Protein Structure Prediction on the Face Centered Cubic Lattice by Local Search

Manuel Cebrián and **Ivan Dotú** and **Pascal Van Hentenryck**
Department of Computer Science
Brown University
Box 1910, Providence, RI 02912

Peter Clote
Biology Department
Boston College
Chestnut Hill, MA 02467

Abstract

Ab initio protein structure prediction is an important problem for which several algorithms have been developed. Algorithms differ by how they represent 3D protein conformations (on-lattice, off-lattice, coarse-grain or fine-grain model), by the energy model they consider, and whether they are heuristic or exact algorithms. This paper presents a local search algorithm to find the native state for the Hydrophobic-Polar (HP) model on the Face Centered Cubic (FCC) lattice; i.e. a self-avoiding walk on the FCC lattice with maximum number of H-H contacts. The algorithm relies on a randomized, structured initialization, a novel fitness function to guide the search, and efficient data structures to obtain self-avoiding walks. Experimental results on benchmark instances show the efficiency and excellent performance of our algorithm, and illustrate the biological pertinence of the FCC lattice.

Introduction

The prediction of 3-dimensional structure of a protein, given only its amino acid sequence, i.e., *protein structure prediction*, remains one of the oldest, most recalcitrant, yet most important problems in computational biology. In 1968, C. Levinthal first raised the question of how a protein can find its *native state*, i.e., its unique 3-dimensional conformation, rapidly (within milliseconds to seconds), although there are exponentially many possible conformations. Subsequently, in a celebrated experiment in which bovine pancreatic ribonuclease A was denatured (unfolded) by the addition of urea, then found to return to its native conformation after removal of denaturant urea, C. B. Anfinsen (1973) provided the first evidence that, at least for a certain class of proteins, the native state of a protein is its minimum free energy conformation, and that no specific folding pathways or chaperone molecules appear to be necessary. In 1972, the Swedish Royal Academy of Sciences granted the 1972 Nobel Prize in Chemistry to Anfinsen for "...studies on ribonuclease, in particular the relationship between the amino acid sequence and the biologically active conformation ..." (Anfinsen 1972).

From Anfinsen's work, it is now generally assumed that the native state of a protein is its minimum free energy (MFE) conformation, and thus is a computational problem, albeit the space of possible conformations is exponentially large. One of the first mathematical models for proteins is the lattice HP-model, first introduced by Lau and Dill (1989) for the 2-dimensional square lattice. Given a sequence of hydrophobic (H) and polar (P), aka hydrophilic, residues, the energy of a self-avoiding walk on the lattice is defined to be minus 1 times the number of non-contiguous H-H contacts at unit distance. For such a simple model, the native state is *degenerate* in the sense that there may be many minimum energy conformations; nevertheless, there is a well defined minimum energy E_0, dependent only on the input HP-sequence, and the formulation of such a clean and simple model stimulated the development of various folding algorithms, as well as efforts to better understand energetics.

Despite its simplicity, finding a minimum energy conformation for the HP-model was shown to be NP-complete for the 2-dimensional lattice by (Crescenzi *et al.* 1998) and for the 3-dimensional cubic lattice by Berger and Leighton (1998). Yue and Dill (1996) applied "constraint-based exhaustive search" to determine the minimum energy conformation(s) of several small proteins including crambin, when represented as HP-sequences on the cubic lattice. Necessarily, any exhaustive search is limited to very small proteins, since the number of conformations for an n-mer on the 3-dimensional cubic lattice is estimated to be approximately 4.5^n (Madras & Slade 1996).

This paper presents a tabu search algorithm to predict protein tertiary structure under the Face Centered Cubic lattice HP-model. The algorithm features a randomized, structured initialization, a one-monomer move neighborhood, and a new fitness function to guide the search. The configurations explored by the algorithm are always feasible, yielding an anytime algorithm for producing 3-dimensional protein structures. The algorithm was applied to the Harvard instances (Yue *et al.* 1995), producing (to our knowledge) the first foldings of these instances on the FCC lattice, Experimental results indicate the fundamental benefits of using a FCC lattice since the resulting foldings have significantly lower energies. Moreover, experimental results show that these foldings can be obtained in reasonable time.

The rest of the paper is organized as follows. It first formalizes the problem and discusses related work. The paper then presents the model, and the local search algorithm. The last two sections present the experimental results, the con-

clusions, and future work.

Problem Formalization

The cubic lattice suffers from a fundamental flaw in modeling real biopolymers; namely, if the parity of the position in the linear chain of any two residues is the same, then they cannot form a contact, i.e., be at unit distance in any conformation. For this reason, this paper considers the face-centered cubic (FCC) lattice which is known to better model biopolymers. Covell and Jernigan (1990) have shown that the FCC lattice, proven to admit the tightest packing of spheres (Cipra 1998), is the most appropriate 3-dimensional lattice for fitting protein C_α-atoms as a self-avoiding walk, and that *root mean square deviation* (`rms`) values are smaller for the FCC lattice than for the cubic, body-centered cubic and tetrahedral lattices. Here `rms` between two C_α-traces $(p_1, p_2, \ldots, p_n)$ and $(q_1, q_2, \ldots, q_n)$, where $p_i, q_i \in \mathbf{R}^3$, is given by $\sqrt{\frac{\sum_{i=1}^{n}(p_i - q_i)^2}{n}}$.

Formally, a *lattice* is defined to be the set of points in $\mathbb{Z}^n$ that are *integral* linear combinations of vectors having integral coordinates; i.e.

$$L = \left\{ \sum_{i=1}^{k} a_i \vec{v}_i : a_i \in \mathbb{Z} \right\} \tag{1}$$

where $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{Z}^n$. In this paper, $n = 3$, i.e., only lattices $L \subseteq \mathbb{Z}^3$ are considered. If k is minimum for which (1) holds, then $\vec{v}_1, \ldots, \vec{v}_k$ form a *basis*, and k is said to be the *dimension* (also called *coordination* or *contact* number) of L. Two lattice points $p, q \in L$ are said to be *in contact* if $q = p + \vec{v}_i$ for some vector $\vec{v}_i$ in the basis of L.

The cubic lattice is formally defined as the closure of the basis vectors $(1, 0, 0)$, $(0, 1, 0)$, $(0, 0, 1)$ under all integral linear combinations. In contrast, the face-centered cubic (FCC) lattice is generated by the following 12 basis vectors, which are identified with compass directions (Will 2005):

$N : (1, 1, 0)$	$S : (-1, -1, 0)$	$W : (-1, 1, 0)$
$E : (1, -1, 0)$	$NW_+ : (0, 1, 1)$	$NW_- : (0, 1, -1)$
$NE_+ : (1, 0, 1)$	$NE_- : (1, 0, _- 1)$	$SE_+ : (0, -1, 1)$
$SW_+ : (-1, 0, 1)$	$SE_- : (0, -1, -1)$	$SW_- : (-1, 0, -1)$.

It follows that the FCC lattice consists of all integer points (x, y, z), such that $(x + y + z) \bmod 2 = 0$. Moreover, $p = (x, y, z)$ and $q = (x', y', z')$ are in **contact**, denoted by $co(p, q)$, if $(x - x') + (y - y') + (z - z') \bmod 2 \equiv 0$, $|x - x'| \leq 1$, $|y - y'| \leq 1$, and $|z - z'| \leq 1$. We will sometimes state that lattice points p, q are at *unit distance*, when we formally mean that they are in contact, hence are at Euclidean distance $\sqrt{2}$ on the FCC lattice.

Given a sequence S of length n, let *HH* denote the set of pairs (i, j) such that $S_i = S_j = H$ and let *CHH* denote the subset of HH for which $j = i + 1$. The protein prediction problem for the HP-model on the FCC lattice can be defined as follows.

> Given a protein sequence S (sequence of amino acids) of length n, find a self-avoiding walk $p_1, \ldots, p_n$ on the FCC lattice that minimizes the energy
>
> $$\sum_{(i,j) \in CHH} E(p_i, p_j) - \sum_{(i,j) \in HH} E(p_i, p_j). \tag{2}$$
>
> Here, p_i is the lattice position of the ith monomer, and energy $E(p_i, p_j) = -1$ if both p_i, p_j are neighbours in the lattice and 0 otherwise, so that equation (2) represents the energy for all non-contiguous H-H contacts.

Related Work

Coarse-grain lattice models have been heavily studied in the context of protein folding. In (Šali, Shakhnovich, & Karplus 1994a; 1994b), Šali et al. measured the average time required to reach the native state, formally the *mean first passage time* (MFPT), for a 27-mer on the $3 \times 3 \times 3$ cubic lattice using Monte Carlo simulation of protein folding. They claimed to have solved the Levinthal paradox by showing that thermodynamics suffices to drive a protein to rapidly find its native state. Subsequently P. Clote (1999) applied Sinclair's work on rapidly mixing Markov chains (Sinclair 1993) towards a mathematical analysis of (Šali, Shakhnovich, & Karplus 1994a).

Yue and Dill (2000) described an improvement to the algorithm presented in (Yue & Dill 1996) with the Constraint Hydrophobic Core Construction (CHCC) algorithm which was benchmarked with sample HP-sequences for the HP-model on the cubic lattice. Hart and Istrail (1996) described a novel approximation algorithm, guaranteed to provide within quadratic time a conformation whose energy is no worse than three-eighths that of the optimal.

Backofen, Will, and Clote (2000) developed a genetic algorithm to fold HP-sequences on arbitrary lattices (including FCC). Using automorphism groups to handle arbitrary lattices, the algorithm supported *pivot* moves to determine optimal conformations, using a *"hydrophobic energy"* term, defined by a contact potential involving normalized polar requirement hydrophobicity values (Woese *et al.* 1966). Using the symmetry-breaking algorithm (Crawford *et al.* 1996), Backofen and Will (2002) developed a constraint-programming algorithm to search for minimum energy conformations on the cubic and face-centered cubic lattice for larger HP-sequences than could be handled by previous algorithms. No results were given on the Harvard instances.

In (Tapia *et al.* 2007) Amato and co-workers applied motion planning from robotics to sample the folding landscape of (simple) proteins using kinetics. Zhang et al. (2007) proposed a new Monte Carlo method, called *fragment regrowth via energy-guided sequential sampling* (FRESS), benchmarked on the HP-model for lattices in two and three dimensions. The algorithm was implemented for the cubic latttice. In (Kou, Oh, & Wong 2006) Kou et al. describe a new equi-energy (EE) sampling approach to estimate the density of states (i.e. histogram of number of conformations have energy $-k$, for all values of k) for the HP-model on the 2-dimensional lattice. Also, Tabu search has been applied with relatice success to the 2D lattice (Jiang *et al.* 2003) and to the cubic lattice (Blazewicz *et al.* 2005).

Our Model

This section presents our model for protein structure prediction. The model associates a decision variable v_i with every amino acid's position on the lattice. In other words, given

a sequence of amino acids S such that $|S| = n$, the variable v_i takes its value in $\mathbb{Z}^3$ and represents the x, y, and z coordinates of the ith amino acid of S in the lattice. These variables must satisfy the following constraints:

- Self-Avoiding Walk: For all $i \neq j$: $v_i \neq v_j$.
- FCC Lattice Constraints: The sum of the coordinates of each point must be even.
- Adjacency: Two consecutive elements i and $i+1$ must be neighbors in the lattice, i.e. in *contact* or at unit distance (as mentioned before, on the FCC, this means at Euclidean distance $\sqrt{2}$).

These are all hard constraints. They will hold initially and be preserved across local moves. In the following, we use σ to denote a complete assignment of the variables v_i that satisfies all the constraints.

The Fitness Function

The HP-model for protein structure predicate features an energy function which is rather poor in guiding the search towards high-quality solutions. Indeed, the number of H-H contacts only increases (decreases) when the algorithm positions (separates) two H amino acids at (from) unit distance; any other does not change the energy. As a result, a local-search algorithm based on such an objective will mostly perform a random walk.

To address this issue, our algorithm introduces a fitness function to guide the algorithm effectively. Define *distance* between two amino acids as $d(i,j)^2 = (x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2$, i.e., the square of the Euclidean distance between the ith and the jth amino acids in the current conformation of a sequence S of length n. Now consider the deviation from the unit distance (to the power of 2) to be $dv(i,j) = d(i,j)^2 - 2$. Our fitness function (or *cost*) is:

$$f(\sigma) = \sum_{i,j:i+1<j}^{n} (dv(i,j))^k \times (s_i = H, s_j = H)$$

where the sum is over i, j such that $i + 1 < j$ and $k \geq 1$ is a parameter of the algorithm. In particular, larger values of k give more weight to unit distances. Observe that these values are only defined when i and j correspond to H-type amino acids. The fitness function f is thus a measure of the deviation from the unit distance for every pair of (non consecutive) H-type amino acids. Therefore, in order to maximize the number of HH contacts, we need to minimize f.

One may view f as a guide towards a compact structure where H-amino acids are close together, thus yielding several HH contacts. It is clear that, in order to achieve unit distance between H-type amino acids, they need to be close to each other. The impact of this fitness function will be better understood in the Experimental Results section. Note that $f(\sigma^*) = 0$ means that all pairs of H-type amino acids are at unit distance in σ^*.

The Alldifferent Constraint

One of the constraints requires that all amino acid positions on the lattice be different. Representing this constraint explicitly is very costly and slows down the search considerably. Instead, the algorithm maintains the constraint implicitly. Each time a local move is performed on v_i, the algorithm only checks those amino acids v_j ($j \neq i$) whose norm is equal to $||v_i||$, since $v_i = v_j \Rightarrow ||v_i|| = |v_j||$. The constraint check is performed in $O(1)$ expected time, since the number of amino acids with the same norm is very low, even in the latest stages of the search process when the molecule is densely packed.

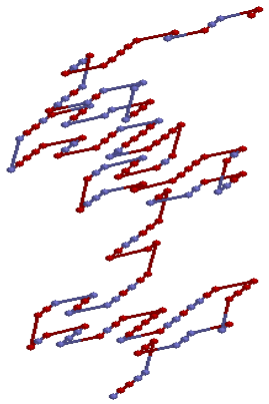


Figure 1: Initial structure for instance S4 (Will 2005, p. 130).

The Neighborhood

In this work, we allowed only one-monomer moves, in which only a single monomer changes position between two successive conformations. Our benchmarks suggest that a one-monomer move set suffices for good results on the FCC lattice, although since this is not the case for the CC lattice, Šali et al. (Šali, Shakhnovich, & Karplus 1994a) considered *crankshaft* (2-monomer) moves as well. If $p_1, \ldots, p_n$ denote current positions of monomers $1, \ldots, n$, then define the neighborhood $N(i)$ of the ith monomer as the set P of points p such that $d(p, p_i)^2 = 2$, $p \in P$. A neighborhood of a tentative solution σ consists in moving monomer i to one of its neighbors, i.e., a point in

$$S(\sigma, i) = \{p \in \mathbb{Z}^3 \mid p \in N(i-1) \wedge p \in N(i+1)\}$$

The neighborhood of σ can then be defined as

$$\mathcal{N}(\sigma) = \{(i,p) \mid 0 < i < n \ \wedge \ p \in S(\sigma, i)\}.$$

A Randomized Initialization

The initial solution has a significant impact on the quality and the speed of the local search algorithm. Given our one-monomer neighborhood and our fitness function, it is important to generate a feasible and compact initial solution with some HH contacts. The initialization iterates the following steps while there are amino acids to place.

1. Repeat a random number of times
   (a) Repeat Forward for a random number of steps.
   (b) Move Left.
   (c) Repeat Backward for a random number of steps.

```
1.  PSPLS(S)
2.    forall i ∈ S
3.      tabu[i] ← {};
4.    σ ← initial configuration;
5.    σ* ← σ;
6.    l ← 0;
7.    s ← 0;
8.    while l ≤ maxIt  do
9.      select (i, p) ∈ N(σ)
          minimizing f(σ[v_i ← p]);
10.     τ ← RANDOM([4,n/2]);
11.     tabu[i] ←
          tabu[i] ∪ {move(i, p, σ)};
12.     σ ← σ[v_1 ↔ p];
13.     if f(σ) < f(σ*) then
14.       σ* ← σ;
15.       s ← 0;
16.     else if s > maxStable then
17.       σ ←random configuration;
18.       s ← 0;
19.       forall i ∈ S do
20.         tabu[i] = {};
21.     else
22.       s++;
23.     l++;
```

Figure 2: The Local Search Algorithm.

2. Move Up.
3. Switch moves with their opposites (e.g., Forward becomes Backward and Left becomes Right).

An initial configuration for the "S4" instance is depicted in figure 1.

The Tabu-Search Algorithm

We are now ready to present the basic local search algorithm. The algorithm, depicted in Figure 2, a tabu search with a restarting component. Lines 2-7 perform the initializations. In particular, the tabu list is initialized in lines 2-3, the initial solution is generated in line 4, while lines 6 and 7 initialize the iteration counter k, and the stability counter s. The initial configuration σ is obtained in the manner explained above. The best solution found so far σ^* is initialized to σ.

The tabu list is distributed across the amino acids and maintains a set of moves. A move is formally defined as

$$move(i, p, \sigma) = p - \sigma(v_{i-1})$$

where $\sigma(v_{i-1})$ denotes the position of amino acid $i-1$ in assignment σ and p is the new position for amino acid i. Note that the subtraction of v_{i-1} from p yields one of the basic vectors previously defined (N,S,W,E, ...). The tabu tenure is randomly selected between 4 and half the length of the sequence.

The core of the algorithm is given in lines 8-23, where local moves are iterated for a number of iterations. The local move is selected in line 9. Here, we use $\sigma[v_i \leftarrow p]$ to denote the solution obtained by changing the value of v_i to p in σ. The key idea is to select the best move in the neighborhood which is not tabu (meaning it has been previously performed) or which improves the best solution. The tabu list is updated in line 11, and the new tentative solution is computed in line 12. Lines 13-15 update the best solution, while lines 16-20 specify the restarting component.

The restarting component simply reinitializes the search from a random configuration whenever the best solution found so far has not been improved upon for *maxStable* iterations. Note that the stability counter s is incremented in line 22 and reset to zero in line 15 (when a new best solution is found) and in line 18 (when the search is restarted).

Dealing with Hs and Ps

The algorithm also differentiates H-type and P-type amino acids, For H-type amino acids, it performs a complete exploration of the neighborhood and chooses the best possible. However, for P-type amino acids, all neighbors have the same the fitness function. Therefore, since we are always choosing the best neighbor, the algorithm moves a P-type amino acid only if all H-type moves yield a solution with at most the same cost. In that case, the algorithm chooses an amino acid and a move completely at random. This simple optimization produces significant reduction in the computational cost of the algorithm.

Experimental Results

All the results presented in this section have been produced by a C implementation of the algorithm, run on a single core of a 60 Intel based, dual-core, dual processor, Dell Poweredge 1855 blade server. Each blade has 8G of memory and a 300G local disk, and each execution was carried out on a single core. The maximum number of iterations was set to 10 million, and the stability parameter to 10000. All tables show results for different values of the k parameter (ranging from 1 to 3). All best results are given as supplemental material to this publication.

The Harvard Instances

Reference (Yue *et al.* 1995) contains a comparison of several methods folding 10 different proteins on the cubic lattice. These proteins are called "Harvard instances". The cubic lattice has been deeply studied as pointed out in the introduction, but the FCC lattice has been shown to admit the tightest packing of spheres (Cipra 1998), indicating that it allows for more complex 3D structures.

Table 1 present the first results (to our knowledge) for the Harvard instances on the FCC lattice. These results are particularly interesting from a biology standpoint. They indicate that the energy of the best solution on the FCC lattice is always at least twice as low as the optimal energy for the cubic lattice, clearly showing the benefits of the FCC lattice for capturing richer 3D information. Since our algorithm is not guaranteed to find the optimal solution, the benefits may be even greater in practice, clearly suggesting that more investigation of the FCC lattice is necessary. Note that, in the FCC lattice, every point has twice as many neighbors as in the cubic lattice (12 instead of 6), thus dramatically increasing the combinatorics of the folding. From an efficiency standpoint, the best results are all obtained in less than 5 minutes and often much less, indicating the potential of the approach.

Seq.	Opt. E CL	k	Lowest E FCC	median time
1	-32	1	-66	2209.67
		2	**-68**	113.05
		3	-67	117.72
2	-34	1	-68	88.04
		2	**-69**	264.56
		3	**-69**	284.54
3	-34	1	-67	11.94
		2	-67	105.44
		3	**-68**	72.16
4	-33	1	**-66**	161.40
		2	-65	35.98
		3	**-66**	44.47
5	-32	1	**-66**	164.38
		2	**-66**	52.80
		3	**-66**	88.82
6	-32	1	-68	3.86
		2	-69	117.111
		3	**-70**	149.29
7	-32	1	**-68**	7.59
		2	**-68**	169.22
		3	-67	63.98
8	-31	1	-63	75.08
		2	**-64**	23.15
		3	**-64**	0.01
9	-34	1	-68	197.19
		2	**-69**	197.06
		3	**-69**	89.48
10	-33	1	**-66**	30.48
		2	**-66**	113.6
		2	**-66**	43.33

Table 1: Results for the Harvard sequences for each value of k. In bold lowest energy found for the FCC lattice. Optimal value for the Cubic lattice is also depicted. Median time to reach the best solution is in seconds.

Other Instances

We also compare our solutions with the only FCC foldings available in the literature. Table 2 shows a comparison for 4 instances found in (Will 2005, p. 130). Optimal results [1] are also shown. Figure 3 depicts a 3D view of the best configuration found for S4 for the various k values. As expected, the hydrophobic amino acids are clustered in the center of the protein. Although it is only an approximation of reality, it is still significant from the biological standpoint. We also present results for the R instances appearing in (Backofen & Will 2006). These instances are proteins of length 200 and are mentioned also in (Will 2005, p. 129) although no optimal configurations were given.

Note that the best results are achieved for parameter $k = 2$, while $k = 3$ yields a faster convergence to a lower quality solution. We interpret this to mean that $k = 3$ gives too high a weight to unit distances, while $k = 2$ represents a smoother weight that carries the search towards higher quality solutions.

Finally, figure 4 depicts the improvement of the solutions of our algorithm over time. The algorithm exhibits a steep descent, followed by a long plateau, and then another steep descent.

[1] Personal Communication with Sebastian Will.

Seq.	Native E	k	Lowest E	median time
S1	-357	1	-315	708.90
		2	**-325**	959.20
		3	-310	0.39
S2	-360	1	-312	548.38
		2	**-315**	1151
		3	-307	0.42
S3	-367	1	-299	704.58
		2	**-307**	68.58
		3	-299	1.8
S4	-370	1	-307	855.75
		2	**-318**	788.55
		3	-290	9.13

Table 2: Results for S sequences for each k. In bold lowest energy found. Time to best solution in seconds.

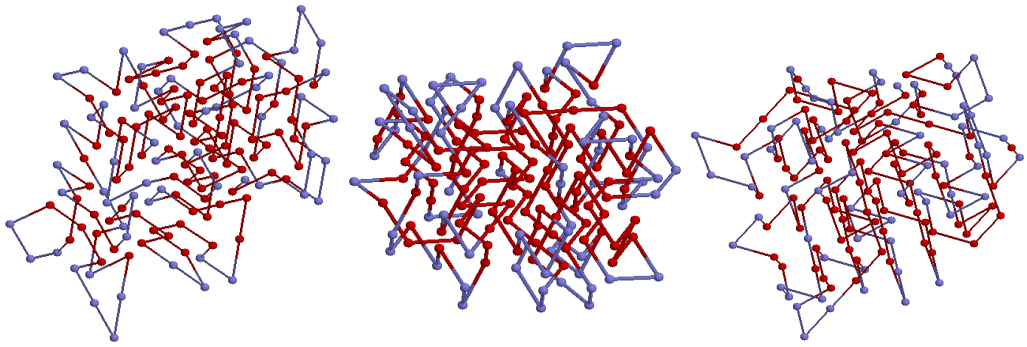
Seq.	Native E	k	Lowest E	median time
R1	-384	1	-261	1.3
		2	-270	2.28
		3	**-284**	125.65
R2	-383	1	-282	47.9
		2	-274	127.92
		3	**-290**	1128.59
R3	-385	1	**-282**	386.98
		2	-278	1.43
		3	-276	2.65

Table 3: Results for R sequences for each k. In bold lowest energy found. Time to best solution in seconds.

Conclusion

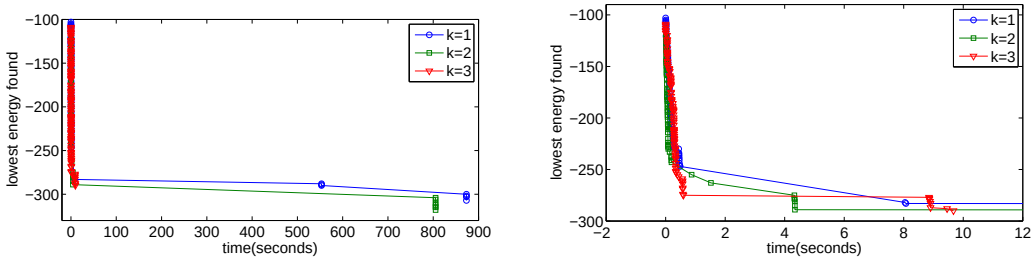
This paper presented a local search algorithm for finding the best self avoiding walk for the Hydrophobic-Polar (HP) energy model on the Face Centred Cubic (FCC) lattice. The algorithm relies on a randomized, structured initialization, a novel fitness function to guide the search, and efficient data structures to obtain self-avoiding walks. Experimental results on standard Harvard instances show the benefits of considering the FCC lattice from a biological standpoint and the efficiency of the approach. In particular, on the well-known Harvard instances, the foldings obtained by the algorithm on the FCC lattice have an energy at least twice as low as the optimal energy for the cubic lattice, clearly showing the benefits of capturing richer 3D information. To our knowledge, these are the first experimental results for the Harvard instances on the FCC lattice.

Our current work explores more complex energy models and off-lattice setups. Preliminary results show that changing the energy (i.e., adding weights to contacts) can be achieved with minimal modification and with similar performance. The algorithm can be adapted to RNA structure prediction, which we are currently exploring and validating from a biological standpoint.



(a) k=1, E=-307. (b) k=2, E=-318. (c) k=3, E=-290.

Figure 3: Lowest energy found for instance S4 (Will 2005, p. 130), with 164 amino acids.



(a) Behaviour over 15 minutes (b) Behaviour over 12 seconds

Figure 4: Algorithm Behavior over Time for instance S4 for each value of k.

Acknowledgements

We would like to thank the reviewers for their useful comments. Also, thanks to Sebastian Will for sharing his results with us. I. Dotú is supported by a "Fundacion Caja Madrid" grant, M. Cebrián is supported by grant TSI 2005-08255-C07-06 of the Spanish Ministry of Education and Science and P. Clote is partially funded by NSF DBI-0543506.

References

Abkevich, V. I.; Gutin, A. M.; and Shakhnovich, E. I. 1997. Computer simulations of prebiotic evolution. *Pac Symp Biocomput.* 0(O):O.

Anfinsen, C. B. 1972. `http://nobelprize.org/nobel_prizes/chemistry/laureates/1972/anfinsen-lecture.pdf`.

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

Backofen, R., and Will, S. 2002. Excluding symmetries in constraint-based search. *Constraints* 7(3):333–349.

Backofen, R., and Will, S. 2006. A constraint-based approach to fast and exact structure prediction in three-dimensional protein models. *Constraints* 11(1):5–30.

Backofen, R.; Will, S.; and Clote, P. 2000. Algorithmic approach to quantifying the hydrophobic force contribution in protein folding. *Pacific Symposium on Biocomputing* 5:92–103.

Berger, B., and Leighton, T. 1998. Protein folding in the hydrophobic-hydrophilic (hp) model is NP-complete. *Journal of Computational Biology* 5:27–40.

Jacek Blazewicz, Piotr Lukasiak, Maciej Milostan. 2005. Application of tabu search strategy for finding low energy structure of protein. *Arti?cial Intelligence in Medicine* 35:135?145.

Cipra, B. 1998. Packing challenge mastered at last. *Science* 281:1267.

Clote, P. 1999. Protein folding, the Levinthal paradox and rapidly mixing Markov chains. In *Automata, Languages and Programming, 26th International Colloquium, ICALP'99*, 240–249.

Covell, D., and Jernigan, R. 1990. Conformations of folded proteins in restricted spaces. *Biochemistry* 27:3287–3294.

Crawford, J.; Ginsberg, M.; Lucs, E.; and Roy, A. 1996. Symmetry-breaking predicates for search problems. *KR'96: Principles of Knowledge Representation and Reasoning*, 148–159.

Crescenzi, P.; Goldman, D.; Papadimitriou, C.; Piccolboni, A.; and Yannakakis, M. 1998. On the complexity of protein folding. *J. Comp. Biol.* 5(3):523–466.

Flamm, C.; Fontana, W.; Hofacker, I.; and Schuster, P. 2000. RNA folding at elementary step resolution. *RNA* 6:325–338.

Hart, W. E., and Istrail, S. C. 1996. Fast protein folding in the hydrophobic-hydrophilic model within three-eighths of optimal. *J. Comput. Biol.* 3(1):53–96.

Tianzi Jiang, Qinghua Cui, Guihua Shi, Songde Ma. 2003. Protein folding simulations of the hydrophobic ? hydrophilic model by combining tabu search with genetic algorithms. *Journal of Chemical Physics* 118(8).

Kou, S. C.; Oh, J.; and Wong, W. H. 2006. A study of density of states and ground states in hydrophobic-hydrophilic protein folding models by equi-energy sampling. *J. Chem. Phys.* 124(24):244903.

Lau, K., and Dill, K. A. 1989. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Journal of the American Chemical Society* 22:3986–3997.

Madras, N., and Slade, G. 1996. *The Self-Avoiding Walk*. Boston: Birkhäuser. Series: Probability and its Applications, 448 p., ISBN: 978-0-8176-3891-7.

Šali, A.; Shakhnovich, E.; and Karplus, M. 1994a. How does a protein fold? *Nature* 369:248–251.

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

Sinclair, A. 1993. *Algorithms for Random Generation and Counting: A Markov Chain Approach*. Birkhäuser.

Tapia, L.; Tang, X.; Thomas, S.; and Amato, N. M. 2007. Kinetics analysis methods for approximate folding landscapes. *Bioinformatics* 23(13):i539–i548.

Will, S. 2005. *Exact, Constraint-Based Structure Prediction in Simple Protein Models*. Ph.D. Dissertation, Friedrich-Schiller-Universität Jena.

Woese, C. R.; Durge, D. H.; Dugre, S. A.; Condo, M.; and Saxinger, W. C. 1966. On the fundamental nature and evolution of the genetic code. *Cold Spring Harbor Symposium on Quantitative Biology* 31:723–736.

Yue, K., and Dill, K. A. 1996. Folding proteins with a simple energy function and extensive conformational searching. *Protein. Sci.* 5(2):254–261.

Yue, K., and Dill, K. A. 2000. Constraint-based assembly of tertiary protein structures from secondary structure elements. *Protein. Sci.* 9(10):1935–1946.

Yue, K.; Fiebig, K.; Thomas, P.; Chan, H.; Shakhinovich, E.; and Dill, K. 1995. A test of lattice protein folding algorithms. In *National Academy of Science*, volume 92, 325–329.

Zhang, J.; Kou, S. C.; and Liu, J. S. 2007. Biopolymer structure simulation and optimization via fragment regrowth Monte Carlo. *J. Chem. Phys.* 126(22):225101.