An Extremal Optimization Search Method for the Protein Folding Problem: the Go-Model Example

Alena Shmygelska
Department of Structural Biology
Stanford University
299 W. Campus Dr.
Stanford, CA, 94305, USA
alenas@stanford.edu

ABSTRACT

The protein folding problem consists of predicting the functional (*native*) structure of the protein given its linear sequence of amino acids. Despite extensive progress made in understanding the process of protein folding, this problem still remains extremely challenging.

In this paper we introduce, implement and evaluate the Extremal Optimization method – a biologically inspired approach which has been applied very successfully to other optimization problems – for the protein folding problem using a widely studied Gō-model of folding.

Standard methods based on the variants of the Monte Carlo method have difficulty exploring low-energy regions efficiently due to the ruggedness of the search landscapes. Most computational methods in the protein folding literature do not keep track of which interactions remain unsatisfied during the search. Instead, in this paper, we propose an adaptive meta-search method which ensures that unexplored promising parts of the search landscape are visited. This is achieved by implementing an adaptive Extremal Optimization meta-search that guides a standard Monte Carlo sampling.

We demonstrate that our Extremal Optimization meta-search compares favorably with currently best-performing Replica Exchange Monte Carlo method in reaching the native state for long proteins under the Gō-model potential. Additionally, we show that our novel approach samples larger ensembles of near-native structures by plotting parts of the energy landscape sampled during the search. Furthermore, we find that it scales well with the increasing sequence length. To our best knowledge this is the first application of Extremal Optimization to the protein folding problem.

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]: Problem Solving, Control Methods, and Search

GECCO'07, July 7–11, 2007, London, England, United Kingdom.
Copyright 2007 ACM 978-1-59593-698-1/07/0007 ...$5.00.

General Terms

Algorithms

Keywords

Extremal Optimization, protein folding, Go-model

1. INTRODUCTION

The *ab initio* protein folding problem is the problem of predicting the functional three dimensional protein structure (the *native* state) from its amino acid sequence for a given energy function. It represents an optimization problem (continuous or discrete, depending on the model assumed). Even for simple lattice models that restrict conformations to a grid, it is an *NP*-hard combinatorial problem [14, 24].

Solving the protein folding problem has enormous implications: computational drug design can be carried out; understanding a number of diseases that are directly caused by protein misfolding can take place, prevention of aggregation and fibrillogenesis will be possible; proteins with desired structure and function can be engineered. Ultimately, simulation of the complete living cell will become possible.

Computational methods for predicting protein structure from sequence are very attractive, since experimental methods (X-ray crystallography and NMR) for protein structure determination are highly labor intensive and require purification and, in the case of X-ray crystallography, crystallization of proteins.

A number of search methods were introduced in the literature for the protein folding problem. The most widely used are Metropolis Monte Carlo (MC) methods and their variants [17, 19, 20, 25].

In canonical Monte Carlo that samples the states of the system according to the Boltzmann probability density function, very high and, more importantly, very low energy configurations are rarely sampled. To overcome these problems, a number of Generalized Ensemble Monte Carlo methods have been developed [12]. These methods strive to perform a random walk in potential energy space by computing the density of states, sampling expanded range of temperatures, or computing other physical quantities affecting transitions between the states during search. Currently, best-performing algorithms for *ab initio* folding are Generalized Ensemble methods [10, 12, 21], particularly Replica Exchange Monte Carlo (REMC) [21], also known as the multiple Markov Chain method and Parallel Tempering [12], which outperforms other search methods.

In REMC, a number of non-interacting copies (replicas) at different temperatures are simulated independently by MC. Every few steps, pairs of replicas are exchanged with a specified transition probability. The weight factor is a product of Boltzmann weights (is essentially known).

The drawback of this method is that as the number of degrees of freedom (f) of the system increases, the required number of replicas also increases ($\sqrt{f}$). Improvements of REMC introduced in the literature include hybrid approaches between REMC (for the weight factor determination) and single chain Generalized Monte Carlo methods such as Multicannonical Algorithm, or Simulated Tempering [12, 15, 21], but difficulty in sampling all relevant parts of the search landscape or the search for the lowest energy conformation still remains.

The performance of stochastic local search algorithms is critically dependent on the properties of the search landscape encountered, such as the degree of landscape ruggedness, connectivity of the landscape, the number and distribution of local minima. Therefore, reactive search strategies that can identify unexplored promising parts of the search landscape and adapt the search accordingly are among the most effective tools for solving optimization problems with complex search landscapes.

Only a few adaptive search strategies for protein folding have been developed so far, these include Energy Landscape Paving (ELP) [10] and Model-based Search (MBS) [7].

Energy Landscape Paving introduces temporary energy surface deformation and the barrier height is decreased proportionally to the time the system stays in the minima (configurations are searched with time-dependent weights, similarly to Tabu Search (TS) that uses a time-dependent adaptive memory in the search [9]).

Model-based Search (MBS) is a stochastic local search that focuses on promising regions of the search space by storing a certain number of conformations L (local minima) in memory and expanding N conformations in every step of the algorithm [7]. During each step of the search, new local optima are stored and all conformations stored are ranked and pruned based on the scoring function that considers their energies and the radius of a local minimum they represent (the radius is estimated by the distance to the nearest neighbors using root mean square deviation). MBS has been shown to outperform in some cases simple MC used in the ROSETTA algorithm [7, 19].

In this paper, we introduce a novel adaptive meta-search method that alternates between two distinct modes of the search process at different levels: the high level which ensures that unexplored promising parts of the search landscape are visited and the low-level search which provides the thorough exploration of local neighborhoods. Generally, meta-search methods perform a higher-level search over the space of candidate solutions by combining multiple search strategies. The idea is that by using multiple search processes in the intelligent way, we are able to search more of the energy landscape in less time.

In this work, we developed a meta-search which utilizes an Extremal Optimization (EO) search [3] – a biologically inspired approach which has been applied very successfully to other optimization problems – at its higher level controlling the standard Markov Chain Monte Carlo search at its lower level.

Most natural systems are self-organizing, self-reshaping and dynamic. These systems often posses a large number of strongly coupled components with similar properties that self organize critically, a concept introduced to describe emergent complexity in physical systems. Recently, Boettcher and Percus proposed a new biologically inspired optimization method called Extremal Optimization [5], that utilizes the self-organized critical state as an efficient search strategy. Currently, EO ranks among the best algorithms for a number of hard combinatorial optimization problems, some of which include the Ising Spin Glasses Problem, the Traveling Salesman Problem (TSP), the Graph Coloring Problem (GCP), the Satisfiability problem (SAT), and many others, *e.g.*, [5, 6].

EO appears to be a very attractive computational method for addressing the protein folding problem, since it combines aspects of the self-organization process and the notion of coupling between system components. These concepts and ideas apply rather naturally to protein folding by: representing pairs of protein amino acids as solution components, guided by energy contributions between pairs of residues as fitness, the system can undergo self-organizing critically phenomena and low-energy conformations are obtained. Determining how well EO performs for protein folding problems motivated this work.

The remainder of the paper is organized as follows: Section 2 introduces an Extremal Optimization meta-search for the G$\bar{o}$-model of protein folding, and describes major components of the outlined method. In Section 3, we detail the model adopted for the problem, describe design of the experiments and provide implementation and environment details. Section 4 compares performance of our EO meta-search with that of best performing algorithms, details performance differences and experimental findings. Finally, in Section 5, we summarize our conclusions and outline some suggested directions for future research.

2. APPROACH – EXTREMAL OPTIMIZATION META-SEARCH

Extremal Optimization (EO) is an optimization heuristic inspired by self-organized critically phenomena captured by the Bak-Sneppen model of evolution [2] which relies on the extremal processes. In this evolutionary model self-organization towards adaptation emerges naturally, from the dynamics of a selection against the extremely "bad" fitness components of the system ("species"). After a sufficient number of steps, the system reaches a highly correlated state known as self-organized criticality (SOC) when almost all components of the system have reached fitness values above a certain threshold [3].

From a computational point of view, EO successively replaces extremely undesirable variables of a single sub-optimal solution with new, random ones. Large fluctuations provided by the local search moves utilized in EO ensure, that the search efficiently explores many local optima. Furthermore, recently it has been shown that EO can be applied successfully to systems with highly connected variables (such as Sherrington-Kirkpatrick Spin Glasses) when systems are not determined by short-range interactions only [4].

This is also the case for the *ab initio* protein folding problem which is of interest here. The problem is formally defined as follows: Given an amino acid sequence $s = s_1 s_2 \ldots s_n$

and an energy function $E(c)$, find an energy-minimizing conformation of s, *i. e.*, find $c^* \in C(s)$ such that $E(c^*) = \min\{E(c) \mid c \in C(s)\}$, where $C(s)$ is the set of all valid conformations for s.

The success of the EO algorithm for a combinatorial problem from any domain is primarily determined by (1) the choice of variables to which fitness values are attributed to and (2) the neighborhood used to modify fitness of low-fit components during the search.

EO is currently one of the best-performing methods for Ising Spin Glasses model of ferromagnetism [6], in which couplings $J_{i,j}$ attempt to align neighboring spins. Fitness values in this problem are attributed to individual spins, and the simplest neighborhood for an update consists of all configurations that could be reached from the current state through the flip of a single spin [3]. For the protein folding problem attributing fitness value to a single amino acid results in a system that does not allow for the instantaneous change of fitness which is required for a sub-sequent update of a conformation c from a given neighborhood $N(c)$. Instead, we propose to attribute fitness λ_{ij} to the pairs of residues (i, j), in a general case there are $\frac{n \cdot (n-1)}{2}$ such pairs for a protein of length n.

Ideally, in EO-type search, the objective function minimized is represented by the sum of fitness values of all system components, taken with a negative sign [3]. For the protein folding problem this relationship is given by the following equation:

$$E(c) = -\sum_{(i,j)} \lambda_{ij} \tag{1}$$

Therefore, we attribute fitness values λ_{ij} with the non-bonded pairwise interactions between amino acids i and j, usually described by means of 10-12 Lennard-Jones potential for van der Waals forces [8] (see Methods Section for details).

Using the above-described fitness assignment, a natural neighborhood to employ is a neighborhood (move-set) which modifies the distance between the selected pair of residues. The distance modification in turn changes the pairwise non-bonded energy between selected residues and therefore leads to a change of the fitness value of the pair. This local move, however, may result in a conformation with high energy. The later situation may happen because in order to bring together two residues that are currently located far apart, the location of other residues may need to be changed significantly, forcing large configurational changes.

To address this, we devised a meta-search algorithm that iteratively conducts:

1. Extremal Optimization search at its higher level to ensure that all parts of the search space, that correspond to different pairwise arrangements of amino acids, are explored

2. Monte Carlo search at its lower level which further minimizes the relaxed conformation in its local neighborhood.

The switch between the two phases, higher-level EO and low-level MC, is performed in the following way: when a selected pair of residues chosen during EO stage is within required distance cut-off the search switches to the MC search to improve the energy of the conformation further; when subsidiary MC search has reached the local optimum and therefore no improvement is observed on the lowest energy for a specified number of steps, *noImpr*, the search switches back to the higher-level EO phase, particular details of these choices are detailed later in this section.

The search iterates through the two phases until the termination criterion is not met (either the specified energy level or the specified time cut-off is reached).

To devise EO meta-search for the protein folding problem, we adopted the model referred as "Gō-model" of protein folding [23, 22], primarily due to the unavailability of a single universal energy function for *ab initio* folding from an extended (denatured) state. Additionally, Gō-model has been widely studied in the protein folding literature [1, 8, 11, 13, 16, 18], and was shown to capture folding events of a number of proteins rather well. For details of the model and the energy potential used see Methods Section.

In the following sub-sections we provide details for each of the two stages of the EO meta-search.

High-level Exploration Phase – Extremal Optimization

Our EO meta-search performs a *high-level exploration phase*, during which low-fitness components are identified, and the search is steered in the direction which modifies their respective fitness values.

Computationally this is achieved as follows: First, fitness value λ_{ij} of each solution component (i, j), where (i, j) is a pair of residues, is calculated (or updated in subsequent iterations). For Gō-model, adopted in this paper, only pairs of native contacts are considered and fitness values λ_{ij} are based on non-bonded pairwise energy contribution between amino acids i and j described by the Lennard-Jones 10-12 potential [8], see Methods Section for details.

Next, components are ordered according to their fitness value, with the worst fitness component ranking as 1 and the best ranking as NC, where NC is number of native contacts in our model. As in [6], lower fitness components are chosen probabilistically according to the scale-free power-law distribution $P_k \propto k^{-\tau_{EO}}$, where k is a rank $(1 \leq k \leq m)$ and τ_{EO} is a parameter specifying weighing of fitness values during the selection process (asymptotic choice of $\tau_{EO} - 1 \propto [ln(NC)]^{-1}$ is often used). Ordering of fitness values is usually approximated by a binary tree of depth $O(log_2 NC)$ with the least-fit components ranking near the root [3] to reduce the complexity of component retrieval and fitness update.

In our implementation the fitness λ_{ij} of a chosen low-fitness component is changed by means of a move-set that uses subsidiary Monte Carlo search with a biasing potential attempting to change fitness λ_{ij} by bringing the selected pair of residues closer to each other. The biasing potential is a simple harmonic potential $K_{bias}(r_{ij} - \sigma_{ij})^2$ added to the main energy potential function that is minimized, see Methods Section for details. Distance r_{ij} denotes current distance between residues i and j; for the Gō-model considered here, σ_{ij} is taken to be the distance between residues i and j in the native state. Thus, the minimum of a biasing potential is attained when the fitness of the component is maximized.

The move completes when either the distance between the selected residues $r_{ij} \leq \sigma_{ij}$ or the number of attempts exceeds a specified number, see Methods Section for details.

The later condition is imposed to bound computational time spent on a single move.

Low-level Exploitation Phase – Monte Carlo Search

The second phase is a *low-level exploitation phase*, when low-temperature Monte Carlo search optimizes the relaxed protein conformation obtained during the exploration phase.

The simplest way to recognize when the search becomes unsatisfactory, *i.e.* it has reached the local (or in some cases a global) optimum, is to record the number of steps during which no improvement on the lowest energy have been observed. Thus, if no improvement on the lowest energy has been seen for a certain number of steps (*noImpr*, which is a parameter of the algorithm, the value is specified in Methods Section) low-level Monte Carlo phase is terminated and the high-level Extremal Optimization phase is initiated.

3. METHODS

In this section, we provide a detailed description of the Gō-model and Gō-energy potential used in this work, give fitness value definition for EO search phase, specify details of the move-set utilized during subsidiary MC, and provide details for experimental analysis performed.

The Gō-model of Protein Folding

In the Gō-model a protein is represented as a linear chain of idealized spheres located at the C_α position (one node per residue). Contacts between residues are classified as either native (present in the three-dimensional native state) or non-native (absent from the native state). Two residues are in contact in the native state if any of the heavy atoms of the side-chain or the C_α of one residue occur within the cut-off radius of 4.55 Å to heavy atoms of the side chain or the C_α of the second residue [8, 11]. As in [8] native contacts between pairs of residues (i, j) with $j \leq i + 3$ are discarded as they interact due to chain connectivity.

Energy potential for the Gō-model has the form of standard energy force-fields employed in a number of molecular dynamics packages, such as AMBER, GROMACS, and CHARMM, with notable difference that the non-bonded interactions (van der Waals and electrostatic terms) are replaced by attractive energy contribution for native contacts, and a non-native repulsion term is usually present [8, 18]. Thus, this potential assigns negative energy contributions only to native interactions, each of these interactions is weighted the same.The energy of a configuration c of a protein with the native state c_0 is given by the expression [8]:

$$\begin{aligned} E(c, c_0) &= \sum_{angles} K_\theta(\theta - \theta_0)^2 \\ &+ \sum_{dihedral} K_\tau^{(n)} \left[1 + cos(n(\tau - \tau_0))\right] \\ &+ \sum_{i<j-3} \epsilon(i,j) \left[5\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - 6\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{10} \right] \\ &+ \sum_{i<j-3} \epsilon_2(i,j) \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} \end{aligned}$$

where θ (θ_0) and τ (τ_0) are the bond and dihedral angles correspondingly. The bond angles are formed between three consecutive residues of the chain, while the dihedral angles are the angles between the normals of the two planes formed by four consecutive residues. The dihedral potential consists of a sum of two terms for every adjacent C^α atoms, with periods $n = 1$ and $n = 3$ as in [8]. The last term represents the non-local native interactions and a short-range repulsive term for non-native pairs. Parameters $\epsilon(i, j) = 1\ kT$ for native interactions and zero otherwise, $\epsilon_2(i, j) = 1\ kT$ for non-native interactions and zero otherwise, σ_{ij} for native interactions in the distance between C^α atoms of residues i and j if the two residues are in contact in the native state and 4 Å for non-native interactions. Weights are set in the following way: $K_\theta = 20\epsilon$, $K_\tau^{(1)} = \epsilon$ and $K_\tau^{(3)} = 0.5\epsilon$. Clementi *et al.* have found that with this choice of parameters the stabilizing energy residing in the non-bonded contacts is approximately twice the stabilizing energy residing in the torsional degrees of freedom [8]. Bond bending is disregarded, and all pseudo-bonds between consecutive C^α atoms are considered to have ideal geometry with length of 3.814 Å. The Gō-type potential ensures that the native state is at the global minimum of the potential, which is a very desirable property of any energy function used.

During the high-level Extremal Optimization phase, the fitness is attributed to each pair of native contacts (i, j). There are NC such pairs. This fitness is defined as the absolute value of 10-12 Lennard-Jones non-bonded potential:

$$\lambda_{ij} = \begin{cases} \left| \epsilon(i,j) \left[5\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - 6\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{10} \right] \right|, & \text{if } r_{ij} > \sigma_{ij} \\ 1, & \text{if } r_{ij} \leq \sigma_{ij} \end{cases} \tag{2}$$

where $0 \leq \lambda_{ij} \leq 1$, with low fitness of 0 being bad, and high fitness of 1 being good.

Search Neighborhood

Initially, we consider an extended polymer where no non-covalent contacts are present; all bond angles are set to 120° and all of the dihedral angles are set to 180°. Local search modifications are performed in the continuous space of the bond and dihedral angle space. There are $n-2$ bond angles (θ) and $n-3$ dihedral angles (τ), which yields $2n-5$ degrees of freedom in total. During the search a uniformly random scan through these degrees of freedom is performed. A move corresponds to a single attempt to modify one of these angles by a displacement drawn from the Gaussian distribution with standard deviation specified (30° is used).

A move-set (search neighborhood) utilized in subsidiary Monte Carlo search during the high-level EO phase is similar to the standard local search moves described above, except that a biasing potential is applied to a specified pair of residues (k, l). Thus, the energy potential used is the Gō-model potential plus $K_{bias}(r_{kl} - \sigma_{kl})^2$, where r_{kl} is the current distance between the pair, and σ_{kl} is the distance between k and l in the native state. Parameter K_{bias} was set to ϵ.

The move is finished when either the distance between the selected residues $r_{ij} \leq \sigma_{ij}$ or the number of attempts exceeds a specified number (20 is used).

Table 1: Set of proteins used in this study. Assignment of class was performed by CATH [17].

	Protein	PDB ID	length	class
1	Ribonuclease T1 (Fungus)	1bu4	104	$\alpha\beta$
2	Acidic fibroblast growth factor-1 (Human)	2afgA	129	mainly β
3	Nuclease (S. aureus)	1stn	136	mainly β
4	Oxy-myoglobin (Sperm whale)	1a6m	151	mainly α
5	Ribonuclease H (E. coli)	2rn2	155	$\alpha\beta$
6	Lysozyme (Coliphage T4)	3lzm	164	mainly α
7	Thermitase (T. vulgaris)	1thm	279	$\alpha\beta$
8	Xylanase (Fungus)	1bg4	302	$\alpha\beta$
9	Lignin peroxidase (Fungus)	1llp	343	mainly α

Empirical Analysis and Implementation Details

For our empirical analysis we used representative protein sequences provided in Table 1.

MC was run at a constant room temperature of $T = 298$ K, REMC was run with 5 replicas with linearly spaced temperatures (ranging from 300 to 700 K), exchanges were attempted every 1000 scans through the chain, similar parameter settings are used in the literature for Gō-models [8, 11]. Low-level MC search in EO meta-search was run at a constant room temperature of $T = 298$ K until number of *noImpr* steps did not exceed 1000 scans through the chain. These parameters were determined from short preliminary runs (not provided here).

EO meta-search has been implemented in C++ and compiled using g++ (version 3.3.6) for the Linux operating system. The same holds for our implementations of simple Monte Carlo (MC) and Replica Exchange Monte Carlo (REMC).

All experiments were performed on PCs with 2.4 GHz Pentium IV CPUs, 256Kb cache, and 1Mb RAM, running Redhat Linux (our reference machine).

4. RESULTS AND DISCUSSION

In the following work, we address the following questions: How does EO meta-search compare with standard MC and REMC methods in finding close-to-native conformations and in exploring the energy landscape of a given protein sequence under the Gō-model? How does its performance scale with sequence length?

We compare our new Extremal Optimization meta-search to Monte Carlo and Replica Exchange Monte Carlo methods. To evaluate algorithms we used a set of 9 proteins of length ranging from 104 to 343 residues representing various CATH structural classes, see Methods Section for details. We conducted 20 independent runs on each protein sequence starting from the completely extended state. Each run was terminated after a fixed CPU time limit had been reached (2 hours CPU time cut-off was used). Only isolated studies of proteins longer than 100 residues exist in the literature [8, 16, 18], most concentrate on unfolding process of the native structure under the Gō-model, therefore results obtained in this work could not be directly compared with the literature.

From the distribution of energy levels over 20 independent runs, we determined the average energy, the average C^α coordinate root mean square deviation (RMSD), the average fraction of native contacts Q, standard deviation of these values, as well as the lowest energy and RMSD, and the highest fraction of native contacts Q reached.

The fraction of native contacts Q was counted as the fraction of contacts between residues (i, j) that were present in the native state. In this context, the contact is regarded to be made if distance between the C^α atoms is shorter than $\gamma = 1.2$ times their native distance σ_{ij} as in [8, 16].

As seen from our results presented in Table 2, EO meta-search outperforms or comes very close to the performance of REMC on all sequences considered. Particularly, performance differences (as captured by the average and lowest energy, average and lowest RMSD, and average and highest Q observed over 20 runs) become more noticeable as the length of the protein sequence increases. Both EO meta-search and REMC outperform MC on longer sequences significantly.

For example, Figures 1 and 2 provide comparison of alignments between lowest-energy structures found within the given CPU cut-off time by EO meta-search, REMC and MC for sequences of Ribonuclease H (pdb id 2rn2, 155 amino acids) and Xylanase (pdb id 1bg4, 302 amino acids in length). For for Ribonuclease H, all methods find low energy structures that have close alignment with the native state, while on a longer sequence of Xylanase both EO meta-search and REMC outperform MC.

It should be noted, however, that we did not substantially optimized parameters for any of the algorithms in the presented experiments, and instead we used commonly adopted parameters for MC and REMC adjusted for longer lengths [11, 16]. Additionally, performance of algorithms can also be affected by the cut-off time used. Therefore, to further evaluate performance of our EO meta-search and the methods known from the literature, we followed the methodology presented in [16] and analyzed energy landscapes sampled by each method. We performed a single long run (5 CPU hours) of each algorithm periodically recording states sampled. We then compared how well search processes explore low-energy regions of the landscape for a given protein by plotting the free energy landscape as a function of the fraction of native contacts Q present and RMSD from the native state. The free energy in this analysis is viewed as a negative logarithm of frequency of sampling a conformation with given Q and RMSD [16]. As seen from the results for a representative Ribonuclease H protein in Figure 3, MC samples a very local part of the landscape, while both REMC and EO sample larger parts. EO meta-search samples more low-RMSD and high-Q states than either REMC or MC.

Table 2: Comparison of the energy levels reached, root mean square deviation, and fraction of the native contacts on a data set of 9 proteins by Monte Carlo (MC), the Replica Exchange Monte Carlo (REMC) algorithm and the Extremal Optimization (EO) meta-search algorithm. All results are based on 20 runs per algorithm per protein sequence.

pdb id	Length	NC	Method	$Energy_{avg} \pm sd$	$Energy_{min}$	$RMSD_{avg} \pm sd$	$RMSD_{min}$	$Q_{avg} \pm sd$	Q_{max}
1bu4	104	244	MC	−107.19 (±25.11)	−144.65	4.10 (±2.61)	1.34	0.79 (±0.11)	0.97
			REMC	−115.12 (±32.75)	−148.38	**2.75** (±1.51)	1.34	**0.85** (±0.20)	0.95
			EO	**-119.62** (±20.79)	**-164.86**	3.31 (±1.48)	**1.15**	**0.85** (±0.06)	**0.98**
2afg	129	363	MC	−98.39 (±27.72)	−166.70	9.61 (±4.90)	2.39	0.59 (±0.09)	0.82
			REMC	−117.99 (±43.02)	−191.99	4.89 (±3.50)	1.45	0.71 (±0.19)	0.86
			EO	**-128.35** (±47.42)	**-207.32**	**4.75** (±3.34)	**1.13**	**0.72** (±0.13)	**0.94**
1stn	136	353	MC	−111.64 (±24.19)	−151.19	7.10 (±3.35)	2.99	0.66 (±0.07)	0.80
			REMC	−102.11 (±34.88)	−145.26	**6.06** (±2.80)	**1.78**	0.68 (±0.19)	0.80
			EO	**-114.51** (±39.07)	**-162.54**	6.30 (±3.06)	2.39	**0.69** (±0.10)	**0.82**
1a6m	151	333	MC	−106.50 (±31.89)	−153.02	6.63 (±4.57)	**1.66**	0.71 (±0.09)	0.86
			REMC	−100.94 (±34.26)	**-169.50**	5.15 (±2.82)	2.02	0.71 (±0.18)	**0.88**
			EO	**-122.66** (±23.59)	−157.20	**4.82** (±2.65)	2.28	**0.73** (±0.06)	0.81
2rn2	155	387	MC	−88.39 (±27.77)	−135.13	9.43 (±4.85)	3.41	0.58 (±0.08)	0.73
			REMC	−79.00 (±26.85)	−120.23	**6.51** (±3.73)	2.80	0.61 (±0.16)	0.73
			EO	**-101.73** (±38.14)	**-186.19**	6.72 (±4.11)	**2.55**	**0.66** (±0.08)	**0.85**
3lzm	164	388	MC	−168.90 (±22.73)	−214.86	**3.17** (±1.42)	1.69	0.79 (±0.05)	0.88
			REMC	−161.48 (±43.67)	−213.94	3.74 (±1.78)	1.97	0.79 (±0.18)	0.85
			EO	**-172.48** (±32.75)	**-231.43**	3.99 (±2.50)	**1.46**	**0.81** (±0.07)	**0.91**
1thm	279	868	MC	−75.46 (±32.25)	−131.13	19.09 (±4.54)	8.29	0.37 (±0.03)	0.43
			REMC	−109.84 (±50.98)	**-197.38**	15.56 (±6.15)	**4.64**	0.44 (±0.12)	**0.55**
			EO	**-150.66** (±21.79)	−183.47	**11.95** (±4.01)	5.70	**0.49** (±0.03)	0.54
1bg4	302	909	MC	−73.17 (±40.31)	−144.35	18.23 (±4.52)	13.03	0.40 (±0.04)	0.46
			REMC	−92.62 (±54.83)	−190.91	13.07 (±4.86)	6.24	0.45 (±0.12)	0.52
			EO	**-142.58** (±77.07)	**-272.09**	**11.99** (±3.91)	**4.03**	**0.51** (±0.07)	**0.67**
1llp	343	953	MC	−15.54 (±45.99)	−82.94	22.44 (±4.53)	12.52	0.35 (±0.03)	0.41
			REMC	−29.97 (±51.01)	−136.23	21.79 (±8.19)	8.13	0.39 (±0.12)	**0.54**
			EO	**-60.34** (±66.57)	**-190.48**	**21.68** (±6.15)	**7.37**	**0.43** (±0.06)	**0.54**

Although it can be argued that this system (G$\bar{o}$-model) is simplistic, the results of our Extremal Optimization meta-search as compared to the results obtained by the Replica Exchange Monte Carlo method are encouraging.

5. CONCLUSION

We have shown that our Extremal Optimization meta-search is successful in finding low-energy states for the G$\bar{o}$-model of protein folding. It compares favorably with the currently best-performing Replica Exchange Monte Carlo method in finding low-energy states of a given protein sequence under the G$\bar{o}$-model potential.

This system, similarly to other more complex protein models, displays highly connected components, where each component is coupled to many others by means of long-range interactions. Despite coupling of the system, EO is highly competitive for this problem when utilized with the subsidiary low-temperature Monte Carlo method. Future work will include the study of EO meta-search for other more complex protein models.

Acknowledgment

This work has been supported by an NSERC Postdoctoral Scholarship (PGS-D). It has also been supported by NIH GM-41455 to Michael Levitt, whom we thank for discussion.

6. REFERENCES

[1] E. Alm and D. Baker. Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. Natl. Acad. Sci. U.S.A.*, 96:11305–11310, 1999.

[2] P. Bak and K. Sneppen. Punctuated equilibrium and criticality in a simple model of evolution. *Phys. Rev. Lett.*, 71:4083–4086, 1993.

[3] S. Boettcher. Extremal optimization: Heuristics via co-evolutionary avalanches. *Computing Sci. and Eng.*, 2:75–82, 2000.

[4] S. Boettcher. Extremal optimization for sherrington-kirkpatrick spin glasses. *Eur. Phys. J. B*, 46:501–505, 2005.

[5] S. Boettcher and A. G. Percus. Extremal optimization: Methods derived from co-evolution. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)*, pages 825–832, 1999.

[6] S. Boettcher and A. G. Percus. Combining local search with co-evolution in a remarkably simple way. In *Proceedings of the 2000 Congress on Evolutionary Computation*, pages 1578–1584, 2000.

[7] T. J. Brunette and O. Brock. Improving protein structure prediction with model-based search. *Bioinformatics Suppl.*, 21:i66–i74, 2005.

[8] C. Clementi, H. Nymeyer, and J. N. Onuchic. Topological and energetic factors: What determines the structural details of the transition state ensemble and en-route intermediates for protein folding? an investigation for small globular proteins. *J. Mol. Biol.*, 298:937–953, 2000.

[9] F. Glover and M. Laguna. *Tabu Search.* Kluwer, Norwell, MA, 1997.

[10] U. H. E. Hansmann. Protein folding simulations in a deformed energy landscape. *Eur. Phys. J. B*, 12:607–611, 1999.

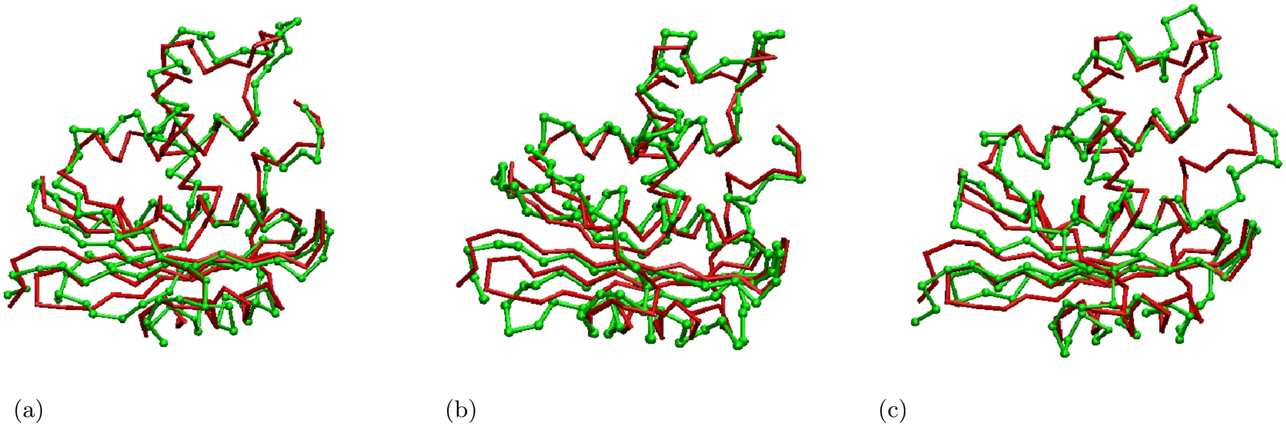


Figure 1: Alignment of the lowest energy structures, (a) -186.19 kT, (b) -120.23 kT and (c)-135.13 kT, found by EO meta-search, REMC and MC respectively for Ribonuclease H protein sequence (pdb id 2rn2) with the native state; RMSD is 2.55 Å, 2.80 Å, and 3.41 Å respectively.

[11] J. Karanicolas and C. L. Brooks III. The origin of asymmetry in the folding transition states of protein l and protein g. *Protein Sci.*, 11:2351–2361, 2002.

[12] A. Mitsutake, Y. Sugita, and Y. Okamoto. Generalized-ensemble algorithms for molecular simulations of biopolymers. *Biopolymers (Peptide Sci.)*, 60:96–123, 2001.

[13] V. Munoz and W. A. Eaton. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl. Acad. Sci. U.S.A.*, 96:11311–11316, 1999.

[14] J. T. Ngo, J. Marks, and M. Karplus. Computational complexity: Protein structure prediction and the levinthal paradox. *Protein Eng.*, 5:313–321, 1992.

[15] Y. Okamoto. Generalized-ensemble algorithms: Enhanced sampling techniques for monte carlo and molecular dynamic simulations. *J. Mol. Graphics Modell.*, 22:425–439, 2004.

[16] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes. The energy landscape perspective. *Annu. Rev. Phys. Chem.*, 48:545–600, 1997.

[17] A. R. Ortiz, A. Kolinski, P. Rotkiewicz, B. Ilkowski, and J. Skolnick. Ab initio folding of proteins using restrains derived from evolutionary information. *Proteins Suppl.*, 3:177–185, 1999.

[18] E. Paci, M. Vendruscolo, and M. Karplus. Validity of go models: Comparison with a solvent-shielded empirical energy decomposition. *Biophys. J.*, 83:3032–3038, 2002.

[19] K. Simons, C. Kooperberg, E. Huang, and D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring function. *J. Mol. Biol.*, 268:209–225, 1997.

[20] J. Skolnick, A. Kolinski, D. Kihara, M. Betancourt, P. Rotkiewicz, and M. Boniecki. Ab initio protein structure prediction via a combination of threading, lattice folding, clustering, and structure refinement. *Proteins Suppl.*, 5:149–156, 2001.

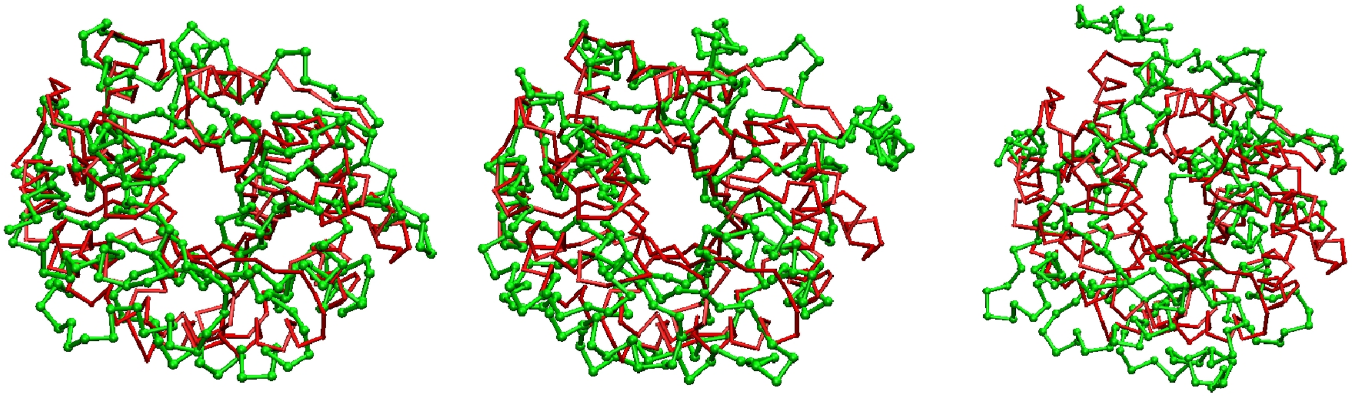
[21] Y. Sugita and Y. Okamoto. Replica-exchange multicanonical algorithm and multicanonical replica-exchange method for simulating systems with rough energy landscape. *Chem. Phys. Lett.*, 329:261–270, 2004.

[22] S. Takada. Go-ing for the prediction of protein folding mechanisms. *Proc. Natl. Acad. Sci. U.S.A.*, 96:11698–11700, 1999.

[23] H. Taketomi, Y. Ueda, and N. Go. Studies on protein folding, unfolding and fluctuations by computer simulation. i. the effect of specific amino acid sequence represented by specific inter-unit interactions. *Int. J. Pept. Protein Res.*, 7:445–459, 1975.

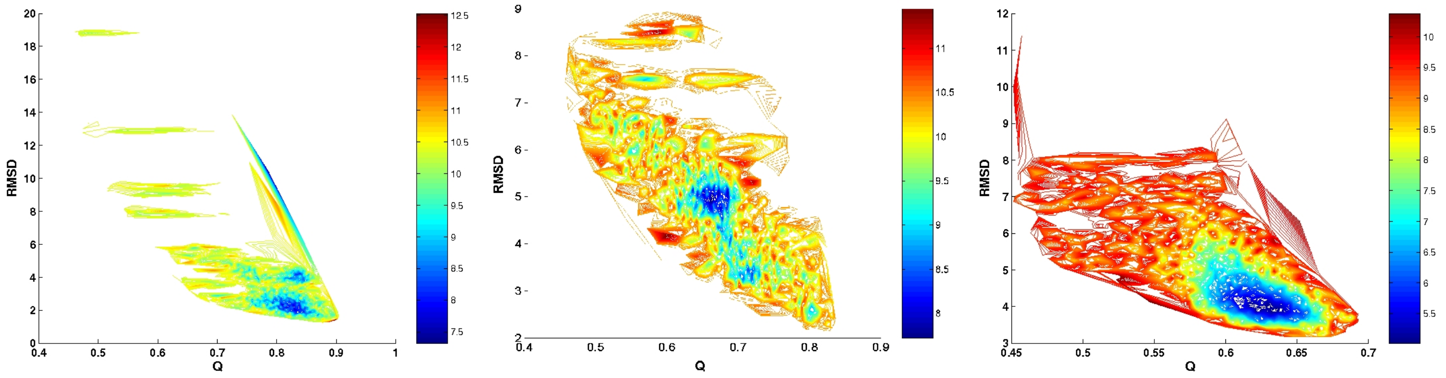
[24] R. Unger and J. Moult. Finding the lowest free energy conformation of a protein is a np-hard problem: Proof and implications. *Bull. Math. Biol.*, 55:1183–1198, 1993.

[25] Y. Zhang, A. K. Arakaki, and J. Skolnick. Tasser: An automated method for the prediction of protein tertiary structures in casp6. *Proteins Suppl.*, 61:91–8, 2005.



(a) (b) (c)

Figure 2: Alignment of the lowest energy structures, (a) -272.09 kT, (b) -190.91 kT and (c) -144.35 kT, found by EO meta-search, REMC and MC respectively for Xylanase protein sequence (pdb id 1bg4) with the native state; RMSD is 4.03 Å, 6.24 Å, and 13.03 Å respectively.



(a) (b) (c)

Figure 3: Contour plots of free energy as a function of fraction of the native contacts and RMSD from the native state for Ribonuclease H protein (pdb id 2rn2) for (a) EO, (b) REMC and (c) MC, cut-off time used was 5 CPU hours.