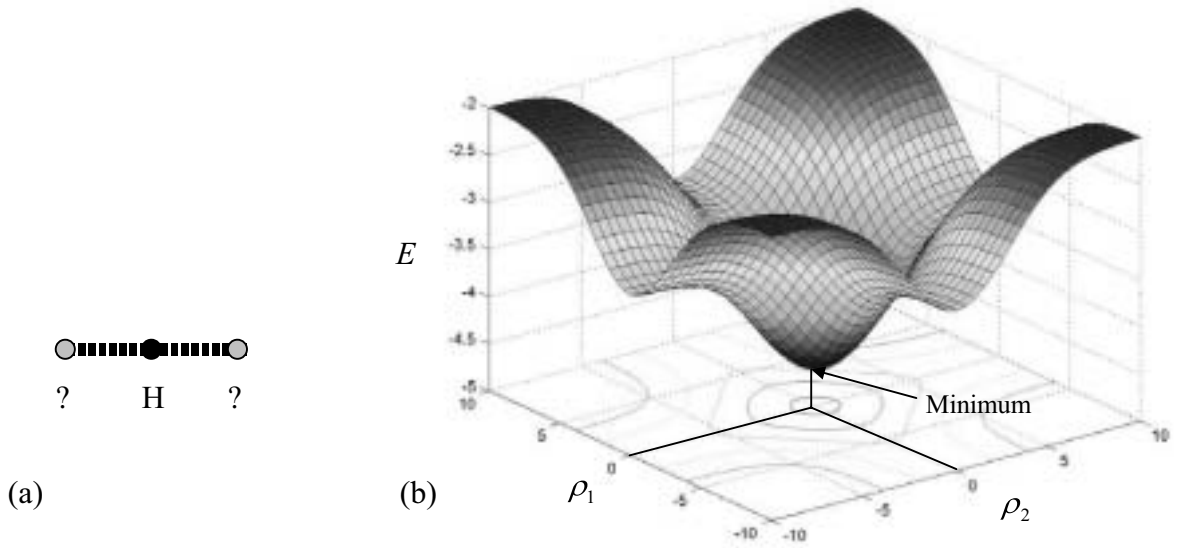

Fig. 3. (a) A hypothetical situation of an H residue interacting with its two neighboring residues whose types are unknown. (b) The continuous energy function wherein the states of the two neighboring sites are continuously interpolated with variables ρ_1 and ρ_2.

Now, consider a lattice model protein chain in a known conformation consisting of N sites. Its total energy can be written as

$$E = 0.5 \left[\sum_{i=1}^{N} \sum_{j=neib_i[1]}^{neib_i[m_i]} \left\{ e_{HP} S_i (1 - S_j) + e_{HP} (1 - S_i) S_j + e_{HH} S_i S_j \right\} \right] \tag{5}$$

where $neib_i$ is an array of neighboring sites, m_i in number, that have interactions with the site i. In the above equation, the serial number of the jth neighboring site is denoted by $neib_i[j]$. Since each pair is counted twice, once as $(i, j)$and again as (j, i), a factor of 0.5 is present in eq. (5).

Alternatively, the total energy can also be written as

$$E = 0.5 \left[\sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} A(i, j)\ e(S_i, S_j) \right] \tag{6a}$$

where

$$A(i, j) = \begin{cases} 1 & \text{if sites } i \text{ and } j \text{ interact} \\ 0 & \text{otherwise} \end{cases} \tag{6b}$$

$$e(S_i, S_j) = e_{HP} S_i (1 - S_j) + e_{HP} (1 - S_i) S_j + e_{HH} S_i S_j . \tag{6c}$$

In view of eq. (1), it is easy to see that the minimum energy given by eqs. (5) or (6a) is obtained when all residues are of the type H. This is true for any CSA conformation on the lattice. A chain with such a sequence, called a homopolymer, is not of interest because it trivially gives the minimum energy for any conformation, and hence it does not have one fixed stable conformation. Thus, it is meaningless to solve the minimum energy problem without a constraint on the number of H residues.[2] Such a constraint can easily be imposed within the framework of the continuous modeling as shown below.

$$\sum_{i=1}^{N} S_i - N_H = 0 \quad \Rightarrow \quad \sum_{i=1}^{N} e^{-(\rho_i/\sigma)^2} - N_H = 0 \tag{7}$$

where N_H is the prescribed number of H residues out of the total number of possible N residues. Just as the expressions for the energy do, the expression for the constraint also allows the possibility of some sites being in the intermediate state between H and P while being accurate when the sites are precisely assigned H or P status (i.e., with $|\rho_i|$ equal to 0 or sufficiently large value, respectively). Similarly, other conditions that need to be considered can be included as constraints expressed in continuous form. If certain sites are preferred to be in a particular state, this can be easily done by excluding the variable associated with it from the minimization procedure.

The problem of sequence design with a prescribed number of H residues can now be written as a constrained minimization problem involving N continuous variables, namely $\{\rho_1, \rho_2, \cdots, \rho_N\}$, each of which can assume any numerical value in the range $(-\infty, \infty)$.

$$\begin{aligned} &\text{Minimize } E = 0.5 \left[\sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} A(i, j)\ e(S_i, S_j) \right] \\ &\text{with respect to } \{\rho_1, \rho_2, \cdots, \rho_N\} \\ &\text{and subject to } \sum_{i=1}^{N} S_i - N_H = 0 . \end{aligned} \tag{8}$$

The method of solution for the above problem, examples, and some features of the continuous formulation and their advantages are presented in Section 4. An alternative formulation without the need to prescribe a constraint on the number of H residues is presented next.

3.1. Energy Gap Criterion

Some researchers argue that the minimum energy criterion is not the best criterion. Deutsche and Kurosky (1996) observed that this minimum energy criterion finds the sequence for a non-degenerate native state of a desired conformation less often than an energy gap criterion that they proposed. Zou and Saven (2000) explored a slightly different energy gap criterion. One form of the energy gap denoted by Δ is given by the following expression

$$\Delta = E - 0.5 \left[\sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \langle A(i, j) \rangle\ e(S_i, S_j) \right] \tag{9}$$

where $\langle A(i, j) \rangle$ is the average taken over all the conformations other than the desired conformation. In this method, the sequence that maximizes Δ is sought. Now, a constraint on the composition of residues (i.e., how many are H) is not necessary. The application of the continuous optimization to maximization of the energy gap is described by Koh and Ananthasuresh (2004) while this paper focuses only on minimizing the energy.

3.2. A Note on Modeling Preferences in This Paper

The continuous modeling presented here works with either of the above criteria and possibly with any other criterion that can be written in a mathematical form. In this paper, only the minimum energy criterion is considered.

2. An alternative criterion based on energy gap where such a constraint is not needed is noted in Section 3.1.

The next preference is concerned with the way inter-residue interactions are modeled. There are different ways to identify the interacting residues. Some consider only the immediate neighbors in the folded conformation that are not adjacent in the chain, i.e., those that are not connected with a peptide bond (Li et al. 1996). Some also include the peptide-bonded neighbors (Deutsche and Kurosky 1996). Others consider residues that are one level farther than the immediate neighbors. Sometimes interactions among three residues (Banavar and Maritan 2003), and higher-order interactions are also considered. In real proteins, a distance (say 7 Å) could be imposed to identify the interacting neighbors. Any of these approaches can be modeled using the framework presented here. In the lattice model based examples included in this paper, only the non-bonded immediate neighbors are assumed to interact. In HP models of real proteins, a distance of 6.5 Å is used to identify the interacting neighbors as done in the past (Miyazawa and Jernigan 1985; Patra and Vishveshwara 2000).

4. Solution Method and Results

The constrained minimization algorithm in eq. (8) can be solved using any of the gradient-based algorithms (Rao 1996). These algorithms use the gradient information to identify a descent direction for every iteration and move along that using a one-dimensional search to lower the objective function while satisfying the constraint. For continuous problems (with C^1 continuity), most algorithms generally converge to a local minimum. There are many robust optimization software programs that can be readily used to solve continuous optimization problems. A routine, entitled fmincon, in the Optimization Toolbox of Matlab (numerical analysis software from Mathworks, Inc., Woburn, MA, USA; see http://www.mathworks.com) is used in this work. This routine combines the Sequential Quadratic Programming and Trust Region algorithms with an efficient one-dimensional search algorithm based on quadratic fitting and golden section algorithm. Instead of using a generic solver such as this, a computational method that is specifically efficient for this problem is also used in this work. This method is along the lines of a class of algorithms called optimality criteria methods, used successfully in structural optimization (Haftka and Gürdal 1992) and the design of compliant mechanisms (Saxena and Ananthasuresh 2000; Yin and Yang 2001). At this point, it is appropriate to note that the state function proposed in this work for protein design was in fact motivated by the way structural topology optimization problems are formulated.

In structural topology optimization problems, material is to be optimally distributed in a given design region to satisfy some constraints and minimize a criterion (Bendsøe and Sigmund 2003). Originally, the material distribution problem is binary in that material may exist (state 1) or not (state 0) at each point within a region of interest. If we discretize the design region into a number of finitely sized cells, it leads to a combinatorial problem. Instead of solving such a discrete problem, a continuous interpolation between 0 and 1 is adopted. Details on this can be found in a review article by Bendsøe and Sigmund (1999) on material interpolation in topology optimization. Now, consider the problem of a structure to be made with three materials. Such a problem will have four states: 0 for no material, 1 for material 1, 2 for material 2, and 3 for material 3. For this, Yin and Ananthasuresh (2001, 2002) proposed a single continuous variable based formulation. This is indeed the basis for the state function defined in eq. (2). Naturally, this state function can be extended beyond HP models to all the 20 amino acid types, perhaps with more than one variable per site.

In structural optimization, very large problems (generally consisting of hundreds and sometimes even thousands of variables) are efficiently solved on a single-processor desktop computer within a few minutes. One such algorithm developed to solve the problem in eq. (8) is described next.

4.1. An Optimality Criteria Method

As is usual in constrained minimization algorithms, the Lagrangian, L, is written for the problem in eq. (8):

$$L = 0.5\left[\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} A(i,\,j)\; e(S_i,\,S_j)\right] + \Lambda\left[\sum_{i=1}^{N} S_i - N_H\right]. \tag{10}$$

Recall that only the state functions S depends on the variables $\{\rho_1, \rho_2, \cdots, \rho_N\}$ and all other quantities in eq. (10) are known except Λ, which is the Lagrange multiplier associated with constraint and is determined as part of the solution. For simplicity of notation, eq. (10) is rewritten by denoting the expression of the objective function as f and that of the constraint as g:

$$L = f + \Lambda g. \tag{11}$$

A necessary condition for a constrained minimum is given by

$$\frac{\partial f}{\partial \rho_i} + \Lambda \frac{\partial g}{\partial \rho_i} = 0 \Rightarrow B_i + \Lambda D_i = 0 \text{ for all } i = 1, 2, \cdots, N \tag{12}$$

where the partial derivatives, denoted by B_i and D_i, can easily be obtained analytically given the simple nature of the continuous functions involved here, and hence easily computed numerically. This is called an optimality criterion that must

be satisfied when the numerical algorithm converges to a minimum. Therefore, each variable ρ_i can be iteratively updated as follows:

$$\rho_i^{new} = \rho_i^{old} - (B_i + \Lambda D_i)\,. \tag{13}$$

The value of Λ needs to be updated in each iteration. The formula for Λ is obtained by substituting eq. (13) into the constraint equation in eq. (8). Since the constraint involved here is nonlinear, the linearized approximation of the constraint is considered before substituting eq. (13). That is, the linearized approximation of the constraint,

$$\left(\sum_{i=1}^{N} S_i - N_H\right)^{old} + \sum_{i=1}^{N} D_i(\rho_i^{new} - \rho_i^{old}) = 0$$

$$\Rightarrow g^{old} + \sum_{i=1}^{N} D_i(\rho_i^{new} - \rho_i^{old}) = 0 \tag{14}$$

with substitution from eq. (13), yields

$$g^{old} - \sum_{i=1}^{N} D_i(B_i + \Lambda D_i) = 0$$

$$\Rightarrow \sum_{i=1}^{N} \Lambda D_i^2 = g^{old} - \sum_{i=1}^{N} D_i B_i \tag{15}$$

from which Λ is obtained as

$$\Lambda = \frac{g^{old} - \sum_{i=1}^{N} D_i B_i}{\sum_{i=1}^{N} D_i^2}. \tag{16}$$

Since a linear approximation of the nonlinear constraint expression is used, the equality constraint may not be satisfied exactly. Hence, an inner loop is used in every iteration to adjust for this. Additionally, a conservative approach is used in practice by imposing limits on how much ρ_i can change. In this work, not more than 10% from the current value was allowed. Even though not essential in the present framework, upper and lower bounds on $\rho_i$are usually imposed for practical reasons. Let the upper and lower bounds be denoted by ρ_u and ρ_l. If the update formula in eq. (13) makes any ρ_i go beyond these limits, that variable will be set to the nearest bound, i.e., the upper or the lower bound. With this additional feature incorporated, eq. (16) can be modified as

$$\Lambda = \frac{g^{old} + \sum_{i\in U} D_i(\rho_u - \rho_i) + \sum_{i\in L} D_i(\rho_l - \rho_i) - \sum_{\substack{i=1 \\ i\notin U\cup L}}^{N} D_i B_i}{\sum_{i=1}^{N} D_i^2}, \tag{17}$$

where U and L denote the sets consisting of serial numbers of variables that have reached the upper or the lower bound, respectively. In the earlier implementation of this algorithm for compliant mechanism design (Yin and Ananthasuresh 2001), it was necessary to gradually decrease the large value of σ used at the beginning of the iterative process to smaller values to completely eliminate the intermediates states. However, the problems solved in this work worked well even with a fixed value of σ.

The variables are updated as described above until convergence. Several convergence criteria can be used. In this work, when the absolute value of the change in the objective function (i.e., the total energy) in successive iterations is less than a specified tolerance (say, 1E-4), the iterative procedure is stopped. By that time, all variables would have been determined such that the H or P states for each site in the chain are obtained. It is important to note that each iteration involves only one evaluation of the energy. Usually, these algorithms converge in 100–1000 iterations. This means that not more than 1000 (often much less than this) sequence permutations are tried to identify a sequence that minimizes the objective criterion. Thus, the efficiency over enumerative methods is immediately apparent. Next, examples with HP lattice models are presented.

4.2. Examples

Consider a 6×6 HP lattice shown in Figure 4(a) with the most designable (Li et al. 1996) conformation. This is the initial guess given to the optimality criteria method described above. With the number of H residues desired to be eight, a result shown in Figure 4(b) was obtained in less than 16 s including time for plotting figures. The algorithm was implemented in a Matlab environment and run on a PC. Instead of uncompiled code such as Matlab, if C or Fortran were to be used, it would have been even faster. The energy of this minimizing sequence was –18.1, which can be verified to be the absolute minimum by visual inspection. That is, if H residues are placed in any other arrangement, the energy will be higher. The value of σ (used to define the continuous state as in eq. (2)) was set at 0.5 in this example and others in this paper. The bounds for the variables were set at –5 and 5. At convergence, the optimized values of ρ_i, $i = 1, \ldots, 36$ are given below in an arrangement that corresponds to the 6×6 lattice sites:

5.0000, 5.0000, 5.0000, 5.0000, 4.0552, 5.0000,

5.0000, 4.0552, 3.5944, 3.4545, 3.5944, 5.0000,

5.0000, **0.0001**, **0.0001**, **0.0001**, **0.0001**, 5.0000,

5.0000, **0.0002**, **0.0273**, **0.0273**, **0.0002**, 5.0000,

5.0000, 4.0197, 2.6855, 2.6855, 4.0197, 5.0000,

5.0000, 5.0000, 5.0000, 5.0000, 5.0000, 5.0000.

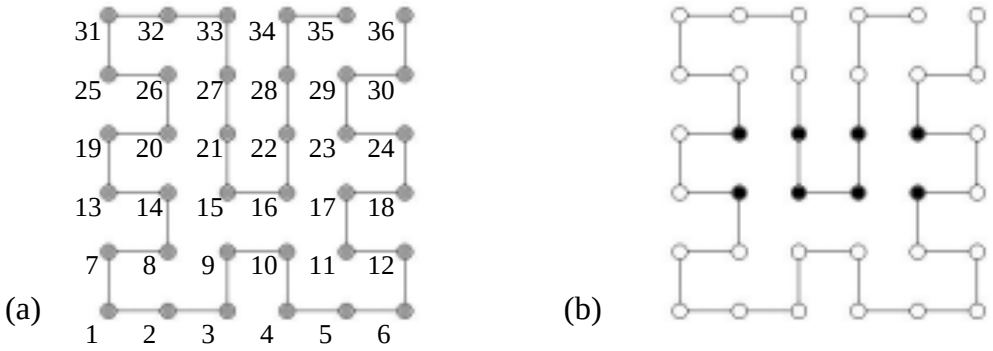

Fig. 4. Sequence design example with 6×6 HP lattice for eight H residues out of 36: (a) initial guess given to the optimization algorithm where all sites are uniformly in an intermediate state; (b) optimum solution with energy equal to –18.1. Circles filled with black, white and gray indicate H, P, and intermediate states, respectively.

It can be seen that the values of the variables (shown above in bold type) at the sites occupied by H residues (see Figure 4(b)) are close to zero.

If N_H is only two, the result shown in Figure 5(a) was obtained. With N_H as 17, the sequence shown in Figure 5(b) was obtained. After examining Figure 5(a), it is natural to question why the optimization algorithm chose sites 21 and 22 instead many other such pairs with energy equal to –4.3. Some of those choices are: {27,28}, {20,14}, {23,17},{26,27}, {28,29}, {14,15}, {16,17}, etc. All of these have one H–H interaction and two H–P interactions making the total energy equal to –4.3. Hence, all of these are local minima for this problem. Yet, the algorithm chooses sites 21 and 22 from an unbiased initial guess (Figure 4(a)), which happen to be the highest scored sites using the graph spectral method that is described in the next section (see Figure 7). The correlation with the scoring is also true for the 17-residue case shown in Figure 5(b). Extending this argument further, sites 20 and 23 that remained in the gray state for the case of three residues (Figure 5(c)) is due to the fact that they both happen to have equal scores. It is best to explain these phenomena (which are related to an unusual characteristic of a local method giving global or more robust minimum) after the graph spectral scoring method is presented in the next section.

5. Priority Ranking of Lattice Sites Using a Graph Spectral Method

The optimization method with continuous modeling just presented solved the problem of eight H residues in a 6×6 lattice in insignificant time of only a few seconds. To compare with the time it would have taken for exhaustive enumeration, note that this case has a total number of sequences equal to (36!)/(28!*8!) = 30,260,340. Out of these, the one with the minimum energy was easily found. There is an even more powerful method that has been applied to sequence design by Sanjeev, Patra, and Vishveshwara (2001). It will be briefly presented here (see Vishveshwara, Brinda, and Kannan 2002 for a detailed review) with some additional insight from the engineering viewpoint.

Consider again the 6×6 lattice of a 36-residue model protein shown in Figure 6(a) with the same conformation as in Figures 4 and 5. As mentioned before, only interactions between non-bonded immediate neighbors are considered in this paper. Thus, Figure 6(b) shows all such interactions as arrays of dashes. These interactions can be alternatively represented as a graph, as shown in Figure 6(c). It consists of 11 disconnected segments labeled A–K. The vertices indicate the sites in the protein model. The edges connect each pair of sites that interact with each other. Three are simply one-vertex graphs with no edges. These correspond to the three corners in the lattice model bonded to two immediate neighbors and hence have no interactions. There are six two-vertex graphs. There is a slightly longer segment with seven vertices. The longest segment has 14 vertices. An *adjacency matrix* **A** of size 36×36 can be constructed for this graph as follows:

$$A_{ij} = \begin{cases} 1 & \text{if vertices } i \text{ and } j \text{ are connected} \\ 0 & \text{otherwise} \end{cases} . \qquad (18)$$

Next, a diagonal matrix, called the degree matrix **D**, is defined. The diagonal elements in **D** are simply the sums of the corresponding rows in **A**. In other words, the ith diagonal entry in **D** represents the number of interactions of the ith vertex. Thus, **D** indicates the degree or the weight of a vertex. The more the weight, the more important that vertex is. Consider now $\mathbf{L} = \mathbf{D} - \mathbf{A}$, which is known as the Laplacian matrix in graph theory. The eigenanalysis of **L** reveals important information as described next.

If there are n disconnected segments in the graph, there will be $n+1$ zero eigenvalues for **L**. The eigenvectors corresponding to the eigenvalues just above zero indicate the segmenta-

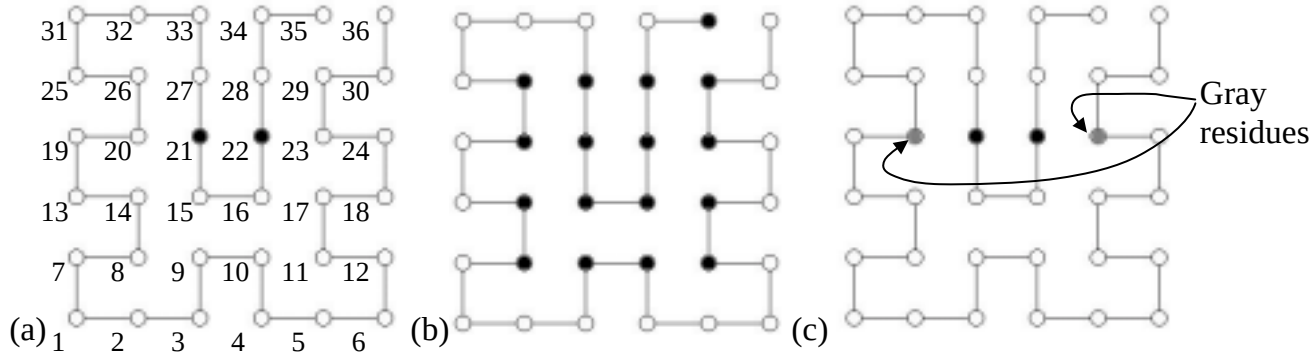

Fig. 5. Optimum HP sequences (a) for two, (b) for 17, and (c) for three H residues.

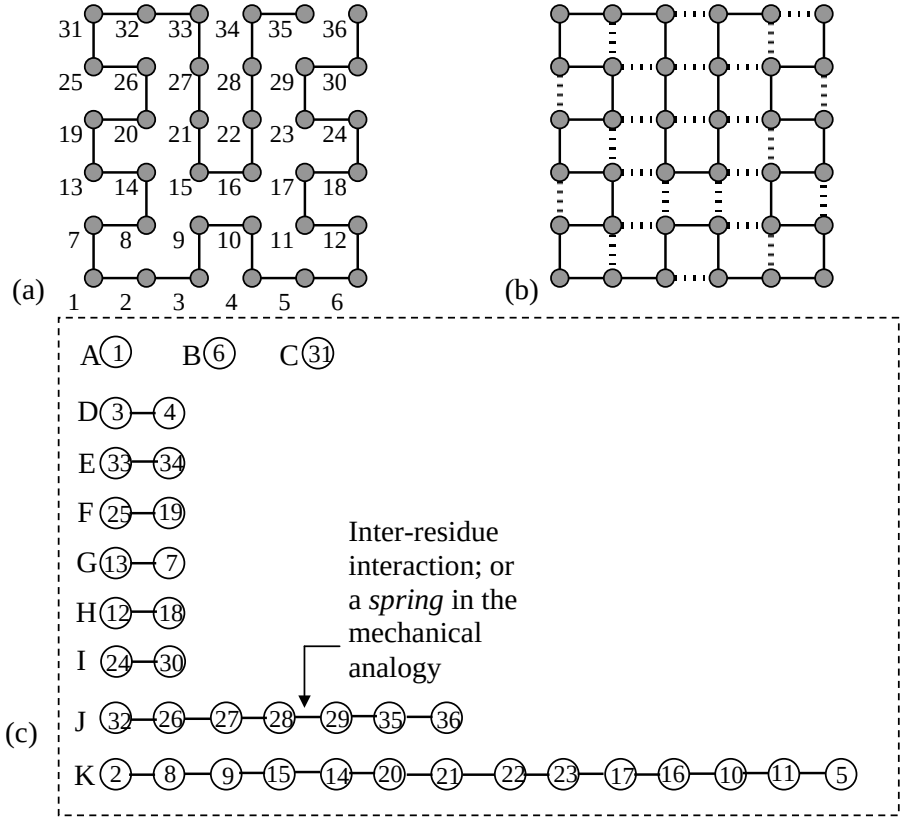

Fig. 6. (a) A CSA 6×6 lattice model protein. (b) Non-bonded inter-residue interactions shown with arrays of dashes. (c) A graph representation of non-bonded interactions.

tion of the graph. That is, an eigenvector will have non-zero values only for those positions that correspond to the vertices in a certain segment. For example, if a graph has two segments, the eigenvector with lowest non-zero eigenvalue will have non-zero values in only those positions that correspond to the first segment. Similarly, the next one may correspond to the second segment. Thus, the analysis of eigenvectors of low eigenvalues reveals segmentation. This can be easily understood by imagining the graph in Figure 6(c) as a system of masses connected by springs and translating along a straight line, where each edge indicating an interaction is a spring of unit spring constant. Then, the Laplacian matrix **L** is nothing but the stiffness matrix (Belegundu and Chandrupatla 2002) of the system, which was also noted by Keskin et al. (2002) and Bahar (1999). If it is a single connected graph, the entire system can translate like a rigid body in one dimension since no mass is fixed. This makes **L** have a rank deficiency of unity. If there are n disconnected segments, **L** will be rank deficient by $n + 1$. To see which mass is more important in terms of its interactions with its neighbors, one thinks of how many masses are perturbed and by what amounts if one mass is given a unit displacement. Clearly, if the masses corresponding to 1, 6, and 31 (in segments A, B, and C, respectively) are perturbed, they have no influence on any other. They are least important and ought to score the lowest. Masses with high influence on many others can be identified from the eigenvectors corresponding to the few highest eigenvalues. This explanation follows.

The eigenvector components [*Hvc*] of the highest few eigenvalues [Hev] contain much more information pertinent to the importance (weight or a score) of a vertex. The connectivity (or the number of non-bonded interactions) of each vertex can be identified from the *Hvc* in general because the vertices that have the high degree (as given by **D**) have high *Hvc*. However, it has been shown that the vertex of high *Hvc* should not necessarily be the one having high connectivity. Hence, the weight in the Sanjeev, Patra, and Vishveshwara (2001) study is constructed by adding the degree of vertex to the *Hvc* of that vertex. The vertices of the same degree are distinguished by their position in the graph. This method of weight assignment works very well for structures represented by a single connected graph. However, when the conformation is represented by multiple disconnected segments, the degree and the *Hvc* of the vertex do not provide a good estimate for its connectivity. The *Hvcs* of each disconnected graph segment may show only the connectivity of the vertex in the concerned segment. One also needs to consider the size of each segment since larger segments have connections with the vertices of higher degree. Therefore, the weight is determined in correlation across the segments so that it takes into account both the size of the segment and the connectivity from the nature of the segment itself. The size of the segment is taken into account by scaling *Hvc* by the fraction of total vertices constituting it. The vectors *Hvcs* are also scaled by Hev of each segment to retain the nature of the graph. The highest eigenvalue of the graph depends on the highest degree in the graph. Therefore, the weight W_i for each vertex is evaluated as follows

$$\begin{aligned} W_i &= W_{ic} + W_{ie} \\ W_{ie} &= e_i \times f_i \times Hvc_i \end{aligned} \tag{19}$$

where W_{ic} is the connectivity of the ith vertex, f_i is the ratio of the number of vertices in the segment to which vertex i belongs to the total number vertices in that segment, e_i is the Hev of the concerned segment, and Hvc_i is the vector component of the vertex i.

The method just described is a type of graph spectral method. The weights computed using eq. (19) were used to rank the sites in the lattice (Sanjeev, Patra, and Vishveshwara 2001). The sites with high weights are preferred candidates for H residues in view of larger importance given to the H residues in the assumed energy levels as in eq. (1). If there were only one H residue in the entire chain, it would go to the site with the highest rank. If there were two, they would go to the top two sites, and so on. Thus, with little computation (just eigenanalysis and computation of the weights), this method can give a good sequence for any composition of H and P residues.

As an example, for the conformation of the 6×6 lattice in Figure 6(a), the weights and rank numbers are given in Table 1 and ranked sites are pictorially depicted in Figure 7. This example helps find correlation between the results obtained with the graph spectral and optimization methods.

5.1. Relationship Between the Ranking and the Optimization Methods

Using Figure 7, it is easy to visualize which sites the H residues ought to occupy when only a few are available. It is interesting to note that when only two residues are allowed, just as this ranking method, the optimization method also preferred the same sites (compare Figures 5(a) and 7). The same is true for the case of 17 residues (compare Figures 5(b) and 7). More interesting than this is that when, say, three H residues were given, optimization let two residues remain in the gray state (see Figure 5(c)). The reason for this is clear from Table 1 where equal weights are enclosed with dashed boxes. This means that all the residues within the box are equally good. That is, all of them contribute equally to the energy minimization. Hence, the optimization algorithm is unable to decide which sites to choose for assigning the H state and leaves them in the gray state. Thus, the optimization method and the graph spectral method not only give the sequence with minimum energy but also indicate many equally good possibilities. This is useful in view of the need for the designed sequence also to be a minimum in the conformation space (see Section 2) because more candidates are made available by either of the two methods. It also helps identify the degeneracy of a given conformation. Another important feature of

Table 1. Priority Ranking of 6×6 Lattice Sites Using the Graph Spectral Method for the Conformation in Figure 6(a)

Rank #	Site #	Weight	Rank #	Site #	Weight	Rank #	Site #	Weight
1	21	2.7693	13	10	2.4119	25	18	1.1048
2	22	2.7693	14	26	2.3285	26	19	1.1048
3	20	2.7308	15	35	2.3285	27	24	1.1048
4	23	2.7308	16	8	2.2556	28	25	1.1048
5	14	2.6554	17	11	2.2556	29	30	1.1048
6	17	2.6554	18	32	1.1172	30	33	1.1048
7	15	2.5475	19	36	1.1172	31	34	1.1048
8	16	2.5475	20	3	1.1048	32	2	1.1048
9	28	2.5268	21	4	1.1048	33	5	1.1048
10	27	2.4747	22	7	1.1048	34	1	0
11	29	2.4747	23	12	1.1048	35	6	0
12	9	2.4119	24	13	1.1048	36	31	0

Note. Weights are also shown. Sites with equal weights are enclosed in dashed boxes.

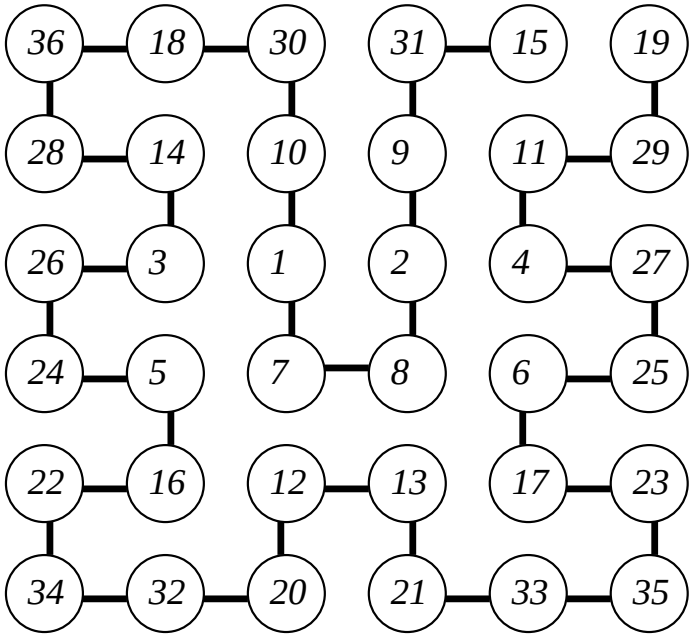

Fig. 7. Ranking of the sites in the 6×6 lattice (italic numerals show the rank number).

the two methods is that they are able to identify the most robust energy-minimizing sequences as in Figure 5(a). As mentioned before, even though many pairs of two residues with the same minimum energy of –4.3 are there, sites 21 and 22 are chosen over the rest. This is because these sites are ranked highest in terms of their interactions with the neighbors. Further explanation of this is given in Section 7.

Computationally, the graph spectral method is much more efficient than the optimization method. The former uses only one eigenanalysis of an $N \times N$ matrix and with that solves all cases of compositions of H and P residues. Sanjeev, Patra, and Vishveshwara (2001), who developed and applied this method to lattice models of proteins, have noticed that the method does not always give the lowest energy sequence, although it comes very close. That is, it sometimes gives sequences with slightly more energy than the lowest possible energy for a conformation. In the next section, the reason for this is identified, which also paves the way for combining this method with the optimization method to yield a combined method that is superior to either method.

6. Combined Graph Spectral and Optimization Method

In this section, first we present the significance of accounting for the magnitudes of interactions (or the weights of edges in the terminology of graph theory) in the graph spectral method and its implication in identifying sequences with a slightly lower energy. In the subsequent subsection, a description follows to explain how both the graph method and the optimization method could be improved by combining them together into one method. The motivation for combining the two methods is threefold. First, it improves the computational efficiency because the graph spectral method is superior to the optimization method in terms of computation time when the number of residues is large. Secondly, the results of the graph method provide a good initial guess to the optimization method, which is a local method and can thus overcome the usual problem of getting trapped at an improper local minimum. Thirdly, the graph method might occasionally miss a few minima with the lowest energy as explained next.

6.1. Why Does The Graph Spectral Method Not Always Identify The Lowest Energy Sequence?

In order to understand why sometimes the graph spectral method is unable to identify the lowest energy sequence, consider the case of the 6×6 lattice conformation for nine H residues. The results for this are shown in Figures 8(a) and (b). The energy of the sequence resulting from the graph spectral method is –20.1, which is slightly higher than the energy of –20.4 by the optimization method. Eight H-residue sites are common to both methods. The ninth site for the graph spectral method is 28, while 27 is equally good as they both have the same weights (see Table 1). The ninth H residue in Figure 8(b), which is the result of the optimization method, is shared more or less equally by sites 9 and 10 as indicated by their gray level state. Referring to Table 1 again, it can be seen that 9 and 10 also have equal weights. The main question here is why the two methods gave dissimilar results with the graph spectral method giving a sequence with energy a little higher than the lowest possible. The reason for this is rooted in the principle behind the graph spectral method as explained below.

The graph spectral method is based purely on the number of inter-residue interactions and not the type of interactions. Hence, the ranking of residues takes place to maximize the number of interactions rather than minimizing the energy. It can be seen that the arrangement in Figure 8(a) has seven H–H interactions and four H–P interactions—a total of 11 interactions. Hence, its energy is given by $7(-2.3) + 4(-1) = -20.1$. This is not the best choice from the energy minimization perspective in view of the assumed energy levels of eq. (1). When only H residues are involved, this method would have chosen all these sites except 28 with seven H–H interactions and two H–P (a total of nine) interactions. To minimize the energy most, the choice for the ninth one is 9 or 10. Placing an H residue at 9 or 10 replaces one H–P interaction (between sites 15 and 9 or 10) with a new H–H interaction, and a new H–P interaction (between sites 9 or 10, and 8 or 11), accompanied by a decrease in energy of 2.3. The total number of interactions now is 10 (eight H–H and two H–P). On the other hand, placing the ninth residue at 28 or 27 adds two H–P interactions with decreases in energy by only 2, although its total number of interactions is 11. Consequently, the choice preferred by the optimization method fares better by 0.3 over the graph spectral method. Hence, maximizing the number of interactions without regard to the type is not always preferable.

While the above simple example illustrates the point, this problem of identifying sequences with slightly higher energy becomes more pronounced as the lattice size (or the number of residues in real proteins) becomes larger. This effect is even more pronounced with 3D lattices and hence important because proteins are after all 3D structures. To show this, cubic HP lattices of size 3, 4, 5, and 6 were analyzed for an arbitrarily chosen, but fixed throughout the analysis, conformation for each case. Figures 9(a)–(d) show the improvement in energy (that is, decrease in energy) that could be achieved for different numbers of H residues in each of the four cases. It should also be noticed that the energy improvement is in steps of 0.3. The reason for this is clear from the above explanation. While the 3-cubic lattice goes up to only 0.6, the other cases reach 1.5, 2.1, and 2.8, respectively. Furthermore, as the size increases, the improvement happens in more and more cases of the number of H residues. Real proteins, which are usually long chains, are likely to show much more significant effect.

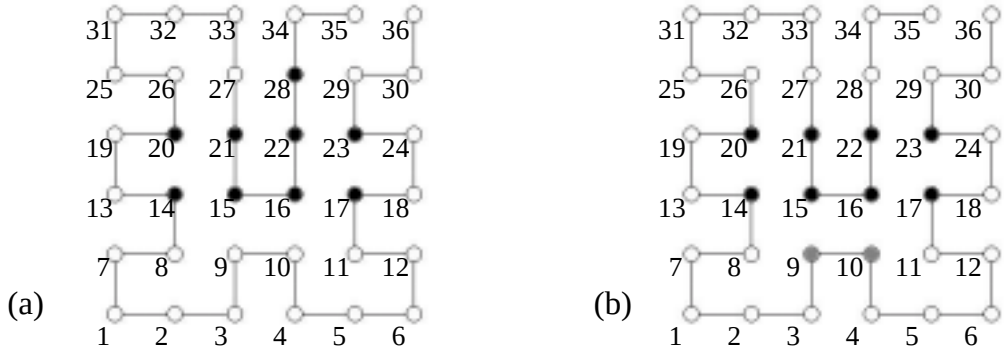

Fig. 8. Positions of nine H residues identified for minimum energy by (a) the graph spectral method and (b) the optimization method. Site 27 is as good as 28 and hence the choice between either of the two is arbitrary and makes no difference to the energy. Two residues with equal contribution (9 and 10) are indicated in gray level by the optimization method. The dissimilarity in the results of the two methods is explained in the text.

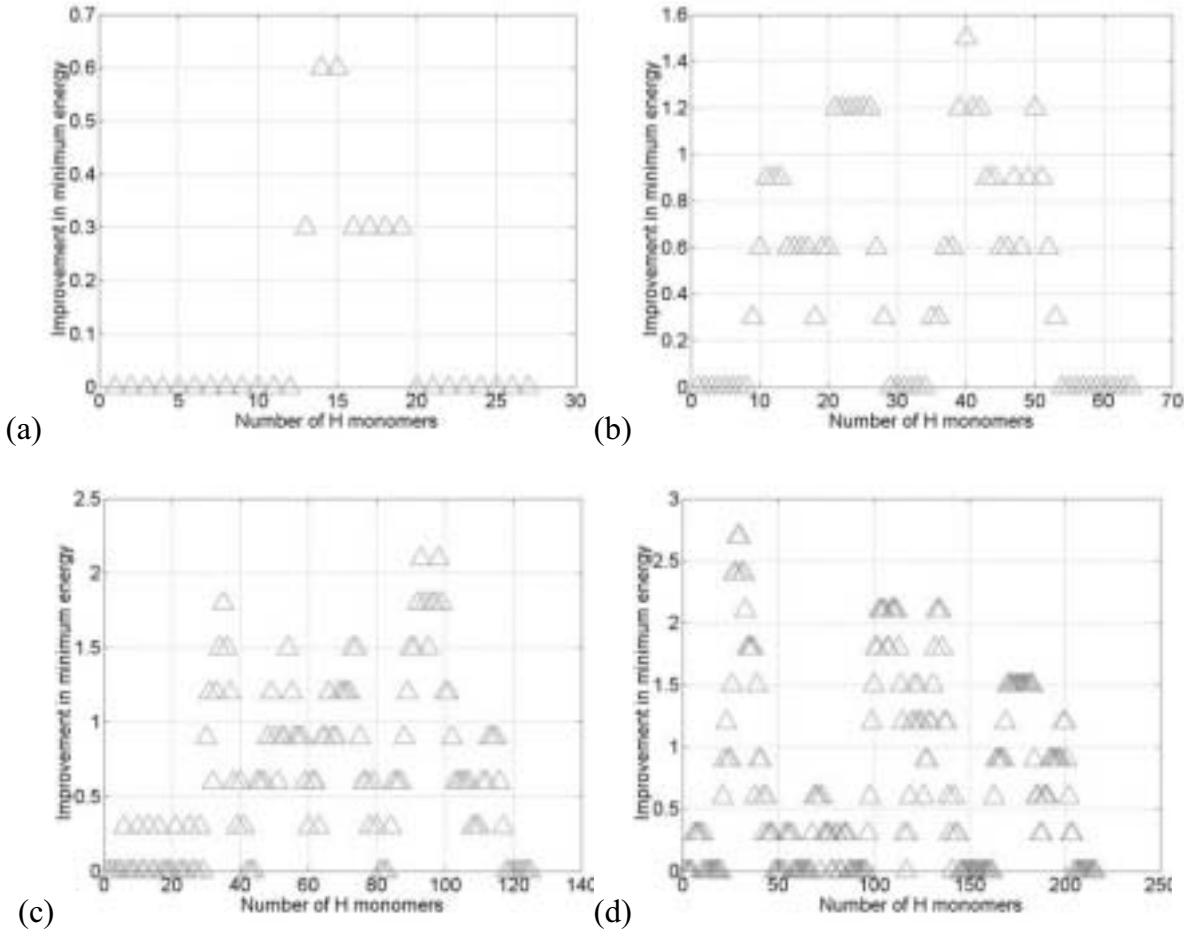

Fig. 9. The improvement in the energy beyond what the graph spectral method gives plotted against the number of H residues for (a) a $3\times3\times3$ lattice conformation 4 in Table 1 of Sanjeev, Patra, and Vishveshwara (2001), (b) a $4\times4\times4$ lattice, (c) a $5\times5\times5$ lattice, and (d) a $6\times6\times6$ lattice.

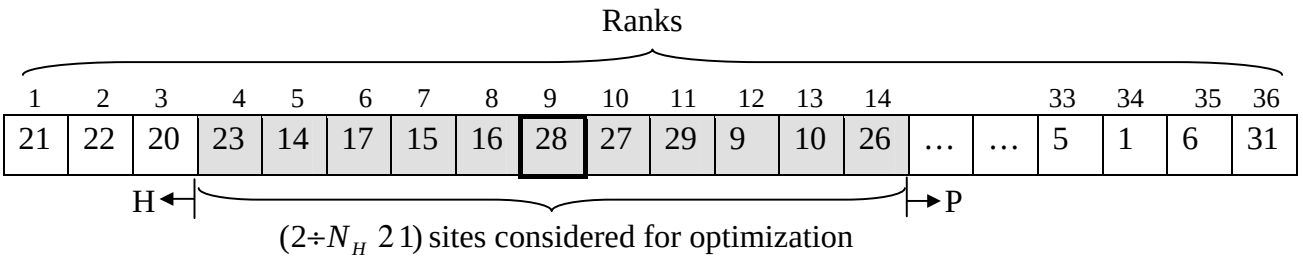

Fig. 10. Selection of sites in the combined method. If ΔN_H= 5, five sites on either side of the site ranked ninth (28) are considered for the case of nine H residues.

It should be noticed that the four cases in Figures 9(a)–(d) involved 27, 64, 125, and 216 residues in their chains. This means that even with H or P states, the respective sequence spaces have $2^{27} \approx 0.13E9$, $2^{64} \approx 1.84E18$, $2^{125} \approx 4.25E37$, and $2^{216} \approx 1.05E65$ sequences. Yet, the data shown in Figures 9(a)–(d) were obtained with relative ease (in about 3 h for all cases put together on a desktop in a Matlab environment). This computational efficiency is a result of combining the graph spectral method and the optimization method, which is the next topic.

6.2. The Combined Method

As noted earlier, the optimization method helps identify one or more sequences with minimum energy. However, it is not as efficient as the graph spectral method in terms of computation time when the size of the chain increases. On the other hand, the graph spectral method is likely to fail in some cases. Fortunately, the two can be combined by drawing upon their respective advantages. Consider a conformation for which the graph spectral method is used to come up with the ranks for all the sites. When a particular case of a residue composition is given, optimization can be performed locally around a selected point in the chain. To elaborate, consider the ranking given in Table 1 for the 6×6 lattice. If nine H residues are specified, a few sites around the ninth position in the ranked list of sites are considered for further scrutiny. Let the number of such sites on either side be denoted by ΔN_H. Then, as shown in Figure 10 where ΔN_H is equal to 5, five sites on either side of the site ranked ninth are considered. The sites above fourth rank are fixed to be in the H state. The sites below the fourteenth rank are fixed in the P state. Then, the remaining 11 sites are chosen for optimization. This helps reduce the size of the problem while being successful in identifying the sequence(s) with the least energy.

The success of the combined method is based on the observation that only local adjustments are necessary to the rankings given by the graph spectral method. That is, instead of creating two additional H–P interactions, one extra H–P interaction and the second replacing an existing H–P interaction with an H–H interaction is all that is necessary to improve the graph spectral method. This can happen in more than one way. So, ΔN_H should be sufficiently high. For the data in Figures 9(a)–(d), ΔN_H was taken to be 21. For a size of 21, using a pure enumerative scheme to determine the above adjustment is computationally inefficient. Hence, the optimization method is used. Next, an example is presented to validate the accuracy of the results obtained with the combined method.

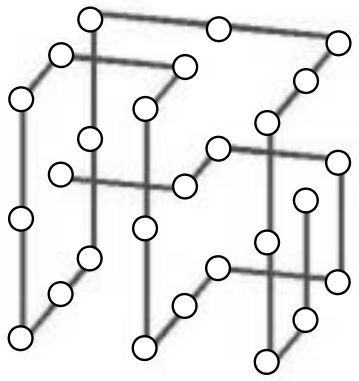

Fig. 11. The most designable conformation of the 3×3×3 HP lattice.

6.3. A Benchmark Example

Consider the conformation shown in Figure 11 for the 3×3×3 lattice. It is reported to be the most designable structure for this lattice (Li, Tang, and Wingreen 2002). This means that many sequences have this conformation as the native state. A sequence with the minimum energy was found for this for every case of the number of H residues over its entire possible range (0–27) using the graph spectral method as well as the combined method. For validation, all the sequences were enumerated, separately for each case of the different numbers of H residues, and the minimum energy and the sequence(s) that possess that value were found. It took several

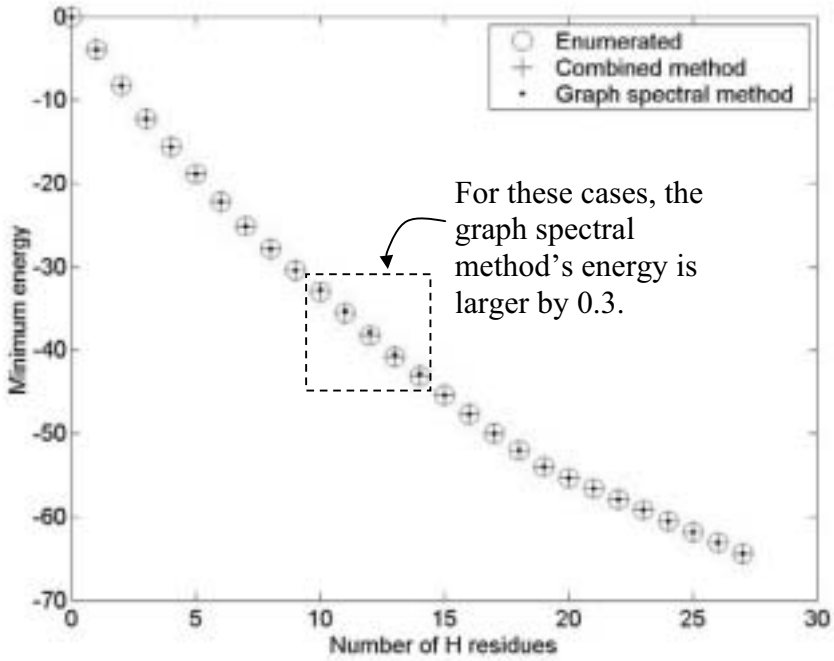

Fig. 12. Minimum energies of the 3×3×3 lattice's most designable conformation with three methods: circles indicate results from exhaustive enumerative scheme; plus symbols indicate results from the combined method; dots indicate results from the graph spectral method.

hours for each case because the number of sequences exponentially increases as the number of H residues approaches half of the total number of residues. In comparison, the combined method took only a few seconds (15 s at the most) for each case. The energies are plotted in Figure 12. The circles (o) indicate the values given by the enumerative procedure, and plus (+) symbols that of the combined method. It can be seen that the combined method never failed to give a sequence with the lowest energy. Furthermore, specific cases of these minimum energies agree with those reported in the literature (e.g., $N_H = 13$ in Zou and Saven 2001). For comparison, the dot symbols indicate the lowest energy given by the graph spectral method. This method too fares quite well but it does fail for five cases (N_H = 10–14), which have minimum energies larger by 0.3. This example confirms that the combined method is not only computationally efficient but also gives the correct results.

6.4. Examples With Real Proteins

It was noted earlier that the combined method was applied to very large lattices including the case of 6×6×6. In fact, the size of the chain does not limit its application as the size of the optimization problem depends only on the value of ΔN_H. Hence, the method can easily be applied to the models of real proteins. However, the limitation comes from the lack of accurate definition of the amino acid residues and their interaction energies. Consequently, the examples of real proteins are also confined to only two states (H and P). Thus, it is a demonstration that large chains and irregular lattice models are possible rather than verifying the sequences of real proteins.

First, the .pdb ASCII text file from the Protein Data Bank (PDB; Berman et al. 2000) was downloaded. This file was then parsed using a simple Matlab program written for this purpose. This program extracts the position coordinates of the C_α atoms in the chain along with the type of amino acid at each residue. Based on the HP categorization of Wang and Wang (1999), the 20 amino acid types were separated into H or P groups. Then, using a distance of 6.5 Å, interacting neighbors in the backbone were identified. This information was used to construct the adjacency matrix. As in other examples of this paper, only the non-bonded neighboring residues are assumed to have an interaction. Once again, the energy levels noted in eq. (1) were used.

Three real proteins were considered. The first is the phosphate-free ribonuclease A (PDB code: 7RSA), which has a single chain consisting of 124 residues folded into a single domain. Its folded conformation is shown in the Richardson's ribbon schematic representation in Figure 13(a) and its backbone is shown in Figure 13(b). The software program RasMol (Molecular Graphics Software; see http://www.RasMol.org) was used to obtain these renderings. The second one is the

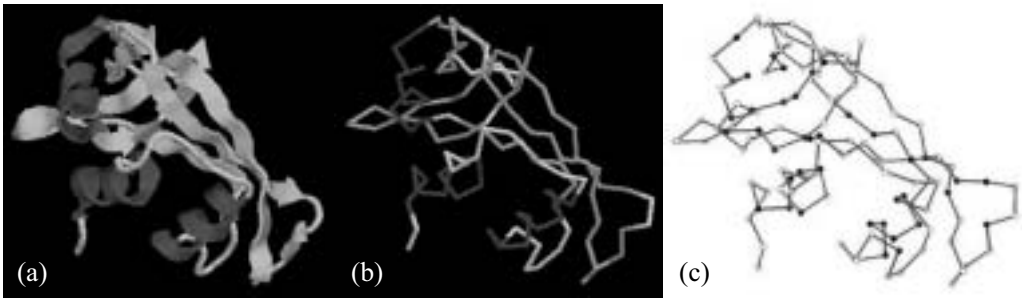

Fig. 13. HP sequence design of phosphate-free robonuclease A (7RSA) protein: (a) ribbon schematic; (b) backbone structure; (c) backbone with designed sequence of minimum energy. Black dots denote H residues and circles P residues.

triosephosphate isomerase complex with sulphate (5TIM), which has two chains labeled A and B (see Figures 14(a) and (b)), each consisting of 249 residues. Only chain A was considered for sequence design with the HP model. The third is the tobacco ring spot virus (1A6C), which is a single-chain folded into multidomain protein consisting of 513 residues (see Figures 15(a) and (b)). The results of the HP sequence design on these with the combined method are presented next.

According to the assumed HP categorization, there were 39 H residues in 7RSA HP model out of 124. With ΔN_H equal to 25, the sequence for a minimum energy given by the combined method is shown in Figure 13(c). The black dots denote the H residues and the circles P residues. The rendering in this figure is from Matlab program written for this purpose. This is shown in approximately the same perspective as the one in RasMol's rendering of Figures 13(a) and (b). The energy due to this sequence is –325.85. Increasing ΔN_H up to 30 did not make any difference to the result.

The HP model of 5TIM has 115 H residues out of 249. Its result is shown in Figure 14(c). The energy of this sequence is –977.05. The value of ΔN_H was 25. Increasing this value up to 50 did not make any difference to the result. For 1A6C, there are 215 H residues out of 513. The obtained sequence (shown in Figure 15(c)) has a minimum energy equal to –1757.95. The value of ΔN_H above 50 did not make any difference to the result.

The computation time depends on the value of ΔN_H because it determines the size of the optimization problem. Specifically, there will be $2\,(\Delta N_H) + 1$ variables. The computation time was less than 5 min even for the case of ΔN_H = 75. For comparison, the minimum energies given by the graph spectral method were noted in each case. They were larger by 2.1, 1.5, and 10.2 in the cases of 7RSA, 5TIM, and 1A6C, respectively.

7. Discussion

The question concerning why the optimization method converges to some sequences skipping some others with equal minimum energy (refer to Figure 5(a)) was not fully answered until now; it was only correlated with the ranking given by the graph spectral method. Now, a more intuitive explanation will be given and its role in making the local optimization method used here apparently globally robust.

The reason for the global robustness of the optimization method stems from the diffuse interpolation of the state of sites between 0 and 1 using the continuous model. At the beginning of optimization, the initial guess (Figure 4(a)) has all sites in between 0 and 1. This has the effect of smoothening out the objective function and, in the process, hiding the unimportant local minima. In the specific example considered (Figure 5(a)), the sites in the other pairs cease to be the preferred H sites when the number of H residues increases. Hence, it may be concluded that they happen to lie in shallow regions of the energy landscape. This argument is pictorially illustrated in Figure 16(a). In this two-dimensional representation of the energy landscape the x-axis in the plot is the "sequence axis" and the y-axis is the energy axis. For small values of σ, the interpolation function has a sharp shape and there is a distinct separation between the state 1 and 0 (see Figure 2), whereas for larger σ values, there is a large portion of intermediate state due to the diffused function form. The solid line in Figure 16, imagined to be the energy landscape with a smaller value of σ, brings out the local minima clearly. However, with larger σ, the dashed line smoothens out the unimportant (shallow) ones leaving the minimum in deep funnel to prevail. Thus, this algorithm shows promise to choose the minimum preferred by a type of energy gap criterion. Or, in other words, more stable ones are chosen over less stable ones. Similarly, Figure 16(b) shows another situation where other local minima with higher energy values can also be skipped by the algorithm when optimization is begun with a large value of σ. A gradual decrease of σ, as the optimization process continues, makes the energy landscape sharper and makes the distinction between other local minima and the one obtained at convergence. Sometimes, this distinction may not be possible because all those local minima may have equal energy. In such a case, degenerate sequences will be identified. An example explains this further.

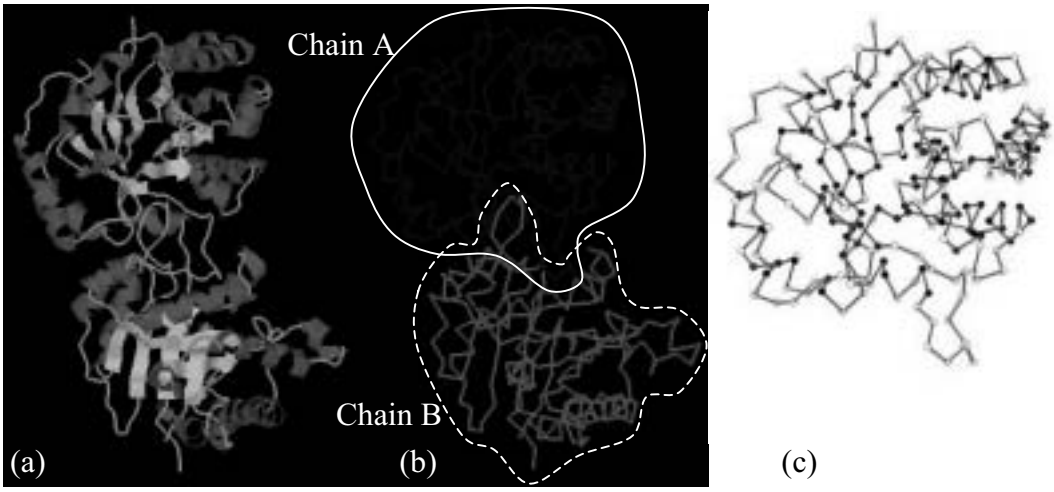

Fig. 14. HP sequence design of chain A of triosephosphate isomerase complex with sulphate (5TIM) protein: (a) ribbon schematic; (b) backbone structure; (c) backbone with designed sequence of minimum energy. Black dots denote H residues and circles P residues.

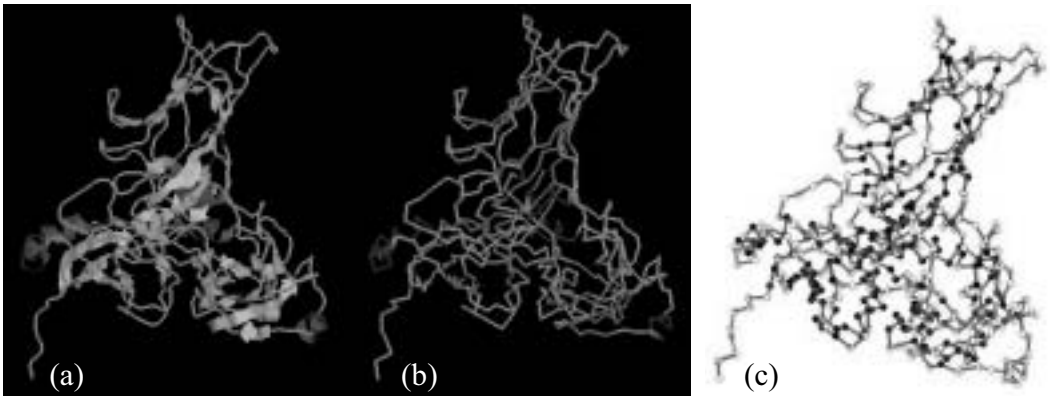

Fig. 15. HP sequence design of tobacco ring spot virus (1A6C) protein: (a) ribbon schematic; (b) backbone structure; (c) backbone with designed sequence of minimum energy. Black dots denote H residues and circles P residues

Consider the case of the $3\times3\times3$ HP lattice in the most designable conformation (Figure 11) with nine H residues. For the purpose of validation, this case was exhaustively enumerated and five degenerate sequences shown in Table 2 were found. They all have the energy equal to -30.4. The sequences given in Table 2 are separated into three parts. The left and the right parts are the same for all, but the italicized middle part is different. The left and right contain eight H residues, and middle part contains the remaining ninth residue. The ninth residue in each of the five cases is underlined. For this problem, the combined method gave the result shown in Figure 17. It can be seen that there are five residues in the gray state. By examining the data in Table 2, it can be observed that the method identified all but the site with the tenth residue in the chain. Thus, four out of the five degenerate sequences are identifiable with this method. The reasons for why the fourth site did not come up as definitively H site and why the tenth residue came out as definitively P site are not clear, but it may be because the sequence resulting with it may be isolated from the rest in the sequence space. Notwithstanding minor uncertainties such as this (which will be explored in the future work), the method shows the ability to find at least some of the degenerate sequences.

7.1. Principal Features of the Combined Method

While running the examples with only the optimization method (i.e., not combining it with the graph spectral method),

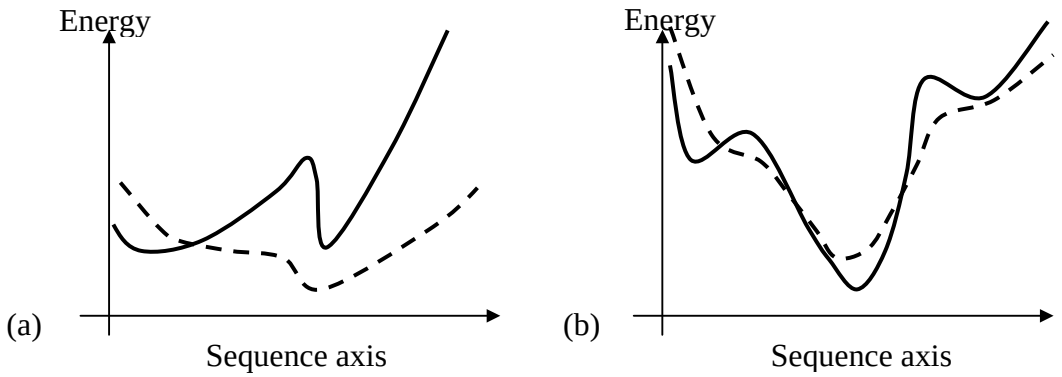

Fig. 16. Hypothetical energy landscapes to understand the effect of the continuous modeling of the states of the residue sites. Dashed curves are due to the smoothening effect. (a) Illustration of how the continuous formulation might hide a shallow minimum of energy is equal to that of the chosen one. (b) The case of many shallow local minima disappearing in the smoothened model.

Table 2. Five Degenerate Minimum Energy Sequences for the HP Model of the Most Designable Conformation of the 3×3×3 Lattice

Sequence Number	Sequence	Number in the Chain of the Site in the Gray State
1	HPPHP *H*PPPPPPPP PPHPHPHPPPHHH	4
2	HPPHP *PH*PPPPPP PPHPHPHPPPHHH	6
3	HPPHP *PPPH*PPPP PPHPHPHPPPHHH	8
4	HPPHP *PPPPPH*PP PPHPHPHPPPHHH	10
5	HPPHP *PPPPPPPH* PPHPHPHPPPHHH	12

an interesting observation was made. Recall that the initial guess for the optimization method has all the sites uniformly in the same gray state. The first few iterations of optimization already indicate the resulting sequence as residues that will eventually be in the P state start to fade while the would-be H residues start to become darker. This is not totally surprising given that the continuous optimization method uses the gradient information, which points to the minimum right from the beginning in the form of the most suitable descent direction. In view of the smoothening effect due to the assumed (large) value of σ, a biophysical interpretation of it may be that minima that are kinetically easily accessible (those that are in the deep funnel regions of the energy landscape) are found rather than those in the shallow regions. At this point, it is only an observation and warrants further investigation and rigorous analysis.

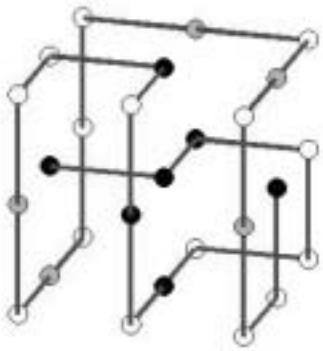

Fig. 17. The sequence given by the combined method for the most designable conformation of the 3×3×3 lattice for nine H residues. The sites in the gray state are discussed in the text.

Next, some of the other features of the combined method and some observations based on the results are summarized.

1. The combined method combines the computational advantage of the graph spectral method and the ability of the optimization method in giving the sequence(s) with minimum energy.

2. The method was validated for the much-studied $3\times3\times3$ HP lattice model by comparing with the results obtained with exhaustive enumeration.

3. When there is more than one sequence with equal minimum energy, the method identifies many such possibilities by leaving a few residues in the gray state. Given the ability of the graph spectral method in screening out some of the lowly ranked possibilities, the number of such residues is small, making it very easy to find all those possibilities.

4. The optimization method using continuous models is somewhat similar to the optimization method used by Zou and Saven (2000) where the variables are site-specific probabilities for each amino acid type. Their method is based on a mean field theory of statistical mechanics. In terms of computation, in their method there will be a number of variables associated with each site whereas in this paper each site has only one variable. Hence, the formulation presented in this paper gives rise to smaller optimization problem. Furthermore, the method of Zou and Saven considers all 20 types of amino acid residues and identifies only the probabilities of each site being occupied by a few most probable amino acid types rather than definitively assigning specific states to the sites. Which approach is better overall has yet to be investigated.

5. To apply the combined method, the backbone of the protein model considered need not be an orderly lattice; it can also be an irregular lattice including that of the real protein. The number of the domains too does not seem to matter, but more examples need to be solved to ensure this.

6. The computation time is dependent on ΔN_H and not on the total number of residues. This is an important feature of the method, which makes it easily scalable to very large proteins. The larger the value of ΔN_H, the longer the computation time. It is not significantly large, and hence the computation time will only be of the order of a few minutes. At this time, a number of values of ΔN_H are to be tried to make sure that as many good candidates with the least energy or close to it are found.

Some future extensions of this promising method are noted below.

7.2 Future Extensions

It may be possible that a few more good candidates that minimize the energy are found if the questionable sites (which contribute to the variables in the optimization method) are identified in more than one location in the chain. That is, the sites for optimization need not only be around the N_Hth site. Perhaps there is a benefit to consider other crucial points in the graph. The method of identification of clusters in a connected graph (Patra and Vishveshwara 2000) may be useful here. Another important extension is to consider more than two states, i.e., go beyond the HP models to the models consisting of all 20 amino acid types. There are two challenges associated with it. First, more types will make optimization difficult but not impossible. The multimaterial structural optimization methods developed by Yin and Ananthasuresh (2001) and level-set methods used by Vese and Chan (2002) and Wang, Wang, and Guo (2003) give useful clues. Secondly, the energy models used for 20 amino acid types need to be reliable because in that case comparison will be made with the real proteins. Statistical analysis based models (Miyazawa and Jernigan 1985) are available, but if fewer than 20 types are considered as a stepping stone to extend this method from its current HP modeling, energy models of reduced residue types are necessary. A few attempts in this direction have already been reported (Wang and Wang 1999; Chan 1999). Finally, the continuous model based optimization should also be able to search in the structure space in addition to the sequence space. All of these, and more, constitute some future investigations along the lines of the method presented in this paper.

8. Conclusions

A novel continuous modeling of the discrete, combinatorial protein sequence design problem is presented in this paper. Towards this, a continuous state function based on the Gaussian distribution function is used to interpolate between the discrete states of amino acid residue sites. This resulted in a continuous energy function, which is minimized to find the sequences with the least energy. A gradient-based optimality criteria method is implemented to find the optimal sequences. The continuous modeling and the optimization method resulted in significant reduction in the computation time compared to other existing techniques that rely upon explicit realization of some or all sequences. Other advantages of this method are noted: its apparent ability to find most stable sequences that are kinetically easily accessible and to identify more than one sequence when there are many good candidates with equal or nearly equal minimum energy. For the purpose of illustration and for validation of results through exhaustive enumeration, two-state (H and P) lattice models were used. Then, a previously reported graph spectral method was reviewed and interpreted from a mechanical engineering perspective. The priority rankings for sites in a given

conformation given by the graph method agree with the results of the continuous optimization method. Thus, the protein topology information obtained as ranks of the nodes from the graph spectral method can be effectively combined with the optimization techniques. Furthermore, the results have emphasized the fact that although the graph spectral method is powerful in obtaining topology based ranking of the vertices, the weights of the vertices and edges should be used for accurate energy evaluation. This led to a fortuitous combination of the optimization and graph spectral methods, which was discussed in detail along with several illustrative examples including three real proteins (PDB codes: 7RSA, 5TIM, and 1A6C). Only H (hydrophobic) and P (polar) states are considered in this paper. The combined method took only a few minutes (10 min being the upper limit and often only 2 or 3 min) in a Matlab environment for HP models of proteins as long 513 residues. Since the current optimization method can efficiently solve the problems with only two different states, the proposed methodology has been demonstrated based on the HP lattice protein models. In order to extend this method to the real protein sequence design with more than two monomer types and more elaborated realistic energy models, an efficient optimization method for more than two different states needs to be developed. In principle, the methodology proposed in this paper can be easily extended to the protein sequence design using realistic energy models. Methodology and results in this direction will be presented in future publications.

Acknowledgments

Helpful discussions with Professor Jeffrey Saven of the Chemistry department at the University of Pennsylvania are gratefully acknowledged. He pointed SKK and GKA to lattice models of proteins and relevant papers in the literature at the beginning stages of this research.

References

Anfinsen, C. 1973. Principles that govern the folding of protein chains. *Science* 181(4096):223–230.

Bahar, I. 1999. Dynamics of proteins and biomolecular complexes: inferring functional motions from structure. *Reviews in Chemical Engineering* 15(4):319–347.

Banavar, J. R., and Maritan, A. M. 2003. Geometrical approach to protein folding: a tube picture. *Reviews of Modern Physics* 75:23–34.

Belegundu, A., and Chandrupatla, T. 2002. *Introduction to Finite Elements in Engineering*. Prentice-Hall, Upper Saddle River, NJ.

Bendsøe, M. P., and Sigmund, O. 1999. Material interpolations in topology optimization. *Archive of Applied Mechanics* 69:645–654.

Bendsøe, M. P., and Sigmund, O. 2003. *Topology Optimization: Theory, Methods, and Applications*. Springer-Verlag, Berlin.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. 2000. The Protein Data Bank. *Nucleic Acids Research* 28:235–242.

Chan, H. S. 1999. Folding alphabets. *Nature Structural Biology* 6(11):994–996.

Chothia, C. 1992. One thousand families for the molecular biologist. *Nature* 357:543–544.

Desjarlais, J. R., and Handel, T. M. 1995. De novo design of the hydrophobic cores of proteins. *Protein Science* 4:2006–2018.

Desmet, J., De Maeyer, M., Hazes, B., and Lasters, I. 1992. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* 356:539–542.

Deutsche, J. M., and Kurosky, T. 1996. New algorithm for protein design. *Physical Review Letters* 76(2):323–326.

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 Science* 4:561–602.

Go, N. 1983. Theoretical studies of protein folding. *Annual Review of Biophysics and Bioengineering* 12:183–210.

Haftka, R. T., and Gürdal, Z. 1992. *Elements of Structural Optimization*. Kluwer Academic, Amsterdam.

Hellinga, H. W., and Richards, F. M. 1994. Optimal selection of sequences f proteins of known structure by simulated evolution. *Proceedings of the National Academy of Sciences* 91:5803–5807.

Keskin, O., Bahar, I., Flatow, D., Covell, D. J., and Jernigan, R. L. 2002. Molecular mechanisms of Chaperonin GroEL-GroES function. *Biochemistry* 41(2):491.

Koh, S. K., and Ananthasuresh, G. K. 2004. Design of HP models of proteins by energy gap criterion using continuous modeling and optimization. *Proceedings of the ASME IDETC 2004 Mechanisms and Robotics Conference*, Salt Lake City, UT, September.

Lau, K. F., and Dill, K. A. 1989. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules* 22:3986.

Li, H., Helling, R., Tang, C., and Wingreen, N. S. 1996. Emergence of preferred structures in a simple model of protein folding. *Science* 273:666–669.

Li, H., Tang, C., and Wingreen, N. S. 1997. Nature of driving force for protein folding: a result from analyzing the statistical potential. *Physical Review Letters* 79(4):765–768.

Li, H., Tang, C., and Wingreen, N. S. 2002. Designability of protein structures: a lattice-model study using the Miyazawa–Jernigan matrix. *Proteins: Structure, Function, and Genetics*. 49:403–412.

Miyazawa, S., and Jernigan, R. L. 1985. Estimation of effective inter-residue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* 18: 534–552.

Pabo, C. 1983. Designing proteins and peptides. *Nature* 301(5897):200.

Park, S., Yang, X., and Saven, J. 2004. Advances in computational protein design. *Current Opinion in Structural Biology* 14(4):487–494.

Patra, S. M., and Vishveshwara, S. 2000. Backbone cluster identification in proteins by a graph theoretical method. *Biophysical Chemistry* 84:13–26.

Phillips, A. T., Rosen, J. B, and Dill, K. A. 2001. Convex global underestimation. *From Local to Global Optimization,* Migdalas, A., editor. Kluwer Academic, Dordrecht, pp. 1–18.

Ponder, J. W., and Richards, F. M. 1987. Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. *Journal of Molecular Biology* 193:775–791.

Rao, S. S. 1996. *Engineering Optimization: Theory and Practice*. Wiley Interscience, New York.

Sali, A., Shakhnovich, E., and Karplus, M. 1994. Kinetics of protein folding: a lattice model study of the requirements for folding to the native state. *Journal of Molecular Biology* 235:1614–1636.

Sanjeev, B. S., Patra, S. M., and Vishveshwara, S. 2001. Sequence design in lattice models by graph theoretical methods. *Journal of Chemical Physics* 114(4):1904–1914.

Saxena, A., and Ananthasuresh, G. K. 2000. On an optimal property of compliant topologies. *Structural and Multidisciplinary Optimization* 19(1):36–49.

Saven, J. G., and Wolynes, P. G. 1997. Statistical mechanics of the combinatorial synthesis and analysis of folding macromolecules. *Journal of Physics and Chemistry B* 101:8375–8389.

Shaknovich, E. I., and Gutin, A. M. 1993. Engineering of stable and fast-folding sequences of model proteins. *Proceedings of the National Academy of Sciences, USA*, 90:7195–7199.

Vese, L. A., and Chan, T. F. 2002. A multiphase level set framework for image segmentation using the Mumford and Shah model. *International Journal of Computer Vision* 50(3):271–293.

Vishveshwara, S., Brinda, K. V., and Kannan, N. 2002. Protein structure: insights from graph theory. *Journal of Theoretical Computational Chemistry* 1(1):187–211.

Wang, J., and Wang, W. 1999. A computational approach to simplifying the protein folding alphabet. *Nature Structural Biology* 6(11):1033–1038.

Wang, M. Y., Wang, X. M., and Guo, D. M. 2003. A level set method for structural topology optimization. *Computer Methods in Applied Mechanics and Engineering* 192(1):227–246.

Yin, L., and Ananthasuresh, G. K. 2001. Topology optimization of compliant mechanisms with multiple materials using a peak function material interpolation scheme. *Structural and Multidisciplinary Optimization* 23(1):49–62.

Yin, L., and Ananthasuresh, G. K. 2002. Novel design technique for electrothermally actuated compliant micromechanisms. *Sensors and Actuators A* 97/98:599–609.

Yin, L., and Yang, W. 2001. Optimality criteria method for the topology optimization under multiple constraints. *Computers and Structures* 79:1839–1850.

Yue, K., and Dill, K. A. 1992. Inverse protein folding problem: designing polymer sequences. *Proceedings of the National Academy of Sciences, USA* 89:4163–4167.

Zou, J., and Saven, J. G. 2000. Statistical theory of combinatorial libraries of folding proteins: energetic discrimination of a target structure. *Journal of Molecular Biology* 296:281–294.