Double Optimization for Design of Protein Energy Function

Seung-Yeon Kim[1] and Julian Lee[2]

[1] Computer Aided Molecular Design Research Center, Soongsil University, 156-743 Seoul, Korea
sykim@ssu.ac.kr
[2] Department of Bioinformatics and Life Science, Soongsil University, 156-743 Seoul, Korea
jul@ssu.ac.kr
http://bioinfo.ssu.ac.kr/~jul/jul_eng.htm

Abstract. We propose an automated protocol for designing the energy landscape suitable for the description of a given set of protein sequences with known structures, by optimizing the parameters of the energy function. The parameters are optimized so that not only the global minimum-energy conformation becomes native-like, but also the conformations distinct from the native structure have higher energies than those close to the native one, for each protein sequence in the set. In order to achieve this goal, one has to sample protein conformations that are local minima of the energy function for given parameters. Then the parameters are optimized using linear approximation, and then local minimum conformations are searched with the new energy parameters. We develop an algorithm that repeats this process of parameter optimization based on linear approximation, and conformational optimization for the current parameters, ultimately leading to the optimization of the energy parameters. We test the feasibility of this algorithm by optimizing a coarse grained energy function, called the UNRES energy function, for a set of ten proteins.

1 Introduction

According to the thermodynamic hypothesis [1], proteins adopt native structures that minimize their free energies. Therefore, obtaining energy function that accurately describes proteins would lead not only to the prediction of three-dimensional structures, but also to the understanding of folding mechanism [2,3].

Energy functions are generally parameterized from quantum mechanical calculations and experimental data on model systems. However, such calculations and data do not determine the parameters with perfect accuracy. The residual errors in energy functions may have significant effects on simulations of macromolecules such as proteins where the total energy is the sum of a large number of interaction terms. Moreover, these terms are known to cancel each other to

D.-S. Huang, K. Li, and G.W. Irwin (Eds.): ICIC 2006, LNBI 4115, pp. 562–570, 2006.

a high degree, making their systematic errors even more significant. Thus it is crucial to refine the parameters of a energy function before it can be successfully used to study protein folding.

In this work, we develop an automated protocol for the parameter optimization, where the parameters are modified so as to make conformations with larger values of root-mean-square deviation (RMSD) have higher values of energy relative to those with smaller values of RMSD. This goal is achieved by repeating two distinct optimization procedures, sampling local minimum-energy conformations for a given parameter set, and parameter optimization using linear approximation. The parameter optimization based on linear approximation is performed by supernodal Cholesky factorization method [4], a Linear Programming algorithm, and the local minimum-energy conformations with low energies for a given parameter are sampled using the conformational space annealing method [5,6,7,8].

We show the feasibility of this algorithm by successfully optimizing the UNRES energy function [9,10], for a set of ten proteins. Our work is an improvement over previous works [3,11,12,13,14,15,16], where either primitive methods of parameter optimization were used, or the optimization was performed for the training set of a small number of protein sequences, at most four of them.

2 Methods

2.1 Constrained and Unconstrained Conformational Searches

In order to check the performance of a energy function for a given set of parameters, one has to perform two types of conformational search, the constrained and unconstrained conformational searches. In the constrained search, the backbone angles of the conformations are fixed to the values of the native conformations, and only the side-chain angles are minimized with respect to the energy. We call the resulting conformations the super-native. The other conformations are obtained from unconstrained conformational search. The conformations obtained from the constrained and unconstrained searches are added to the structural database of local minimum-energy conformations for each protein. The search algorithm we use is the conformational space annealing (CSA) method [5,6,7,8]. The CSA method can be considered as a genetic algorithm that enforces a broad sampling in its early stages and gradually allows the conformational search to be focused into narrow conformational space in its later stages. As a consequence, many low-energy local minima including the GMEC of the benchmark protein can be identified for a given parameter set.

2.2 Parameter Refinement Using Linear Programming

The changes of energy gaps are estimated by the linear approximation of the energy function in terms of parameters. Among the conformations with non-zero RMSD values in the structural database, 50 (an arbitrary number) conformations with the smallest RMSD values are selected as the native-like conformations,

while the rest are considered as the non-native ones. Since an energy function can be considered to describe the nature correctly if native-like structures have lower energies than non-native ones, the parameters are optimized to minimize the energy gaps $E_{\mathrm{gap}}^{(1)}$ and $E_{\mathrm{gap}}^{(2)}$,

$$\begin{aligned} E_{\mathrm{gap}}^{(1)} &= E^{\mathrm{N}} - E^{\mathrm{NN}} \\ E_{\mathrm{gap}}^{(2)} &= E^{\mathrm{SN}} - E^{\mathrm{NN}} \end{aligned} \tag{1}$$

for each protein in the training set, where E^{N} and E^{SN} are the highest energies of the native-like and super-native conformations, respectively, and E^{NN} is the lowest energy of the non-native conformations. The energies are the modified ones that are weighted with the RMSD values of the conformations:

$$E_{\mathrm{modified}} = E + 0.3 \text{ RMSD}. \tag{2}$$

Weighting the energies with the RMSD values makes the large RMSD conformations have high energies compared to ones with small RMSD values. The parameter optimization is carried out by minimizing the energy gaps $E_{\mathrm{gap}}^{(1)}$ and $E_{\mathrm{gap}}^{(2)}$ of each protein in turn, while imposing the constraints that all the other energy gaps, including those from the other proteins, do not increase.

In this work, we adjust the 715 linear parameters of the UNRES energy function [9,10], a coarse-grained protein energy function. The energy of a local minimum-energy conformation can be written as:

$$E = \sum_j p_j e_j(\mathbf{x_{min}}) \tag{3}$$

where e_i's are the energy components evaluated with the coordinates $\mathbf{x_{min}}$ of a local minimum-energy conformation. Since the positions of local minima also depend on the parameters, the full parameter dependence of the energy gaps are nonlinear. However, if the parameters are changed by small amounts, the energy with the new parameters can be estimated by the linear approximation:

$$E^{\mathrm{new}} \approx E^{\mathrm{old}} + \sum_i (p_i^{\mathrm{new}} - p_i^{\mathrm{old}}) e_i(\mathbf{x_{min}}) \tag{4}$$

where the p_i^{old} and p_i^{new} terms represent the parameters before and after the modification, respectively. The parameter dependence on the position of the local minimum can be neglected in the linear approximation, since the derivative in the conformational space vanishes at a local minimum. The additional term 0.3 RMSD of Eq.(2) vanishes in these expressions due to the same reason. The changes of the energy gaps are estimated as:

$$\begin{aligned} \Delta E_{\mathrm{gap}}^{(1)} &= E_{\mathrm{gap}}^{(1)}(\{p_j^{\mathrm{new}}\}) - E_{\mathrm{gap}}^{(1)}(\{p_j^{\mathrm{old}}\}) \\ &= (E^{\mathrm{N}}(\{p_j^{\mathrm{new}}\}) - E^{\mathrm{NN}}(\{p_j^{\mathrm{new}}\})) - (E^{\mathrm{N}}(\{p_j^{\mathrm{old}}\}) - E^{\mathrm{NN}}(\{p_j^{\mathrm{old}}\})) \\ &= \sum_j (e_j^{\mathrm{N}} - e_j^{\mathrm{NN}})(p_j^{\mathrm{new}} - p_j^{\mathrm{old}}) \end{aligned} \tag{5}$$

$$\begin{aligned}\Delta E^{(2)}_{\rm gap} &= E^{(2)}_{\rm gap}(\{p^{\rm new}_j\}) - E^{(2)}_{\rm gap}(\{p^{\rm old}_j\}) \\ &= (E^{\rm SN}(\{p^{\rm new}_j\}) - E^{\rm NN}(\{p^{\rm new}_j\})) - (E^{\rm SN}(\{p^{\rm old}_j\}) - E^{\rm NN}(\{p^{\rm old}_j\})) \\ &= \sum_j (e^{\rm SN}_j - e^{\rm NN}_j)(p^{\rm new}_j - p^{\rm old}_j) \end{aligned} \tag{6}$$

The magnitude of the parameter change $\delta p_j \equiv p^{\rm new}_j - p^{\rm old}_j$ is bounded by a certain fraction ϵ of $p^{\rm old}_j$. We use $\epsilon = 0.01$ in this study. First, the vector δp_j is chosen within the bound to decrease the energy gap $\Delta E^{(1)}_{\rm gap}$ of the selected protein as much as possible while imposing the constraints that any positive values among $E^{(2)}_{\rm gap}$ and the energy gaps of the other proteins do not increase and negative values do not become positive. Denoting the energy gaps of the k-th protein as $E^{(p=1,2)}_{\rm gap}(k)$ and assuming the i-th protein is selected for the decrease of the energy gap, this problem can be phrased as follows:

`Minimize`

$\Delta E^{(1)}_{\rm gap}(i) = \sum_j (e^{\rm N}_j(i) - e^{\rm NN}_j(i))(p^{\rm new}_j - p^{\rm old}_j)$

`with constraints`

$|\delta p_i| \le \epsilon$

$$\Delta E^{(2)}_{\rm gap}(i) = \sum_j (e^{\rm SN}_j(i) - e^{\rm NN}_j(i))(p^{\rm new}_j - p^{\rm old}_j) \le \begin{cases} 0 & \text{if } E^{(2)}_{\rm gap}(i) > 0 \\ -E^{(2)}_{\rm gap}(i) & \text{otherwise} \end{cases}$$

$$\Delta E^{(p=1,2)}_{\rm gap}(k \ne i) = \sum_j (e^{\rm (S)N}_j(k) - e^{\rm NN}_j(k))(p^{\rm new}_j - p^{\rm old}_j) \le \begin{cases} 0 & \text{if } E^{(p)}_{\rm gap}(k) > 0 \\ -E^{(p)}_{\rm gap}(k) & \text{otherwise} \end{cases}$$

This is a global optimization problem where the linear parameters p_j are the *variables*. The object function to minimize, and the constraints, are all linear in p_j. This type of the optimization problem is called the Linear Programming. It can be solved exactly, and many algorithms have been developed for solving the Linear Programming problem. We use the primal-dual method with supernodal Cholesky factorization [4] in this work, which finds an accurate answer with reasonably computational costs.

After minimizing $\Delta E^{(1)}_{\rm gap}(i)$, we solve the same form of linear programming where now $\Delta E^{(2)}_{\rm gap}(i)$ is the objective function and the other energy gaps become constrained. Then we select another protein and repeat this procedure (300 times in this work) of minimizing $\Delta E^{(1)}_{\rm gap}$ and $\Delta E^{(2)}_{\rm gap}$ in turn.

The algorithm of parameter optimization based on linear approximation, using repeated Linear Programming, is summarized in algorithm 1.

2.3 Reminimization and New Conformational Search

Since the procedure of the previous section was based on the linear approximation Eqs.(5) and (6), we now have to evaluate the true energy gaps using the newly obtained parameters. The breakdown of the linear approximation may

Algorithm 1. Parameter optimization for conformations in the structural database, using linear programming

N_p = Number of protein sequences in the dataset (constant)
N_{it} = Maximum number of iteration (300 in this work)
$\mathbf{p}(i)$ = The initial parameters at the start of this sub-algorithm
for $i = 0$ to N_{it} **do**
  Calculate energy gaps $E_{\text{gap}}^{(1)}(p)(p = 1, \cdots, N_p)$
  for $p = 1$ to N_p **do**
    minimize $\Delta E_{\text{gap}}^{(1)}(p)$, while constraining the other energy gaps (Linear Programming)
    minimize $\Delta E_{\text{gap}}^{(2)}(p)$, while constraining the other energy gaps (Linear Programming)
  end for
  if Energy gaps are negative for all of the N_p proteins **then**
    End this sub-algorithm, since parameters are optimized for the conformations in the structural database.
  end if
end for

come from two sources. First, the conformations corresponding to the local minima of the energy for the original set of parameters are no longer necessarily so for the new parameter set. For this reason, we reminimize the energy of these conformations with the new parameters. Since super-native conformations are not local minimum-energy conformations, even with the original parameters, the unconstrained reminimization of these conformations with the new parameters may furnish low-lying local minima with small values of RMSD. Second, the local minima obtained from conformational searches with the original parameter set may constitute only a small fraction of low-lying local minima. After the modification of the parameters, some of the local minima which were not considered due to their relatively high energies, can now have low energies for the new parameter set. It is even possible that entirely distinct low-energy local minima appear. Therefore these new minima are taken into account by performing subsequent conformational searches with the newly obtained parameter set.

2.4 Update of the Structural Database and Iterative Refinement of Parameters

The low-lying local energy minima found in the new conformational searches are added into the energy-reminimized conformations to form a structural database of local energy minima. The conformations in the database are used to obtain the energy gaps, which are used for the new round of parameter refinement. As the procedure of [conformational search → parameter refinement → energy reminimization] is repeated, the number of conformations in the structural database increases. This iterative procedure is continued until sufficiently good native-like conformations are found from the unconstrained conformational search. The whole procedure is summarized in algorithm 2.

Algorithm 2. The algorithm for protein energy function parameter optimization

N_p = Number of protein sequences in the dataset (constant)
N_{it} = Maximum number of iteration (constant)
$\mathbf{p}(i)$ = the initial parameters
for $i = 0$ to N_{it} **do**
  for $p = 1$ to N_p **do**
    For the protein sequence p, sample low-lying local minimum-energy conformations with no constraints, and save their coordinates and energy components.
    For the protein sequence p, sample low-lying local minimum-energy native-like conformations (constrained sampling), and save their coordinates and energy components
  end for
  if Low-lying conformations found from unconstrained search are native-like for all of the N_p sequences **then**
    Parameter optimization accomplished. End the algorithm.
  end if
  if $i == 0$ **then**
    structural database $\Leftarrow$ low-lying conformations obtained from unconstrained search + low-lying native-like conformations obtained from constrained search (coordinates and energy components)
  else
    structural database $\Leftarrow$ structural database + low-lying conformations obtained from unconstrained search + low-lying native-like conformations obtained from constrained search (coordinates and energy components)
  end if
  Optimize parameters using Linear Programming, so that the energy gap $E_g = E_{na} - E_{nn}$ between the maximum energy among native-like conformations, E_{na}, and the minimum energy among non-native conformations, E_{nn} (See alg. 1), is minimized for each of the N_p sequences in the training set $\Rightarrow$ $\mathbf{p}(i+1)$ (new parameters)
  Reminimize structural database with respect to the energy function with the new parameters
end for

3 Results and Discussions

3.1 Ten Proteins in the Training Set and Two Proteins in the Test Set

We apply our protocol to the optimization of UNRES energy function, for a training set consisting of ten proteins, that belong to the structural class of α proteins. The PDB codes of these proteins are, 1BBA(36), 1BDD(60), 1EDI(56), 1EDK(56), 1HNR(47), 1IDY(54), 1PRB(53), 1PRU(56), 1VII(36), and 1ZDB(38), where the number inside parentheses are the lengths of these proteins. The initial parameter set is the one used in CASP3[17,18]. The optimized parameter set obtained using the training set above, is useful for the protein folding study of an α protein, including the tertiary structure prediction, where the secondary structure content

Table 1. The C_α RMSD of GMEC, for the ten proteins in the training set and the two proteins in the test set. The number in parantheses are the smallest RMSD values found in the fifty low-energy conformations.

	1BBA	1BDD	1EDI	1EDK	1HNR	1L2Y
initial parameters	8.9 (8.1)	9.8 (7.2)	7.8 (5.0)	7.6 (4.9)	9.9 (6.6)	6.5 (4.6)
optimized parameters	9.3 (4.3)	3.9 (2.9)	3.8 (2.4)	3.9 (2.4)	9.4 (5.3)	3.6 (3.1)
	1IDY	1PRB	1PRU	1VII	1ZDB	1F4I
initial parameters	11.2 (6.9)	10.1 (7.5)	11.5 (6.9)	6.3 (4.9)	7.7 (6.7)	6.8 (5.1)
optimized parameters	10.1 (4.9)	6.2 (5.9)	6.7 (5.8)	5.3 (3.5)	7.6 (3.0)	5.4 (4.2)

can be determined relatively easily using experimental methods such as Circular Dichroism (CD) or Nuclear Magnetic Resonances (NMR). It is of course possible to obtain energy function parameters suitable for the general description of proteins regardless of their structural classes, using training set consisting of proteins that belong to diverse structural classes[13,14].

The performance of UNRES energy function with the optimized parameters was tested, by sampling low-energy conformations of two α proteins not included in the training set, 1F4I(40) and 1L2Y(20).

For proteins both in training and test sets, fifty conformations were sampled in each conformational search. The RMSDs of C_α coordinates from those of native structures for the global minimum-energy conformations (GMEC) are shown in Table 1, along with the smallest values of RMSD found among the fifty low-energy conformations, obtained with the initial and optimized parameters. Five iterations of linear optimization were performed in order to obtain the optimized parameter set. The energies are not displayed since their numerical

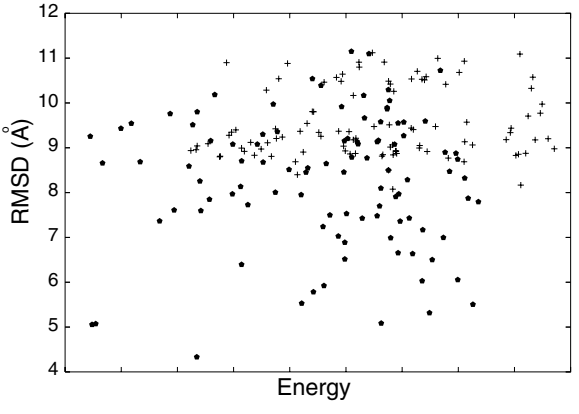

Fig. 1. Plots of the energy and C^α RMSD values of fifty low-energy conformations for the protein 1BBA, obtained using the initial (*plus signs*) and the optimized parameters (*filled circles*). Although the RMSD of the GMEC is smaller for conformations obtained with the initial parameters, much more native-like low-energy conformations are obtained with the optimized parameters.

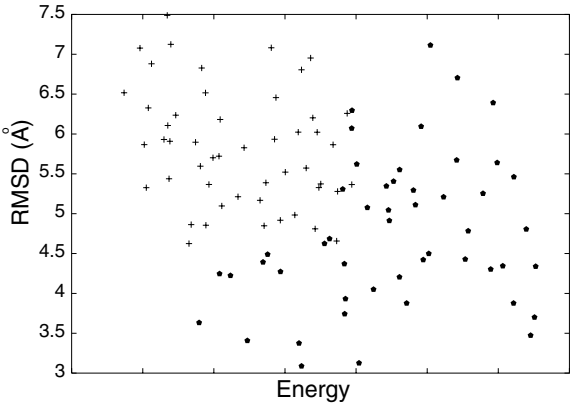

Fig. 2. Plots of the energy and C^{α} RMSD values of fifty low-energy conformations for the protein 1L2Y, obtained using the initial (*plus signs*) and the optimized parameters (*filled circles*). The GMEC found with the optimized parameters has smaller RMSD value than that found with the initial parameters. Also, much more native-like low-energy conformations are obtained with the optimized parameters.

values have no physical meaning, due to the fact that the overall scale of the linear parameters is not fixed in our protocol. We see from the data that after five iteration, the parameters are indeed optimized for the ten proteins in the training set. The smallest RMSD values found among the fifty low-energy conformations decreased for all of the proteins in the training set, and the RMSD values of the GMECs also decreased for all of them except 1BBA. Since the native structure of a protein is usually predicted by constructing several models, rather than selecting just one GMEC, GMEC not being native-like is not a serious problem, as long as there are sufficient number of native-like conformations in the final set of low-energy conformations obtained by the sampling algorithm. In fact, even for 1BBA where RMSD value of the GMEC increased, those of the second and third lowest energy conformations all decreased, and there are much more native-like low-energy conformations obtained with the optimized parameters, as can be seen in Fig. 1 where the RMSD values of the fifty low-energy conformations are plotted against their energy values.

We see that also for the two proteins in the test set, the lowest RMSD value found from the fifty low-energy conformations, as well as that of GMEC, decrease as the parameters are optimized. Again, the low-energy conformations become more native-like overall, as can be seen from the plot of RMSD against energy values for 1L2Y(Fig. 2). The result suggests that the optimized parameters are transferable to α proteins not included in the training set, and can be used for the study of protein folding of such proteins.

Acknowledgement

This work was supported by the Soongsil University Research Fund.

References

1. Anfinsen, C. B.: Principles that govern the folding of protein chains. Science vol. 181 (1973) 223–230
2. Kim, S.-Y., Lee, J., Lee, J.: Folding of small proteins using a single continuous potential. J. Chem. Phys. vol.120 (2004) 8271–8276
3. Lee, J., Kim, S.-Y., Lee, J.: Optimization of potential-energy parameters for folding of several proteins. J. Kor. Phys. Soc. vol. 44 (2004) 594–598
4. Mészáros, C. A.: Fast Cholesky Factorization for Interior Point Methods of Linear Programming. Computers & Mathematics with Applications vol. 31 (1995) 49–54
5. Lee, J., Scheraga, H. A., Rackovsky. S.: New optimization method for conformational energy calculations on polypeptides: Conformational space annealing. J. Comput. Chem. vol. 18 (1997) 1222–1232
6. Lee, J., Lee, I.H., Lee, J.: Unbiased Global Optimization of Lennard-Jones Clusters for $N >= 201$ Using the Conformational Space Annealing Method. Phys. Rev. Lett. vol. 91 (2003) 080201-1–4
7. Kim S.-Y., Lee S.J., Lee J.: Conformational space annealing and an off-lattice frustrated model protein Conformational space annealing and an off-lattice frustrated model protein. J. Chem. Phys. vol. 119 (2003) 10274–10279
8. Kim S.-Y., Lee S.B., Lee J.: Structure optimization by conformational space annealing in an off-lattice protein model. Phys. Rev. E vol. 72 (2003) 011916-1–6
9. Liwo, A., Oldziej, S., Pincus, M. R., Wawak, R. J., Rackovsky, S., Scheraga, H. A.: A united-residue force field for off-lattice protein-structure simulations. I: Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. J. Comput. Chem. vol. 18 (1997) 849–873
10. Liwo, A., Pincus, M. R., Wawak, R. J., Rackovsky, S., Oldziej, S., Scheraga, H. A.: A united-residue force field for off-lattice protein-structure simulations. II: Parameterization of local interactions and determination of the weights of energy terms by Z-score optimization. J. Comput. Chem. vol. 18 (1997) 874 – 887
11. Lee, J., Ripoll, D. R., Czaplewski, C., Pillardy, J., Wedemeyer, W. J., Scheraga, H. A.: Optimization in Macromolecular Potential Enrgy Functions by Conformational space Annealing. J. Phys. Chem. B vol. 105 (2001) 7291–7298
12. Liwo, A., Arlukowicz, P., Czaplewski, C., Oldziej, S., Pillardy, J, Scheraga, H. A. : A method for optimizing potential-energy functions by a hierarchical design of the potential-energy landscape: Applicaiton to the UNRES for field. Proc. Natl. Acad. Sci. U.S.A. vol. 99 (2002) 1937–1942
13. Lee, J.; Park, K.; Lee, J.: Full Optimization of Linear Parameters of a United Residue Protein Potential. J. Phys. Chem. B vol. 106 (2002) 11647–11657
14. Lee, J., Kim, S-.Y., Lee, J.: Design of a protein potential energy landscape by parameter optimization. J. Phys. Chem. B vol. 108 (2004) 4525–4534
15. Lee, J., Kim, S.-Y., Lee, J. Protein structure prediction based on fragment assembly and parameter optimization. Biophys. Chem. vol. 115 (2005) 209–214
16. Lee, J., Kim, S.-Y., Lee, J. Protein structure prediction based on fragment assembly and beta-strand pairing energy function. J. Korean Phys. Soc. vol. 46 (2005) 707–712
17. Lee, J., Liwo, A., Ripoll, D. R., Pillardy, J., Scheraga, H. A.: Calculation of Protein Conformation by Global Optimization of a potential energy function. Proteins. Suppl. **3** (1999) 204–208
18. Third Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction; Asilomar Conference Center, December 13-17, 1998; http://predictioncenter.org/casp3/Casp3.html.