Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 EECS Building
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 06-012

Effective Optimization Algorithms for Fragment-assembly based Protein Structure Prediction

Kevin Deronne and George Karypis

March 27, 2006

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
27 MAR 2006		**00-00-2006 to 00-00-2006**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Effective Optimization Algorithms for Fragment-assembly based Protein Structure Prediction	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Department of Computer Science and Engineering,University of Minnesota,200 Union Street SE,Minneapolis,MN,55455-0159**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**11**	

Effective Optimization Algorithms for Fragment-assembly based Protein Structure Prediction*

Kevin W. DeRonne and George Karypis

Department of Computer Science & Engineering,
Digital Technology Center, Army HPC Research Center
University of Minnesota, Minneapolis, MN 55455

{deronne, karypis}@cs.umn.edu

Abstract

Despite recent developments in protein structure prediction, an accurate new fold prediction algorithm remains elusive. One of the challenges facing current techniques is the size and complexity of the space containing possible structures for a query sequence. Traditionally, to explore this space fragment assembly approaches to new fold prediction have used stochastic optimization techniques. Here we examine deterministic algorithms for optimizing scoring functions in protein structure prediction.

Two previously unused techniques are applied to the problem, called the Greedy algorithm and the Hill-climbing algorithm. The main difference between the two is that the latter implements a technique to overcome local minima. Experiments on a diverse set of 276 proteins show that the Hill-climbing algorithms consistently outperform existing approaches based on Simulated Annealing optimization (a traditional stochastic technique) in optimizing the root mean squared deviation (RMSD) between native and working structures.

1 Introduction

Reliably predicting protein structure from amino acid sequence remains a challenge in bioinformatics. Although the number of known structures continues to grow, many new sequences still lack a known homolog in the PDB [2], which makes it harder to predict structures for these sequences. The conditional existence of a known structural homolog to a query sequence commonly delineates a set of subproblems within the greater arena of protein structure prediction. For example, the biennial CASP competition[1] breaks down structure prediction as follows. In homologous fold recognition the structure of the query sequence is similar to a known structure for some other sequence. However, these two sequences have only a low (though detectable) similarity. In analogous fold recognition there exists a known structure similar to the correct structure of the query, but the sequence of that structure has no detectable similarity to the query sequence. Still more challenging is the problem of predicting the structure of a query sequence lacking a known structural relative, which is called new fold (NF) prediction.

Within the context of the NF problem knowledge-based methods have attracted increasing attention over the last decade. In CASP, prediction approaches that assemble fragments of known structures into a candidate structure [18, 7, 10] have consistently outperformed alternative methods, such as those based largely on explicit modeling of physical forces. Fragment assembly for a query protein begins with the selection of structural fragments based on sequence information. These fragments are then successively inserted into the query protein's structure, replacing the coordinates of the query with those of the fragment. The quality of this new structure is assessed by a scoring function. If the scoring function is a reliable measure of how close the working structure is to the native fold of the protein, then optimizing the function through fragment insertions will produce a good structure prediction. Thus, building a structure in this manner can break down into three main components: a fragment selection technique, an optimizer for the scoring function, and the scoring function itself.

To optimize the scoring function, all the leading assembly-based approaches use an algorithm involving a stochastic search (e.g. Simulated Annealing [18], genetic algorithms [7], or conformational space annealing [10]). One potential drawback of such techniques is that they can require extensive parameter tuning before producing good solutions.

*This work was supported in part by NSF EIA-9986042, ACI-0133464, IIS-0431135, and NIH RLM008713A; the Digital Technology Center at the University of Minnesota; and by the Army High Performance Computing Research Center (AHPCRC) under the auspices of the Department of the Army, Army Research Laboratory (ARL) under Cooperative Agreement number DAAD19-01-2-0014. The content of which does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred. Access to research and computing facilities was provided by the Digital Technology Center and the Minnesota Supercomputing Institute.

[1] http://predictioncenter.org/

In this paper we wish to examine the relative performance of deterministic and stochastic techniques to optimize a scoring function. The new algorithms presented below are inspired by techniques originally developed in the context of graph partitioning [4], and do not depend on a random element. The Greedy approach examines all possible fragment insertions at a given point and chooses the best one available. The Hill-climbing algorithm follows a similar strategy but allows for moves that reduce the score locally, provided that they lead to a better global score.

Several variables can affect the performance of optimization algorithms in the context of fragment-based *ab initio* structure prediction. For example, how many fragments per position are available to the optimizer, how long the fragments are, if they should be multiple sizes at different stages [18] or all different sizes used together [7], and other parameters specific to the optimizer can all influence the quality of the resulting structures.

Taking the above into account, we varied fragment length and number of fragments per position when comparing the performance of our optimization algorithms to that of a tuned Simulated Annealing approach. Our experiments test these algorithms on a diverse set of 276 protein domains derived from SCOP 1.69 [14]. The results of these experiments show that the Hill-climbing-based approaches are very effective in producing high-quality structures in a moderate amount of time, and that they generally outperform Simulated Annealing. On the average, Hill-climbing is able to produce structures that are 6% to 20% better (as measured by the root mean square deviation (RMSD) between the computed and its actual structure), and the relative advantage of Hill-climbing-based approaches improves with the length of the proteins.

2 Materials and Methods

2.1 Data

The performance of the optimization algorithms studied in this paper were evaluated using a set of proteins with known structure that was derived from SCOP 1.69 [14] as follows. Starting from the set of domains in SCOP, we first removed all membrane and cell surface proteins, and then used Astral's tools [3] to construct a set of proteins with less than 25% sequence identity. This set was further reduced by keeping only the structures that were determined by X-ray crystallography, filtering out any proteins with a resolution greater than 2.5Å, and removing any proteins with a $C_\alpha - C_\alpha$ distance greater than 3.8Å times their sequential separation[2].

[2]No bond lengths were modified to fit this constraint; proteins not satisfying it were simply removed from consideration.

Table 1: Number of sequences at various length intervals and SCOP class.

	Sequence Length			
SCOP Class	< 100	100–200	> 200	total
alpha	23	40	6	69
beta	23	27	18	69
alpha/beta	4	26	39	69
alpha+beta	15	36	17	69

The above steps resulted in a set of 2817 proteins. From this set, we selected a subset of 276 proteins (roughly 10%) to be used in evaluating the performance of the various optimization algorithms (i.e., a test set), whereas the remaining 2541 sequences were used as the database from whence to derive the structural fragments (i.e., a training set).[3] The test sequences, whose characteristics are summarized in Table 1, were selected to be diverse in length and secondary structure composition.

2.2 Neighbor Lists

As the search space for fragment assembly is much too vast, fragment-based *ab initio* structure prediction approaches must reduce the number of possible structures that they consider. They accomplish this primarily by restricting the number of structural fragments that can be used to replace each k-mer of the query sequence. In evaluating the various optimization algorithms developed in this work, we followed a methodology for identifying these structural fragments that is similar in spirit to that used by the Rosetta [18] system.

Consider a query sequence X of length l. For each position i, we identify a list (L_i) of n structural fragments by comparing the query sequence against the sequences of the proteins in the training set. For fragments of length k, these comparisons involve the k-mer of X starting at position i ($0 \leq i \leq l-k+1$) and all k-mers in the training set. The n structural fragments are selected so that their corresponding sequences have the highest profile-based score with the query sequence's k-mer. Throughout the rest of this paper, we will refer to the list L_i as the *neighbor list* of position i.

In our study we used neighbor lists containing fragments of a single length as well as neighbor lists containing fragments of different lengths. In the latter case we consider two different approaches to leveraging the varied length fragments. The first, referred to as *scan*, uses the fragment lengths in decreasing order. For example, if the neighbor lists contain structural fragments of length three, six, and nine, the algorithm starts by first optimizing the structure using only fragments of length nine, then fragments of length six, and finally fragments

[3]This dataset is available at http://www.cs.umn.edu/~deronne/supplement/optimize

of length three. Each one of these optimization phases terminates when the algorithm has finished (i.e., reached a local optimum or performed a predetermined number of iterations), and the resulting structure becomes the input to the subsequent optimization phase. The second approach for combining different length fragments is referred to as *pool*, and it optimizes the structure once, selecting fragments from any available length. Using any single length fragment in isolation, or using either scan or pool will be referred to as a *fragment selection scheme*.

2.2.1 Sequence Profiles

The comparisons between the query and the training sequences take advantage of evolutionary information by utilizing PSI-BLAST [1] generated sequence profiles.

The profile of a sequence X of length l is represented by two $l \times 20$ matrices. The first is its position-specific scoring matrix PSSM_X that is computed directly by PSI-BLAST. The rows of this matrix correspond to the various positions in X, while the columns correspond to the 20 distinct amino acids. The second matrix is its position-specific *frequency* matrix PSFM_X that contains the frequencies used by PSI-BLAST to derive PSSM_X. These frequencies (also referred to as *target frequencies* [13]) contain both the sequence-weighted observed frequencies (also referred to as *effective frequencies* [13]) and the BLOSUM62 [6] derived-pseudocounts [1]. For each row of a PSFM, the frequencies are scaled so that they add up to one. In the cases where PSI-BLAST could not produce meaningful alignments for a given position of X, the corresponding rows of the two matrices are derived from the scores and frequencies of BLOSUM62.

For our study, we used the version of the PSI-BLAST algorithm available in NCBI's blast release 2.2.10 to generate profiles for both the test and training sequences. These profiles were derived from the multiple sequence alignment constructed after five iterations using an e value of 10^{-2}. The PSI-BLAST search was performed against NCBI's nr database that was downloaded in November of 2004 and which contained 2,171,938 sequences.

2.2.2 Profile-to-Profile Scoring Method

The similarity score between a pair of k-mers (one from the query sequence and one from a sequence in the training set) was computed as the ungapped alignment score of the two k-mers whose aligned positions were scored using profile information.

Many different schemes have been developed for determining the similarity between profiles that combine information from the original sequence, position-specific scoring matrix, or position-specific target and/or effective frequencies [13, 21, 11]. In our work we use a scheme that is derived from PICASSO [5, 13] that was recently used in developing effective remote homology prediction and fold recognition algorithms [16]. Specifically, the similarity score between the ith position of protein X's profile, and the jth position of protein Y's profile is given by

$$S_{X,Y}(i,j) = \sum_{l=1}^{20} \text{PSFM}_X(i,l)\,\text{PSSM}_Y(j,l) \; + \sum_{l=1}^{20} \text{PSFM}_Y(j,l)\,\text{PSSM}_X(i,l), \tag{1}$$

where $\text{PSFM}_X(i,l)$ and $\text{PSSM}_X(i,l)$ are the values corresponding to the lth amino acid at the ith position of X's position-specific scoring and frequency matrices. $\text{PSFM}_Y(j,l)$ and $\text{PSSM}_Y(j,l)$ are defined in a similar fashion.

Equation 1 determines the similarity between two profile positions by weighting the position-specific scores of the first sequence according to the frequency at which the corresponding amino acid occurs in the second sequence's profile. The key difference between Equation 1 and the corresponding scheme used in [13] (therein referred to as PICASSO3), is that our measure uses the target frequencies, whereas the scheme of [13] is based on effective frequencies.

2.3 Protein Structure Representation

Internally, we consider only the positions of the C_α atoms, and we use a vector representation of the protein in lieu of ϕ and ψ backbone angles. Our protein construction approach uses the actual coordinates of the atoms in each fragment, rotated and translated into the reference frame of the working structure. Fragments are taken directly from known structures, and are chosen from the training dataset using the above profile-profile scoring methods.

2.4 Scoring Function

As the focus of this work is to develop and evaluate new optimization techniques, we use the RMSD between the predicted and native structure of a protein as the scoring function. Although such a function cannot serve as a predictive measure, we believe that using this as a scoring function allows for a clearer differentiation between the optimization process and the scoring function. In effect, we assume an ideal scoring function in order to test the optimization techniques.

2.5 Optimization Algorithms

In this study we compare the performance of three different optimization algorithms in the context of fragment assembly-based approaches for *ab initio* structure predictions. One of these algorithms, Simulated Annealing

[8], is currently a widely used method to solve such problems, whereas the other two algorithms, Greedy and Hill-climbing, are newly developed for this work.

The key operation in all three of these algorithms is the replacement of a k-mer starting at a particular position i, with that of a neighbor structure. We will refer to this operation as a *move*. A move is considered valid if, after inserting the fragment, it does not create any steric conflicts. A structure is considered to have a steric conflict if it contains a pair of C_α atoms within 2.5Å of one another. Also, for each valid move, its *gain* is defined as the improvement in the value of the scoring function between the working structure and the native structure of the protein.

2.5.1 Simulated Annealing (SA)

Simulated Annealing [8] is a generalization of the Monte Carlo [12] method for discrete optimization problems. This optimization approach is designed to mimic the process by which a material such as metal or glass cools. At high temperatures, the atoms of a metal can adopt configurations not available to them at lower temperatures—e.g., a metal can be a liquid rather than a solid. As the system cools, the atoms arrange themselves into more stable states, forming a stronger substance.

The Simulated Annealing (SA) algorithm proceeds in a series of discrete steps. In each step it randomly selects a valid move and performs it (i.e., inserts the selected fragment into the structure). This move can either improve or degrade the quality of the structure. If the move improves the quality, then the move is accepted. If it degrades the quality, then the move will still be accepted with probability

$$p = e^{\left(\frac{S_{old}-S_{new}}{T}\right)}, \qquad (2)$$

where T is the current temperature of the system, q_{old} is the score of the last state, and q_{new} is the score of the state in question. From Equation 2 we see that the likelihood of accepting a bad move is inversely related to the temperature and how much worse the new structure is from the current structure. That is, the optimizer will accept a very bad move with a higher probability if the temperature is high than if the temperature is low.

The algorithm begins with a high system temperature which it progressively decreases according to an *annealing schedule*. As the optimization must use finite steps, the cooling of the system cannot be continuous, but the annealing schedule can be modified to increase its smoothness. The annealing schedule depends on a combination of the number of total allowed moves and the number of steps in which to make those moves. Our implementation of Simulated Annealing, following the general framework employed in Rosetta [18], uses an annealing schedule that linearly decreases the temperature of the system to zero over a fixed number of cycles.

Simulated Annealing is a highly tunable optimization framework. The starting temperature and the annealing schedule can be varied to improve performance, and the performance of the algorithm depends greatly on these parameters. Section 3.2.1 describes how we arrive at the values for these parameters of SA as implemented in this study.

2.5.2 The Greedy Algorithm (G)

One of the characteristics of the Simulated Annealing algorithm is that it considers moves for insertion at random, irrespective of their gains. The Greedy algorithm that we present here selects maximum gain moves.

Specifically, the algorithm consists of two phases. In the first phase, called *initial structure generation*, the algorithm starts from a structure corresponding to a fully extended chain, and attempts to make a valid move at each position of the protein. This is achieved by scoring all neighbors in each neighbor list and inserting the best neighbor (i.e. the neighbor with the highest gain) from each list. If some positions have no valid moves on the first pass, the algorithm attempts to make moves at these positions after trying all positions once. This ensures that the algorithm makes moves at nearly every position down a chain, and also provides a good starting point for the next phase.

In the second phase, called *progressive refinement*, the algorithm repeatedly finds the maximum gain valid move over all positions of the chain, and if this move leads to a positive gain—i.e. it improves the value of the scoring function—the algorithm makes the move. This progressive refinement phase terminates upon failing to find any move to make. The Greedy algorithm is guaranteed to finish the progressive refinement phase in at least a local optimum.

2.5.3 Hill-Climbing (HC)

The Hill-climbing algorithm was developed to allow the Greedy algorithm to effectively climb out of locally optimal solutions. The key idea behind Hill-climbing is to not stop after achieving a local optimum but to continue performing valid moves in the hope of finding a better local or a (hopefully) global optimum.

Specifically, the Hill-climbing algorithm works as follows. The algorithm begins by applying the Greedy algorithm in order to reach a local optimum. At this point, it begins a sequence of iterations consisting of a *hill-climbing* phase, followed by a progressive refinement phase (as in the Greedy approach). In the hill-climbing phase, the algorithm performs a series of moves, each time selecting the highest gain valid move irrespective of whether or not it leads to a positive gain. If at any

point during this series of moves, the working structure achieves a score that is better than that of the structure at the beginning of the hill-climbing phase, this phase terminates and the algorithm enters the progressive refinement phase. The above sequence of iterations terminates when the hill-climbing phase is unable to produce a better structure after successively performing all best scoring valid moves.

Since the hill-climbing phase starts at a local optimum, its initial set of moves will lead to a structure whose quality (as measured by the scoring function) is worse than that at the beginning of the hill-climbing phase. However, subsequent moves can potentially lead to improvements that outweigh the initial quality degradation; thus allowing the algorithm to climb out of locally optimal solutions.

Move Locking As Hill-climbing allows negative gain moves, the algorithm can potentially oscillate between a local optimum and a non-optimal solution. To prevent this from happening, we implement a notion of move locking. After each move, a *lock* is placed on the move to prevent the algorithm from making this move again within the same phase. By doing so, we ensure the algorithm does not repeatedly perform the same sequence of moves; thus guaranteeing its termination after a finite number of moves. All locks are cleared at the end of a hill-climbing phase, allowing the search maximum freedom to proceed.

We investigate two different locking methods. The first, referred to as *fine-grain locking*, locks the single move made. The algorithm can subsequently select a different neighbor for insertion at this position. The second, referred to as *coarse-grain locking*, locks the position of the query sequence itself; preventing any further insertions at that position. In the case of pooling, coarse locking locks moves of all sizes.

Since fine-grain locking is less restrictive, we expect it to lead to better quality solutions. However, the advantage of coarse-grain locking is that each successive fragment insertion significantly reduces the set of fragments that need to be considered for future insertions; thus, leading to a faster optimization algorithm.

2.5.4 Efficient Checking of Steric Conflicts

One characteristic of the Greedy and Hill-climbing algorithms is their need to evaluate the validity of every available move after every insertion. This proves necessary because each insertion can potentially introduce new proximity conflicts. In an attempt to assuage the time requirement for this process, we have developed an efficient formulation for validity checking.

Recall that a valid move brings no two C_α atoms within 2.5Å of each other. To quickly determine if this proximity constraint holds, we impose a three-dimensional grid over the structure being built with boxes 2.5Å on each side. As each move is made, its atoms are added to the grid, and for each addition the surrounding 26 boxes are checked for atoms violating the proximity constraint. In this fashion we limit the number of actual distances that must be computed.

We further decrease the required time by sequentially checking neighbors at each position down the amino acid chain. All atoms upstream of the insertion point must be internally valid, as they have previously passed proximity checks. Thus, we need only examine those atoms at or downstream from the insertion. This saves on computation time within one iteration of checking all possible moves.

3 Experimental Evaluation

3.1 Performance of the Greedy and Hill-climbing Algorithms

To compare the effectiveness of the Greedy and Hill-climbing optimization techniques, we report results from a series of experiments in which we vary a number of parameters. Table 2 shows results for the Greedy and Hill-climbing optimization techniques using k-mer sizes of 9, 6, and 3 individually, as well as using the scan and pool techniques to combine them. Average times are also reported for each of these five fragment selection schemes.

Examining Table 2, we see that the Hill-climbing algorithm consistently outperforms the Greedy algorithm. As Hill-climbing includes running Greedy to convergence, the result is not surprising, and neither is the increased run-time that Hill-climbing requires. Both schemes seem to take advantage of the increased flexibility of smaller fragments and greater numbers of fragments per position. For example, on the average the 3-mer results are 20.6%, 27.4%, and 31.3% better than the corresponding 9-mer results for Greedy, Hill-climbing (coarse) (hereafter HC_c) and Hill-climbing (fine) (hereafter HC_f), respectively. Similarly, increasing the neighbor lists from 25 to 100 yields a 15.1%, 19.4%, and 28.2% improvement for Greedy, HC_c, and HC_f, respectively. These results also show that the search algorithms embedded in Greedy, HC_c, and HC_f are progressively more powerful as the size of the overall search space increases.

With respect to locking, a less restrictive fine-grained approach generally yields better results than a coarse-grained scheme. For example, averaging over all experiments, fine-grained locking yields a 13.7% improvement over coarse-grained locking. However, this increased performance comes at the cost of an increase in run-time of 700% on the average.

Comparing the performance of the scan and pooling

Table 2: Average values over 276 proteins optimized using Hill-climbing and different locking schemes. Times are in seconds and scores are in Å. Lower is better in both cases.

		$n = 25$		$n = 50$		$n = 75$		$n = 100$	
		Score	Time	Score	Time	Score	Time	Score	Time
Greedy	$k = 9$	12.17	14	11.14	16	10.74	18	10.44	20
	$k = 6$	10.84	15	10.28	19	9.66	24	9.44	27
	$k = 3$	9.51	25	8.99	34	8.58	45	8.25	58
	Scan	9.29	38	8.29	42	7.89	46	7.52	51
	Pool	9.06	54	8.36	69	7.81	215	7.61	106
Hill-climbing (coarse) (HC_c)	$k = 9$	10.18	22	9.32	30	8.73	55	8.38	72
	$k = 6$	8.92	30	8.03	56	7.51	98	7.24	116
	$k = 3$	7.31	59	6.86	129	6.29	231	6.11	313
	Scan	6.89	93	6.17	159	5.53	243	5.23	336
	Pool	6.89	175	6.32	379	5.79	543	5.54	959
Hill-climbing (fine) (HC_f)	$k = 9$	9.59	51	8.45	162	7.88	381	7.45	649
	$k = 6$	8.15	117	7.13	426	6.55	996	6.16	1731
	$k = 3$	6.84	261	5.94	1093	5.40	2175	4.83	3169
	Scan	6.43	330	5.41	1162	4.82	2028	4.33	2801
	Pool	6.00	903	5.03	3586	4.46	5246	4.11	7151

Table 3: SCOP classes and lengths for the tuning set.

SCOP identifier	length	SCOP class
d1jiwi_	105	beta
d1kpf__	111	alpha+beta
d2mcm__	112	beta
d1bea__	116	alpha
d1ca1_2	121	beta
d1jiga_	146	alpha
d1nbca_	155	beta
d1yaca_	204	alpha/beta
d1a8d_2	205	beta
d1aoza2	209	beta

methods to combine variable length k-mers we see that only in the HC_f algorithm does pool perform consistently better than scan (an average of 6.6%). The other schemes either have mixed results, or in the case of HC_c, pool performs somewhat worse. Results from the pool and scan settings indicate that Greedy and HC_c are not as effective at exploring the search space as HC_f.

3.2 Comparison with Simulated Annealing

3.2.1 Tuning the Performance of SA

Due to the sensitivity of Simulated Annealing to specific values for various parameters, we performed a search on a subset of the test proteins in an attempt to maximize the ability of SA to optimize the test structures. Specifically, we attempted to find values for two governing factors: the initial temperature T_0 and the number of moves nm. To this end, we selected ten medium length proteins of diverse secondary structural classification (see Table 3), and optimized them over various initial temperatures. The initial temperature that yielded the best average optimized RMSD was $T_0 = 0.1$ and we used this value in all subsequent experiments.

In addition to an initial temperature, when using Simulated Annealing one must select an appropriate annealing schedule. Our annealing schedule decreases the temperature linearly over 3500 cycles. This allows for a smooth cooling of the system. Over the course of these cycles, the algorithm attempts $\alpha \times (l \times n)$ moves, where α is an empirically determined scaling factor, l is the number of amino acids in the query protein, and n is the number of neighbors per position. Note that for the scan and pool techniques (see Section 2.2), we allow SA three times the number of attempted moves because the total number of neighbors is that much larger. In order to produce comparable run-times to the G, HC_c and HC_f schemes, α values of 20, 60, and 250 are employed, respectively.

In addition to the above experimental framework, we also investigated the extent to which the performance of SA improves if the total number of allowed moves were split across a number of independent runs (in this case the final structure will be the one that achieved the best score over the different runs). Our results showed that this approach resulted in comparable or somewhat worse results and we did not pursue it any further. Finally, following recent work [17] we allowed for a temporary increase in the temperature after 150 consecutive rejected moves.

3.2.2 Results

The Simulated Annealing results are summarized in Table 4. As we see in this table, Simulated Annealing consistently outperforms the Greedy scheme. For example, the average performance of SA with $\alpha = 20$ is 16.0% better than that obtained by G. These performance comparisons are obtained by averaging the ratios between the two schemes of the corresponding RMSDs over all fragment selection schemes and values of n. The superior performance of Simulated Annealing over Greedy is to be expected, as Greedy lacks any sort of hill-climbing ability, whereas the stochastic nature of Simulated Annealing allows it a chance of overcoming locally optimal solutions. In contrast, both the fine and coarse-

Table 4: Average values over 276 proteins optimized using Simulated Annealing. Times are in seconds and scores are in Å. Lower is better in both cases.

		$n = 25$		$n = 50$		$n = 75$		$n = 100$	
		Score	Time	Score	Time	Score	Time	Score	Time
$\alpha = 20$	$k = 9$	11.46	19	10.22	26	9.47	33	9.11	40
	$k = 6$	9.57	19	8.47	27	7.99	34	7.74	42
	$k = 3$	7.60	20	6.84	28	6.58	36	6.44	44
	Scan	7.59	59	6.76	82	6.31	105	6.32	129
	Pool	7.38	59	6.93	83	7.03	105	7.18	129
$\alpha = 60$	$k = 9$	10.73	33	9.41	55	8.85	77	8.63	99
	$k = 6$	8.72	34	7.95	57	7.71	80	7.57	103
	$k = 3$	6.85	36	6.57	60	6.64	85	6.66	109
	Scan	6.91	105	6.56	177	6.42	252	6.47	316
	Pool	7.26	105	7.20	176	7.23	246	7.28	317
$\alpha = 250$	$k = 9$	9.80	103	8.88	196	8.70	289	8.62	381
	$k = 6$	8.15	107	7.85	204	7.75	301	7.77	397
	$k = 3$	6.87	114	6.89	219	6.94	325	6.92	433
	Scan	6.87	327	6.65	621	6.63	933	6.61	1217
	Pool	7.39	329	7.30	626	7.40	924	7.36	1218

The values of α in the above table scale the number of moves Simulated Annealing is allowed to make. In our case, the total number of moves is $\alpha \times (l \times n)$ where l is the length of the protein being optimized and n is the number of neighbors per position.

Table 5: Average values over the longest 138 proteins optimized using Hill-climbing and different locking schemes. Times are in seconds and scores are in Å. Lower is better in both cases.

		$n = 25$		$n = 50$		$n = 75$		$n = 100$	
		Score	Time	Score	Time	Score	Time	Score	Time
Greedy	$k = 9$	14.67	19	13.46	24	13.12	27	12.80	32
	$k = 6$	13.48	23	13.11	30	12.22	38	11.98	45
	$k = 3$	11.99	42	11.74	58	11.32	79	11.00	103
	Scan	11.86	72	10.66	79	10.23	97	10.07	109
	Pool	11.62	86	10.97	114	10.23	144	10.22	183
Hill-climbing (coarse) (HC_c)	$k = 9$	12.22	37	11.17	50	10.50	100	10.07	133
	$k = 6$	11.09	51	9.97	101	9.25	183	9.06	218
	$k = 3$	9.10	106	8.92	242	8.23	439	8.13	598
	Scan	8.77	149	7.53	296	7.29	481	6.90	627
	Pool	8.69	322	8.21	719	7.43	1010	7.25	1859
Hill-climbing (fine) (HC_f)	$k = 9$	11.57	91	10.15	302	9.46	720	8.94	1229
	$k = 6$	10.01	218	8.81	809	8.03	1904	7.52	3315
	$k = 3$	8.60	490	7.70	2085	7.14	5167	6.54	8892
	Scan	8.13	597	6.79	2024	6.19	4431	5.45	5976
	Pool	7.44	1720	6.17	6936	5.43	10046	5.01	16807

locking versions of Hill-climbing outperform SA. More concretely, on the average HC_c performs 6.2% better than SA with $\alpha = 60$, and HC_f performs 19.2% better than SA with $\alpha = 250$.

Analyzing the performance of Simulated Annealing with respect to the value of α, we see that while Simulated Annealing shows an average improvement of 3.0% when α is increased from 20 to 60, the performance deteriorates by an average of 0.2% when α is increased from 60 to 250. This indicates that further increasing the value of α may not lead to performance comparable to that of the Greedy and Hill-climbing schemes.

Also note that in some of the results shown in Table 4, the performance occasionally decreases as the α value increases. This ostensibly strange result comes from the dependence of the cooling process on the number of allowed moves, in which the value of α plays a role. For all entries in Table 4 the annealing schedule will cool the system over a fixed number of steps, but the number of moves made will vary greatly. Thus, in order to keep the cooling of the system linear we vary the number of moves allowed before the system reduces its temperature. As a result, different values of α can lead to different randomly chosen optimization paths.

Comparing the performance of the various optimization schemes with respect to the various fragment selection schemes, we see two interesting trends. First, Greedy, HC_c, and HC_f tend to produce better results (i.e., lower RMSD values) using the scan fragment selection scheme as compared to those obtained for $k = 3$; whereas the performance of Simulated Annealing does not significantly improve. For example, on the average HC_f with scan does 9.0% better than HC_f with $k = 3$ over the different values of n; but the corresponding average improvement of SA is only 1.3%. Second, the performance of SA deteriorates (by 10.0% on the average) when the different length k-mers are used via the pool method, whereas the performance of HC_f improves

Table 6: Average values over the longerst 138 proteins optimized using Simulated Annealing. Times are in seconds and scores are in Å. Lower is better in both cases.

		$n = 25$		$n = 50$		$n = 75$		$n = 100$	
		Score	Time	Score	Time	Score	Time	Score	Time
$\alpha = 20$	$k = 9$	13.83	30	12.44	43	11.39	56	11.19	69
	$k = 6$	11.73	31	10.55	44	9.93	57	9.66	71
	$k = 3$	9.61	31	8.74	46	8.50	60	8.33	74
	Scan	9.67	94	8.73	135	8.10	175	8.23	218
	Pool	9.34	94	8.93	136	9.15	176	9.44	218
$\alpha = 60$	$k = 9$	12.93	56	11.50	96	10.77	135	10.77	175
	$k = 6$	10.60	58	9.98	99	9.76	139	9.66	181
	$k = 3$	8.65	60	8.56	104	8.68	148	8.69	191
	Scan	8.93	176	8.57	306	8.28	440	8.35	552
	Pool	9.46	176	9.37	303	9.41	429	9.39	555
$\alpha = 250$	$k = 9$	11.94	182	11.00	348	10.89	515	10.86	680
	$k = 6$	10.20	189	10.01	362	9.90	536	9.93	709
	$k = 3$	8.88	200	8.91	388	8.94	578	8.93	771
	Scan	8.97	575	8.49	1102	8.49	1662	8.45	2168
	Pool	9.51	578	9.40	1110	9.57	1642	9.46	2167

The values of α in the above table scale the number of moves Simulated Annealing is allowed to make. In our case, the total number of moves is $\alpha \times (l \times n)$ where l is the length of the protein being optimized and n is the number of neighbors per position.

(by 6.6% on average). We are currently investigating the sources of these behaviors, but one possible explanation of the latter observation is that Simulated Annealing has a bias towards smaller fragments. This bias might result because an insertion of a bad 3-mer will degrade the structure less than that of a bad 9-mer, and as a result, the likelihood of accepting the former move will be higher (Equation 2). This may reduce the optimizers ability to effectively utilize the variable length k-mers.

Performance on Longest Sequences In order to gain a better understanding of how the optimization schemes perform, we focus on the longer half of the test proteins. Average RMSDs and times for the Greedy and Hill-climbing schemes are shown in Table 5, and average RMSDs and times for Simulated Annealing are shown in Table 6.

In general, the trends in these tables agree with the trends in the average values over all the proteins. However, one key difference is that the relative improvement of the Hill-climbing scheme over Simulated Annealing is higher, while that of Greedy is lower. For example, comparing G and SA for $\alpha = 20$, SA performs 17.0% better, as opposed to 16.0% for the full average. Comparing with SA for $\alpha = 60$, HC_c performs 7.0% better as opposed to 6.2% for the full average. Finally, comparing with SA for $\alpha = 250$, HC_f is 21.1% better, as opposed to 19.2% for the full average. These results suggest that, in the context of a larger search space, a hill-climbing ability is important, and that the hill-climbing abilities of HC_c and HC_f are better than those of SA.

4 Discussion and Conclusions

This paper presents two new techniques for optimizing scoring functions for protein structure prediction. One of these approaches, HC_c, using the scan technique, reaches better solutions than Simulated Annealing in comparable time. The performance of SA seems to saturate beyond $\alpha = 60$, but HC_f will make use of an increased time allowance, finding the best solutions of all the examined algorithms. Furthermore, experiments with variations on the number of moves available to the optimizer demonstrate that the Hill-climbing approach makes better use of an expanded search space than Simulated Annealing. Additionally, Simulated Annealing requires the hand-tuning of several parameters, including the total number of moves, the initial temperature, and the annealing schedule. One of the main advantages of using schemes like Greedy and Hill-climbing is that they do not rely on such parameters.

Recently, greedy techniques have been applied to problems similar to the one this paper addresses. The first problem is to determine a set of representative fragments for use in decoy structure construction [15, 9]. The second problem is to reconstruct a native protein fold given such a set of representative fragments [19, 20]. The greedy approaches used for both these problems traverse the query sequence in order, inserting the best found fragment for each position. As an extension, the algorithms build multiple structures simultaneously in the search for a better structure. While such approaches have the ability to avoid local minima, they lack an explicit notion of hill-climbing.

The techniques this paper describes could be modified to solve either of the above two problems. To build a representative set of fragments, one could track the frequency of fragment use within multiple Hill-climbing optimizations of different proteins. This would yield a large set of fragments, which could serve as input to a clustering algorithm. The centroids of these clusters could then be used in decoy construction. In order to construct a native fold from these fragments one need only

restrict the move options of Hill-climbing to the representative set. We are currently working on adapting our algorithms to solve these problems.

References

[1] S. F. Altschul, L. T. Madden, A. A. Schffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–402, 1997.

[2] H.M. Berman, J.Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The protein data bank. *Nucleic Acids Research*, 2000.

[3] J. Chandonia, G. Hon, N. S. Walker, L. Lo Conte, P. Koehl, M. Levitt, and S. E. Brenner. The astral compendium in 2004. *Nucleic Acids Research*, 2004.

[4] C.M. Fiduccia and R.M. Mattheyses. A linear-time heuristic for improving network partitions. *Proceedings of the 19th Design Automation Conference*, 1982.

[5] A. Heger and L. Holm. Picasso: generating a covering set of protein family profiles. *Bioinformatics*, 2001.

[6] S. Henikoff and J. G. Henikoff. Amino acid subsitution matrices from protein blocks. *PNAS*, 89:10915–10919, 1992.

[7] K. Karplus, R. Karchin, J. Draper, J. Casper, Y. Mandel-Gutfreund, M. Diekhans, and R. Hughey. Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *PROTEINS: Structure, Function and Genetics*, 2003.

[8] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[9] R. Kolodny, P. Koehl, L. Guibas, and M. Levitt. Small libraries of protein fragments model native protein structures accurately. *Journal of Molecular Biology*, 323:297–307, 2002.

[10] J. Lee, S. Kim, K. Joo, I. Kim, and J. Lee. Prediction of protein tertiary structure using profesy, a novel method based on fragment assembly and conformational space annealing. *PROTEINS: Structure, function and bioinformatics*, 2004.

[11] M. Marti-Renom, M. Madhusudhan, and A. Sali. Alignment of protein sequences by their profiles. *Protein Science*, 13:1071–1087, 2004.

[12] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[13] D. Mittelman, R. Sadreyev, and N. Grishin. Probabilistic scoring measures for profile-profile comparison yield more accurate short seed alignments. *Bioinformatics*, 19(12):1531–1539, 2003.

[14] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536–540, 1995.

[15] B. H. Park and M. Levitt. The complexity and accuracy of discrete state models of protein structure. *Journal of Molecular Biology*, 249:493–507, 1995.

[16] H. Rangwala and G. Karypis. Profile based direct kernels for remote homology detection and fold recognition. *Bioinformatics*, 21:4239–4247, 2005.

[17] C. A. Rohl, C. E. M. Strauss, K. M. S. Misura, and D. Baker. Protein structure prediction using rosetta. *Methods in Enzymology*, 2004.

[18] K. T. 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 functions. *Journal of Molecular Biology*, 1997.

[19] P. Tuffery and P. Derreumaux. Dependency between consecutive local conformations helps assemble protein structures from secondary structures using go potential and greedy algorithm. *Protein Science*, 61:732–740, 2005.

[20] P. Tuffery, F. Guyon, and P. Derreumaux. Improved greedy algorithm for protein structure reconstruction. *Journal of Computational Chemistry*, 26:506–513, 2005.

[21] G. Wang and R. L. Dunbrack JR. Scoring profile-to-profile sequence alignments. *Protein Science*, 13:1612–1626, 2004.