# Random-Walk: A Stagnation Recovery Technique for Simplified Protein Structure Prediction

Mahmood A Rashid[1,2]*, Swakkhar Shatabda[1,2], M.A.Hakim Newton[1,2], Md Tamjidul Hoque[3], Duc Nghia Pham[1,2] and Abdul Sattar[1,2] †

[1]Queensland Research Laboratory, NICTA‡
[2] Institute for Integrated and Intelligent Systems, Griffith University
[3] Computer Science, University of New Orleans, USA

## ABSTRACT

Protein structure prediction is a challenging optimisation problem to the computer scientists. A large number of existing (meta-)heuristic search algorithms attempt to solve the problem by exploring possible structures and finding the one with minimum free energy. However, these algorithms often get stuck in local minima and thus perform poorly on large sized proteins. In this paper, we present a random-walk based stagnation recovery approach. We tested our approach on tabu-based local search as well as population based genetic algorithms. The experimental results show that, random-walk is very effective for escaping from local minima for protein structure prediction on face-centred-cubic lattice and hydrophobic-polar energy model.

## Categories and Subject Descriptors

I.2 [**Computing Methodologies**]: Artificial Intelligence; [**Problem Solving, Control Methods, and Search**]: *Heuristic methods*;
J.3 [**Computer Applications**]: Life and Medical Sciences; [**Biology and Genetics**]: *protein structure prediction*

## General Terms

Algorithms

## Keywords

Protein Structure Prediction, Search Stagnation, Local Minima, Random-walk

## 1. INTRODUCTION

Proteins are essentially sequences of amino acids. They adopt specific folded three-dimensional structures to perform specific tasks. The function of a given protein is determined by its *native* structure, which has the lowest possible free energy level. Nevertheless, misfolded proteins cause many critical diseases such as Alzheimer's disease, Parkinson's disease, and Cancer [1, 6]. Protein structures are important in drug design and biotechnology.

Protein structure prediction (PSP) is computationally a very hard problem [11]. Given a protein's amino acid sequence, the problem is to find a three dimensional structure of the protein such that the total interaction energy amongst the amino acids in the sequence is minimised.

A large number of existing (meta-)heuristic search algorithms attempt to solve the problem by exploring possible structures and finding the one with minimum free energy but they suffer from stagnancy for larger proteins. Stagnation is the situation when the search algorithms get stuck or stall in local minima: trap in a valley or loss way out in plateaus. Therefore, finding an effective stagnation recovery technique is essential for further progress in conformational search.

Search-based optimisation algorithms for PSP deal stagnation in different ways. In local search (LS), random restart is widely used in stagnation recovery. In population based algorithms, such as genetic algorithms (GA), the similarity within the population increases with progressive generations. In GA, typically, the individuals that are very similar are removed from the current population and the vacancies are filled up with randomly generated solutions in a stall situation.

In this paper, we present a random-walk (RW) based stagnation recovery technique. A random-walk is a process of exploring feasible solutions from a base solution with minimal changes to it. We tested our technique with tabu guided local search as well as population based genetic algorithms for PSP on FCC lattice and HP energy model. Finally, we show the effectiveness of random-walk in search optimisation for simplified PSP experimentally.

*corresponding author

†{mahmood.rashid, swakkhar.shatabda, hakim.newton, duc-nghia.pham, abdul.sattar}@nicta.com.au, thoque@cs.uno.edu

‡NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

ACM-BCB'12, October 7-10, 2012, Orlando, FL, USA
ACM 978-1-4503-1670-5/12/10.

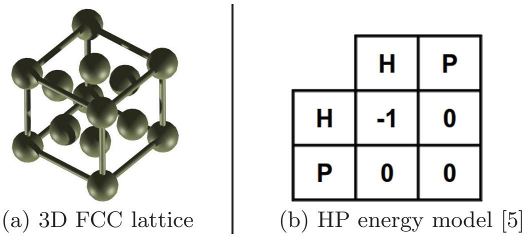
## 2. BACKGROUND

*Homology modeling*, *protein threading* and *ab initio* are three computational approaches used in protein structure prediction. Prediction quality of *homology modeling* and *protein threading* depends on the sequential similarity of previously known protein structures. However, our work is based on the *ab initio* approach that only depends on the amino acid sequence of the target protein. Levinthal's paradox [10] and Anfensen's hypothesis [2] are the basis of *ab initio* method for PSP. The idea was originated in 1970 when it was demonstrated that all information needed to fold a protein resides in its amino acid sequence. In our simplified protein structure prediction model, we use 3D FCC lattice for conformation mapping, HP energy model for conformation evaluation. We use LS and GA as search optimisation algorithms.

**LS:** Starting from an initial solution, local search algorithms move from one solution to another to find a better solution. Local search algorithms are well known for efficiently producing high quality solutions, which are difficult for systematic search approaches. However, they are incomplete [3], and suffer from revisitation and stagnation.

**GA:** A genetic algorithm maintains a set of solutions known as population. In each *generation*, it generates a new population from the current population using a given set of genetic operators known as *crossover* and *mutation*. It then replaces inferior solutions by superior newly generated solutions to get a better current population.

**3D FCC Lattice:** The FCC lattice has the highest packing density compared to the other existing lattices [7]. In FCC (Figure 1a), each lattice point has 12 neighbours with 12 *basis vectors*: $(1, 1, 0)$, $(-1, -1, 0)$, $(-1, 1, 0)$, $(1, 1, 0)$, $(0, 1, 1)$, $(0, 1, -1)$, $(1, 1, 0)$, $(1, 0, -1)$, $(0, -1, 1)$, $(-1, 0, 1)$, $(0, -1, -1)$, and $(-1, 0, -1)$.



(a) 3D FCC lattice

|   | H | P |
|---|---|---|
| H | -1 | 0 |
| P | 0 | 0 |

(b) HP energy model [5]

**Figure 1: a) 3D FCC lattice and b) HP energy model.**

**HP Energy Model:** In the HP model [5], when two non-consecutive hydrophobic amino acids become topologically neighbours, they release a certain amount of energy, which for simplicity is shown as $-1$ in Figure 1b. The total free energy ($E$) of a conformation based on the HP model becomes the sum of the contributions of all pairs of non-consecutive hydrophobic amino acids as shown in Equation 1.

$$E = \sum_{i<j-1} c_{ij}.e_{ij} \quad (1)$$

Here, $c_{ij} = 1$ if amino acids $i$ and $j$ are non-consecutive neighbours on the lattice, otherwise 0; and $e_{ij} = -1$ if $i^{\text{th}}$ and $j^{\text{th}}$ amino acids are hydrophobic, otherwise 0.

## 3. RELATED WORK

In LS, restart from current best solution by resetting all parameters and tabu-list was used in [4] to escape from local minima. However, in GA, a twin-removal operator [8] was used to remove duplicates from the population to recover the search from stalling.

## 4. OUR APPROACH

In search stagnation, we apply a random-walk algorithm to change the current solution. We refer random-walk as a process of exploring feasible solutions from a base solution with minimal changes to it. In random-walk, both bad and good solutions are traversed by making minimal changes to the current solution.

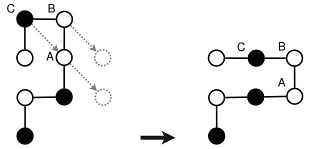


**Figure 2: The pull move operator used random-walk; for easy comprehension, presented in 2D space.**

Local minima in PSP are encountered when premature hydrophobic-core (H-core) is formed. To break the premature H-core we apply a random-walk algorithm (the *pseudocode* in Algorithm 1). This algorithm uses pull moves [9] (as shown in Figure 2). We use pull-moves because they are complete, local, and reversible. Successful pull moves never generate infeasible conformations. During pulling, energy level and structural diversification are observed to maintain balance among these two. We allow energy level to change within 5% to 10% that changes the structure from 10% to 75% of the original. We try to accept the conformation that is close to the current conformation in terms of energy level, and has possible maximum structural diversity from the current conformation.

**Algorithm 1:** randomWalk()

```
isFound ⟵ false
while (!isFound) do
    for (pos=1 to seqLength) do
        applyPullMove(pos)
    end
    isFound ⟵ checkDiversity()
end
```

We evaluate and compare current solution with the so-far global-best solution at the end of each iteration (generation in GA). For LS, when the number of non-improving iteration is reached to the *maxtry-threshold* that was set initially, the random-walk algorithm is applied. Instead of iteration, non-improving generation count is used in GA to apply the random-walk algorithm.

## 5. EXPERIMENTAL RESULTS

In our experiment, we use the protein sequences (as shown in Table 1 and 2) available from Peter Clote laboratory website[1]. Cebrian *et al.* [4] used these instances to test their algorithm. We present two sets of results: Table 1 presents the

[1] Peter Clote Lab: http://bioinformatics.bc.edu/clotelab/FCCproteinStructure/

effectiveness of random-walk in LS and Table 2 in GA. The column *LB-E* presents the lower bound of free energy and = in column *Best* for both the tables implies that the lower bound of the free energy is obtained (i.e., *Best*=*LB-E*). The results are calculated over 50 different runs (2 hours/run) for both LS and GA.

| Protein Info | | | **Local Search** | | | | |
|---|---|---|---|---|---|---|---|
| | | | Without RW (r) | | With RW (t) | | |
| **Seq** | **Size** | **LB-E** | **Best** | **Avg** | **Best** | **Avg** | **RI** |
| F1 | | -168 | -162 | -157 | = | **-167** | **91%** |
| F2 | | -168 | -159 | -157 | **-167** | **-164** | 65% |
| F3 | 91 | -167 | -161 | -156 | = | **-165** | 81% |
| F4 | | -168 | -162 | -158 | = | **-165** | 77% |
| F5 | | -167 | -161 | -158 | = | **-165** | 76% |
| S1 | 135 | -357 | -340 | -330 | **-355** | **-347** | 62% |
| S2 | 151 | -360 | -337 | -330 | **-354** | **-347** | **57%** |
| S3 | 162 | -367 | -346 | -328 | **-359** | **-350** | **57%** |
| S4 | 164 | -370 | -339 | -321 | **-358** | **-350** | 59% |

Table 1: Random-walk with local search.

| Protein Info | | | **Genetic Algorithms** | | | | |
|---|---|---|---|---|---|---|---|
| | | | Without RW (r) | | With RW (t) | | |
| **Seq** | **Size** | **LB-E** | **Best** | **Avg** | **Best** | **Avg** | **RI** |
| F1 | | -168 | -165 | -159 | = | **-166** | 78% |
| F2 | | -168 | -164 | -159 | = | **-165** | 67% |
| F3 | 91 | -167 | -163 | -158 | = | **-164** | 67% |
| F4 | | -168 | -165 | -160 | = | **-165** | 63% |
| F5 | | -167 | -166 | -160 | = | **-166** | **86%** |
| S1 | 135 | -357 | -344 | -336 | **-355** | **-348** | 57% |
| S2 | 151 | -360 | -346 | -335 | **-356** | **-349** | 56% |
| S3 | 162 | -367 | -347 | -334 | **-361** | **-349** | **45%** |
| S4 | 164 | -370 | -350 | -333 | **-364** | **-352** | 51% |

Table 2: Random-walk with genetic algorithms.

### 5.1 Analysis

In Tables 1 and 2, we present relative improvements (RI) on average conformation quality. We compare results between *target* (with random-walk) and *reference* (without random-walk) for both LS and GA. For each protein, the RI of the target ($t$) w.r.t. the reference ($r$) is calculated using the formula in Equation 2, where $E_t$ and $E_r$ denote the average energy values achieved by the target and the reference respectively, and $E_l$ is the lower bound of free energy for the protein in HP model.

$$RI = \frac{E_t - E_r}{E_l - E_r} * 100\% \qquad (2)$$

The results show that the minimum and maximum RI of using random-walk in LS are 57% and 91% respectively. These values are 45% and 86% for GA.

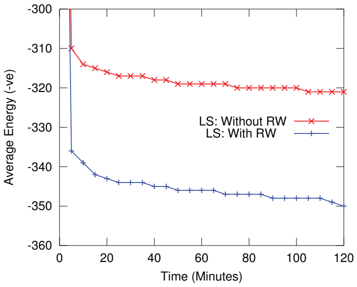


Figure 3: Search progress for protein sequence S4 w.r.t. time between two variants of LS.

### 5.2 Search Progress

We compare the average search progresses between two variants of LS (Figure 3), and two variants of GA (Figure 4) over time (2 hours) for protein sequence S4. We observe that all of the algorithms achieve very good progress initially, but with increasing time, algorithms using random-walk make more progress than others.

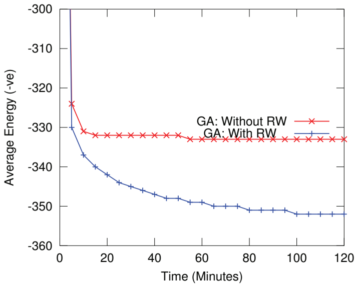


Figure 4: Search progress for protein sequence S4 w.r.t. time between two variants of GA.

## 6. CONCLUSION AND FUTURE WORK

In this paper, we present a random-walk-based stagnation recovery technique. We experimentally show that for both tabu-based single point local search and population based genetic algorithms, random-walk is very effective in escaping stagnation for conformation search. We aim to apply random-walk with different algorithms in high resolution PSP in future.

## 7. REFERENCES

[1] Adam Smith. Protein Misfolding. *Nature Reviews Drug Discovery*, 426(6968):78–102, December 2003.

[2] C. B. Anfinsen. The Principles that Govern the Folding of Protein Chains. *Science*, 181(4096):223–230, 1973.

[3] B. Berger and T. Leightont. Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete. *Journal of Computational Biology*, 5(1):27–40, 1998.

[4] M. Cebrián, I. Dotú, P. Van Hentenryck, and P. Clote. Protein structure prediction on the face centered cubic lattice by local search. In *National Conference on Artificial Intelligence - Volume 1*, pages 241–246, 2008.

[5] K. A. Dill. Theory for the folding and stability of globular proteins. *Biochemistry*, 24(6):1501–1509, 1985.

[6] C. M. Dobso. Protein folding and misfolding. *Nature*, 426(6968):884–890, 2003.

[7] T. Hales. A proof of the Kepler conjecture. *The Annals of Mathematics*, 162(3):1065–1185, 2005.

[8] M. T. Hoque, M. Chetty, A. Lewis, and A. Sattar. Twin Removal in Genetic Algorithms for Protein Structure Prediction using Low-Resolution Model. *Transactions on Computational Biology and Bioinformatics*, 8(1):234–245, 2011.

[9] N. Lesh, M. Mitzenmacher, and S. Whitesides. A complete and effective move set for simplified protein folding. In *Research in Comp. Mol. Biology (RECOMB)*, 2003.

[10] C. Levinthal. Are there pathways for protein folding? *Journal of Medical Physics*, 65(1):44–45, 1968.

[11] The Science Editorial. So Much More to Know. *The Science*, 309(5731):78–102, July 2005.