Continuous Optimization Method for the Sequence Design of Protein Models Consisting of Multiple Monomer Types

S. K. Koh[*], G. K. Ananthasuresh[*,†], X. Yang[‡], and J. Saven[‡]

[*]Mechanical Engineering and Applied Mechanics, University of Pennsylvania, Philadelphia, PA, USA
[†]Mechanical Engineering, Indian Institute of Science, Bangalore 560012, India,
suresh@mecheng.iisc.ernet.in
[‡]Chemistry, University of Pennsylvania, Philadelphia, PA, USA

ABSTRACT

In our earlier work, we proposed continuous modeling of discrete H (hydrophobic) and P (polar) types of the amino acid monomers so that continuous and deterministic gradient based optimization methods could be applied to the originally combinatorial problem of protein sequence design for a given folded 3-D conformation. In this paper, we extend this approach to handle multiple monomer types so that real proteins with all 20 amino acid monomers could be designed. For this, two continuous variables are defined for each residue site. Treating the amino acid monomers as different states of the residue site, the two variables continuously interpolate the states. When optimization is carried out using these variables, the monomer type at each residue site is determined so that the energy for the desired folded conformation is minimized. The initial guess for the optimization method is determined using either our previously reported quadratic programming formulation or a statistical theory based method. The highlight of the approach presented here is the computational efficiency wherein the sequences of large, real proteins could be synthesized within minutes on a desktop computer. Verification of the global minimality of the results is shown using exhaustive enumeration of simplified models or known results.

Keywords: protein sequence design, deterministic optimization, continuous modeling of amino acid types, drug molecular design

1 INTRODUCTION

The sequence design of proteins, i.e., the inverse folding problem, entails the determination of the type of amino acid monomer at each residue site in the protein chain for a given or desired folded conformation. This is often done by finding the sequence with the lowest energy wherein energy interactions among non-bonded interacting neighboring residue sites are considered. Since each site can have anyone of the 20 amino acid monomer types, the sequence design leads to combinatorial explosion: a chain with 60 residues will have $20^{60} = 1.15\times10^{78}$ sequences, which is out of the scope of exhaustive enumeration. Many stochastic and discrete optimization methods have been applied to solve this problem [e.g., 1-3]. There also exist a few deterministic methods such as statistical mean field theory-based [4,5], mixed integer and linear programming [6], and graph spectral [7] methods. In general, well-formulated continuous optimization problems solved using gradient-based deterministic optimization algorithms lead to superior computational efficiency when compared with stochastic methods. Large-scale structural and mechanism optimization problems have been solved using such techniques in the engineering literature [8,9]. Motivated by these techniques, in our earlier work [10, 11], we introduced continuous modeling of discrete sequence space so that the combinatorial explosion of the sequence space is circumvented.

In [10], only two types of monomers (H: hydrophobic and P: polar) were considered for each site. By respectively assigning numerical values one and zero to H and P 'states' of every residue, the following 'state-interpolation' [12] was used.

$$S_i(\rho_i) = \exp\left\{-(\rho_i/\sigma)^2\right\} \qquad (1)$$

where the value of S_i continuously interpolates the state of i^{th} residue between zero and one as a function of the variable ρ_i. The tuning parameter σ determines the smoothness of the interpolation. For very small values of σ, S_i practically becomes a sharp selection between zero (P) and one (H) states but the problem remains continuous nevertheless. The energy interaction between a pair of sites i and j is then modeled continuously as follows.

$$e_{ij} = S_iS_je_{HH} + (1-S_i)S_je_{PH} + S_i(1-S_j)e_{HP} + (1-S_i)(1-S_j)e_{PP} \qquad (2)$$

When the sum of pair wise energy interactions, which is now a continuous function of variables ρ_i with $i = 1, 2, \cdots, N$, is minimized the minimizing set of values of ρ_i is obtained, which in turn, determines the HP sequence of amino acids. The ability of this method in finding the minimizing sequences was demonstrated for HP lattice models for which globally minimum energy is known via exhaustive enumeration or by other means. Even for long chains (as big as 513 residues), the computation time on a desktop computer was less than 10 minutes.

In [11], a linear interpolation of the states was used so that any number of monomer types (i.e., beyond H and P categorization) could be handled. The linear interpolation leads to Nm variables when m monomer types are considered but has a nice mathematically tractable form of a quadratic programming (QP) problem. A minor shortcoming of this is that due to linear interpolation and

the absence of a tuning parameter like σ, some residue sites may remain in an intermediate state in the minimizing sequence. That is, there will be a 'hybrid monomer' type at some sites, which is not realistic. But the solution of the QP serves as a good initial guess for our earlier method based on the nonlinear interpolation of Eq. 1. But then, as mentioned above, that method can only handle H-P states. In this paper, we extend that nonlinear interpolation to multiple monomer types so as to solve more realistic models of real proteins.

An additional extension of this paper is the combination of the statistical mechanics based methods [4,5] with the continuous optimization method described above. The former gives the probabilities of each monomer type at each site to minimize the overall energy. By taking the most probable m monomer types for each site, the latter method is applied to identify a specific type for each site.

The efficacy of the two-stage method is demonstrated with numerical examples in Section 4. Section 3 has a more detailed description of the two-stage optimization procedure. Next, in Section 2, the extension of the nonlinear two-state interpolation of Eq. 1 to multiple states is presented.

2 CONTINUOUS INTERPOLATION OF MULTIPLE MONOMER TYPES

In Eq. 1, one variable per residue site was used to interpolate two types, H and P. For this we used a normalized Gaussian distribution function that provides a peak at $\rho_i = 0$ for the i^{th} site. For extending this to multiple types, we take two variables, θ_i and ϕ_i, per site and define the peaks on the surface of a sphere as opposed to on a straight line as done in Eq. 1. Thus, $(1, \theta_i, \phi_i)$ are the spherical coordinates using which a unit vector $\hat{\boldsymbol{\rho}}(\theta_j, \phi_j)$ pointing to any point on the sphere could be specified. Next, we define a peak function to indicate interpolated state of j^{th} ($j = 1, 2, \cdots, m-1$) monomer type at the i^{th} site as S_i^j.

$$S_i^j = P_1^j + P_2^j \tag{3a}$$

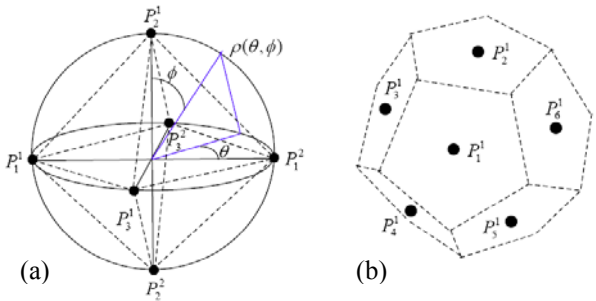
$$P_k^j = \exp\left\{-\frac{\arccos^2(<\hat{\boldsymbol{\rho}}(\theta_j, \phi_j), \hat{\boldsymbol{\rho}}_k^j>)}{\sigma^2}\right\} \tag{3b}$$

where $\hat{\boldsymbol{\rho}}_k^j$ denotes a unit vector at which a peak is located. For symmetry, a pair of peaks is arranged diametrically opposite to one another leading to $k = 1, 2$. The state of the m^{th} monomer type at the i^{th} site is given by

$$S_i^m = 1 - \sum_{j=1}^{m-1} S_i^j \tag{4}$$

Thus, when we have m monomer types, we need $2(m-1)$ unit vectors $\hat{\boldsymbol{\rho}}_j^k, (k = 1, 2; j = 1, 2, \cdots, m-1)$ defined a priori. These vectors have to be defined such that from each peak of one type the peaks of all other types are equally accessible (barring the m^{th} type which is the entire surface of the sphere) in addition to being equally spaced on the spherical surface. This can be achieved by means of polyhedra and regular polyhedra (Platonic solids) in particular.

As shown in Figure 1a, an octahedron has six vertices that are divided into three pairs to locate the peaks of three monomer types. So, it is applicable when m = 4. It can be seen that from any peak the others, except the 4th type, are equally accessible as the variables θ_i and ϕ_i take on different values. Next, as shown in Figure 1b, by placing the peaks at the centers of the 12 faces of a dodecahedron, seven monomer types (m = 7) could be handled. Other cases of m are possible with similar arrangements with other Platonic solids. In Figures 2a and 2b, the state-interpolation for m = 4 using peaks arranged on an octahedron is shown for two values of the tuning parameter σ. As can be seen, as σ decreases the peaks become sharper, the selection of the monomer type becomes clearer as per the optimized values of θ_i and ϕ_i.



(a) (b)

Figure 1: Locations for peaks on the sphere using (a) octahedron for m = 4 (b) dodecahedron for m = 7

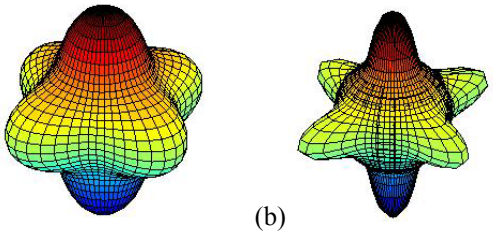

(a) (b)

Figure 2: Interpolated shape variation with different tuning parameter (a) σ = 0.4 (b) σ = 0.2

Using the above multiple-monomer state interpolation, the interpolated continuous energy can be written as

$$E = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{p=1}^{m}\sum_{q=1}^{m} A_{ij} e_{pq} S_i^p S_j^q \tag{5}$$

where A_{ij} is the $(i, j)^{\text{th}}$ entry in the adjacency matrix $\mathbf{A}$ such that A_{ij} = 1 if i^{th} and j^{th} sites have an energetic interaction and A_{ij} = 0 otherwise, and e_{pq} is the value of the energy between p^{th} and q^{th} monomer types. While there are many ways to obtain the value of e_{pq}, the Miyazawa-Jernigan [13] matrix is the easiest to implement. Thus, the values of e_{pq} come from the 20×20 MJ matrix. Furthermore, a cut of distance of 6.5 Å was used to identify

the energetically interacting, non-bonded neighboring residues to construct the adjacency matrix.

Next, we use the above energy expression to find the energy-minimizing sequence.

3 OPTIMIZATION PROBLEM

Based on the foregoing, the energy minimization problem can be posed as follows.

$$\begin{aligned} &\textit{Minimize} \quad E = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{p=1}^{m}\sum_{q=1}^{m} A_{ij} e_{pq} S_i^p S_j^q \\ &\textit{w.r.t.} \quad \{\phi_1, \theta_1, \phi_2, \theta_2, \cdots, \phi_N, \theta_N\} \end{aligned} \tag{6}$$

If we take the case of $m = 4$, we can use the peaks arranged using the octahedron. Then, each residue site needs to have four monomer types available for it. These are identified by a method developed by Saven's group [5], which gives the probabilities of each monomer type at every site using statistical mean field theory [4]. In those site-specific probabilities, we choose the four most probable monomer types for each site. At this point, it should be noted that there are still 4^N possible combinations and it is still computationally intractable for exhaustive enumeration for large values of N found in real proteins.

The problem is solved in two stages using continuous optimization as explained below. In some cases, as will be seen in the numerical examples, the Stage I could be skipped going directly to Stage II. In Stage I, a quadratic programming (QP) problem is solved wherein the state interpolation is linear [11] as mentioned in Section 1. This problem is stated as follows.

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

Subject to

$$0 \le x_p \le \text{for } p = 1, 2, \cdots, Nm$$

$$\sum_{k=1}^{m} x_{(k-1)N+l} = 1 \quad \textit{for } l = 1, 2, \cdots, N$$

and

$$\sum_{l=1}^{N} x_{(k-1)N+l} = N_\alpha \quad \textit{for } k = 1, 2, \cdots, m \tag{7}$$

where the last two constraints refer to the composition of the number of different monomer types in the total of m monomer types, and $e(\alpha_k, \alpha_l)$ is MJ-based energy interaction for monomer types α_k and α_l. This problem involves m variables but it is still easy to solve because it has the structure of a QP problem. The QP problem is solved using the routine QUADPROG in Matlab's Optimization Toolbox [14]. While the aforementioned site-specific probabilities are used as an initial guess for the QP problem, the solution of the QP problem is used as an initial guess for the Stage II solution. This is done because any local optimization algorithm needs a good initial guess. In Stage II, we solve the multi-state (MS) minimization problem in Eq. 6. The MS problem is solved using routine FMINUNC in Matlab.

4 NUMERICAL EXAMPLES

4.1 Example of a lattice model

First, we consider a 4×4 lattice model because with this the global minimum of the obtained solution can be validated by exhaustive enumeration. It has $4^{16} \simeq 4.3\times10^9$ combinations since we considered four monomer types namely cystine, alanine, serine, and aspartic acid. Thus, exhaustive enumeration is practical even though it takes considerable computing time. Figure 3a shows the lattice model with the numbering of the residue sites while Figure 3b shows the probabilities given by the statistical mean field theory based method. It can be noticed that sites 6, 9, and 14 have equal probabilities for all four types of monomers because they do not have energy interactions with their non-bonded immediate neighbors.

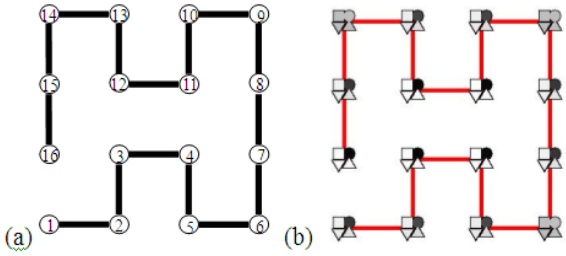


Figure 3: 4×4 Lattice model with four monomer types: ■: Aspartic Acid, ●: Cysteine, ▲:Alanine, ▼:Serine, (a) numbering of the residue sites (b) probabilities given by statistical mechanics method and indicated in gray scale.

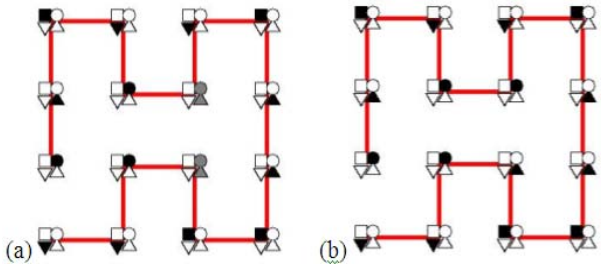


Figure 4: (a) solution of the QP problem with $E_{\min}$ = -16.82 kcal/mol (b) solution of the MS problem $E_{\min}$ = -16.67 kcal/mol. The composition of each of the four types is fixed at four.

Figures 4a and 4b respectively show the solutions obtained using the QP problem in Stage I and MS problem in Stage II. It can be seen that this being a small protein model, QP solution found the global minimum but has two sites (4 and 11) in the intermediate state. The MS problem fixes this problem because it can sharply select the type of the monomer. The energies are indicated in the figures. The energy of the MS solution is slightly larger than that of QP but it is due to the nature of QP and its intermediate states. The properties of the QP solution are analyzed in [11].

4.2 Examples of real proteins

The computational effectiveness of the QP and MS problem becomes apparent when large real proteins are considered. Here, we present the designed sequences of three proteins whose PDB (protein database) codes are given by 1SRL (56 residues), 1POH (85 residues), and

1JWO (97 residues). Since exhaustive enumeration is not practical for these cases, we compare the solutions obtained with QP and MS with that by another method developed by Saven's group. This is called exchange replica (ER) method, which is a stochastic sequence determination algorithm that uses a Monte Carlo method for a sequence update. Unlike a simulated annealing method in which temperature continuously decreases until convergence, the temperature for energy calculation in Exchange Replica method varies over a finite set of temperatures.

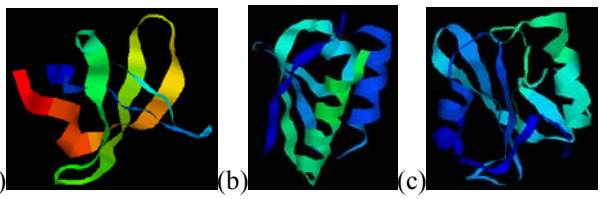


Figure 5: Ribbon schematics of the real proteins solved here, with PDB codes (a) 1SRL, (b) 1 POH, (c) 1JWO

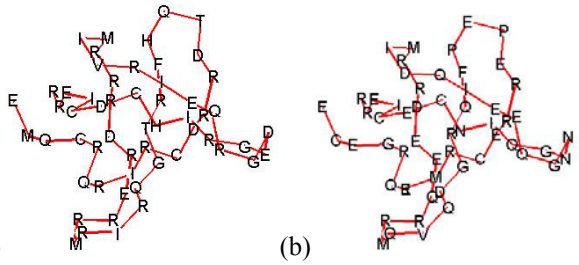


Figure 6: Solution of 1SRL with the sequence indicated by single-letter codes (a) single-stage MS solution with initial guess from site-specific probabilities, $E_{\min}$ = -764.96 kcal/mol (b) with a random initial guess, $E_{\min}$ = -691.30 kcal/mol.

Figures 6a and 6b show the sequences obtained by the MS method in a single stage for 1SRL. The single-letter codes of amino acid monomers are used to indicate the sequence in the schematic of the C_α backbone. The sequence in Figure 6a was obtained using the site-specific probabilities given by the statistical theory-based method, as the initial guess while the one in Figure 6b was with a random initial guess. It can be seen that the former is better in terms of minimized energy. This was observed in other examples as well. This example was solved with QP and ER methods, which gave the same sequence, EMQCRQRIRCDRRDTQHFRTGQRIRMRRERDRRRMIVRIE QRGDEGRIHCDCIERR, as the one shown in Figure 6a. The computation times are shown in Table 1. As can be seen in the table, the number of sequences evaluated is insignificantly low for MS and QP but ER shows better minima in all cases than MS. QP performed better on both counts but may contain intermediate monomer types, which is not realistic. Thus, each method has its own merits. Future work will be focused on combining the strengths of the three methods.

5 CONCLUSIONS

In this paper, building upon our previous work on continuous modeling of the discrete combinatorial problem of protein sequence design, we extended the ability of the method to handle multiple monomer types. The procedure presented here combines several methods: statistical theory based method that gives site-specific probabilities, quadratic programming problem, multi-state optimization problem (new contribution of this paper), and stochastic exchange replica method. With the combined procedure, real proteins with all amino acids considered can be designed as illustrated with examples. The MS and QP methods need to evaluate only a few (tens to hundreds) sequences to identify a minimum.

Table 1: Comparison of performance of three methods

Protein PDB	**Energy (kcal/mol)**			**Sequences evaluated**		
	$-E_{MS}$	$-E_{QP}$	$-E_{ER}$	n_{MS}	n_{QP}	n_{ER}
1SRL	764.96	764.96	764.96	27	279	10^7
1POH	857.76	920.53	920.53	35	271	10^7
1JWO	979.99	1044.5	1044.6	92	349	10^7

REFERENCES

[1] Sun, S., Brem, R., Chan, H. S., and Dill, K. A., *Protein Eng.*, **8**, 1995, pp. 1205-1213.

[2] Jones, D. T., *Protein Sci.*, **3**, 1994, pp. 567-574.

[3] Hellinga, H. W. and Richards, F. M., *Proc. Natl. Acd. Sci.*, **91**, 1994, pp. 5803-5807.

[4] Saven, J. G. and Wolynes, P. G. 1997. *J. Phys. Chem. B.* 101:8375-8389.

[5] Zou, J. and Saven, J. G. (2003) *The Journal of Chemical Physics*, **118**, pp. 3843–3854.

[6] Singh, M., *Special session on geometry of protein modeling in 248th Regional Meeting of the American Mathematical Society*, Lawrenceville, NJ, April 17-19, 2004.

[7] Sanjeev, B. S., Patra, S. M., and Vishveshwara, S., *J. Chem. Phys.*, **114**, 2001, pp. 1906-1914.

[8] Bendsoe, M.P. and Sigmund, O., *Topology Optimization: Theory, methods, and Applications*, Springer, Berlin, 2003.

[9] Ananthasuresh, G. K. (ed.), *Optimal Synthesis Methods for MEMS*, Kluwer Academic Publishers, Boston, 2003.

[10] Koh, S. K., Ananthasuresh, G. K., and Vishveshwara, S., *International Journal of Robotics Research*, 24(2), pp. 109-130.

[11] Koh, S. K., Ananthasuresh, G. K., and Croke, C., *Journal of Mechanical Design*, to appear in July 2005.

[12] Yin, L. and Ananthasuresh, G. K., *Structural and Multidisciplinary Optimization*, **23** (1), 2001, pp. 49-62.

[13] Miyazawa, S. and Jernigan, R., *Macromolecules*, **18**, 1985, pp. 534-552.

[14] Matlab Technical Computing Software, 2005, URL: http://www.mathworks.com.