A Genetic Algorithm with Backtracking for Protein Structure Prediction

Clayton Matthew Johnson
Department of Math and Computer Science
California State University East Bay
Hayward, California 94542
matt.johnson@csueastbay.edu

Anitha Katikireddy
Department of Math and Computer Science
California State University East Bay
Hayward, California 94542
akatikireddy@horizon.csueastbay.edu

ABSTRACT

In this paper, we propose a simple genetic algorithm for finding the optimal conformation of a protein using the three-dimensional square HP model. A backtracking procedure is used to resolve the positional collisions and illegal conformations that occur during the course of genetic search. Backtracking is shown to be a simple and efficient means of collision repair that requires little overhead. Empirical results show that a genetic algorithm using backtracking can obtain the lowest energy structure of an amino acid sequence in fewer energy evaluations than earlier approaches.

Categories and Subject Descriptors

I.2.8 [**Artificial Intelligence**]: Problem Solving, Control Methods and Search – *Heuristic Methods*; J.3 [**Life and Medical Sciences**]: Biology and genetics.

General Terms

Algorithms, Design.

Keywords

Genetic Algorithms, HP Model, Protein Structure Prediction.

1. PROTEIN STRUCTURE PREDICTION

The *primary structure* of a protein is the amino acid sequence of its polypeptide chain, while the *secondary structure* is the local arrangement of a polypeptide's backbone atoms without regard to the conformations of its side chains. Under certain physiological conditions, the primary structure of a protein spontaneously folds into a precise three-dimensional form called its *tertiary structure* or *native state* that determines its functional properties.

The search for a set of rules that would derive a protein's tertiary structure from its primary structure is known as the *Protein Folding Problem*. Currently, the primary structures of approximately 40,000 proteins are known. Only a small percentage of these have known native states.

Thermodynamic measurements indicate that a protein in its native state has minimum energy. Efforts aimed at solving the Protein Folding Problem have involved the optimization of a potential energy function that approximates the thermodynamic state of a protein macromolecule. Since an algorithm using such a potential function does not give insight into how a protein folds, these approaches are instead known as *Protein Structure Prediction*.

2. THE HP MODEL

The *hydrophobic-hydrophilic model* (HP model) by Dill [1] is a simple abstraction that captures the essence of the important concepts of Protein Structure Prediction. In the HP model, amino acids are divided into two categories: *hydrophobic* (H) and *hydrophilic* (P). The *primary sequence* of a protein is therefore S $\in \{H, P\}^+$. Using this simplification, optimization models can be developed that seek to maximize interactions between adjacent pairs of hydrophobic amino acids (or *hydrophobes*). Adjacency is considered only in the cardinal directions of a lattice upon which the sequence is embedded.

In an HP lattice, vertices represent amino acids and edges represent connecting bonds. Black squares at the vertices indicate hydrophobes, while white squares indicate hydrophilic amino acids. A lattice can be two or three dimensional, and either square, cubic or triangular.

The hydrophobic-hydrophobic (HH) contacts are the basis for the evaluation function. Every pair of hydrophobes that are adjacent on the lattice and not consecutive in the primary sequence is awarded a value ε (usually –1). The sum of all such values gives the energy of the conformation. Figure 1 shows a 13-length sequence embedded on a square lattice, with HH contacts indicated by gray double arrows.

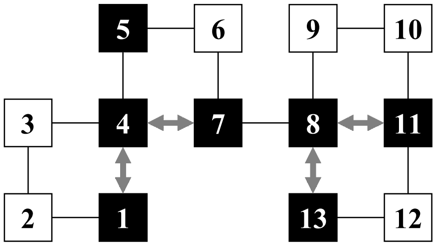

Figure 1. HP sequence on a square lattice with energy -4.

3. EXPERIMENT

This investigation considers the efficacy of a GA utilizing a simple backtracking procedure to avoid illegal conformations.

GECCO'06, July 8–12, 2006, Seattle, Washington, USA.
ACM 1-59593-186-4/06/0007.

The GA is initiated with a population of randomly generated structures containing the three-dimensional internal coordinates {F, L, R, U, D} (forward, left, right, up, and down) of each residue in a protein. Each chromosome therefore contains the sequence (F, L, R, U, D)$^{n-2}$ for a peptide sequence of length *n*, as there is no directional offset at either end of the chain. This structure is plotted on a three-dimensional lattice in order to evaluate the chromosome's fitness.

A *collision* is the embedding of two different peptides onto the same vertex of the lattice. As each member of the initial population is randomly generated, it may represent an illegal conformation resulting in one or more collisions when embedded. Similarly, the application of the mutation and crossover operators to legal conformations may produce additional collisions. The GA implements a relatively straightforward backtracking method to resolve these situations.

When a collision occurs, the backtracking routine repairs the chromosome by considering alternative positional offsets. As there are four possible new directions in which the fold might proceed after collision, each of these directions is tested in random order until an empty lattice point is discovered. If no such vertex exists, the algorithm backtracks to the previous peptide in the sequence and investigates the alternative directional choices from this position. This process is iteratively repeated until a proper placement can be achieved, at which point embedding of the remaining amino acid residues in the sequence proceeds.

The GA implements roulette wheel selection, single-point crossover at a rate of 0.95, and bitwise mutation at a rate of either 0.001 or 0.0001. Elitism is used to replicate 2% of each generation directly into the next, with a population size set between 1000 and 1600. Each sequence is run for 300 generations.

4. RESULTS

For comparison purposes, we have run our algorithm on the same 27-length sequences used by Unger and Moult [3] and Patton *et al.* [2].

Table 1 gives a comparison of the results. The numbers reported for our approach are the best results achieved from five separate trials. All runs of the algorithm, however, were able to discover the lowest energy conformation of each sequence in fewer energy evaluations.

The "backtrack count" indicated in the final column of Table 1 is a tally of each directional offset examined during all backtracks for the sequence. It should be noted that each count is a Θ(1) operation, as it only requires the attempted placement of a single residue. By comparison, each energy evaluation is linearly bounded, as the entire sequence must be embedded onto the lattice and checked for adjacent H-H pairs.

5. CONCLUSIONS

When comparing the difference in performance between our GA and earlier approaches, see a significant drop in the number of energy evaluations needed. Thus, a genetic algorithm using backtracking to resolve collisions outperforms earlier approaches for the 27-length test sequences.

The backtracking heuristic itself is well-behaved with little overhead. On average, the method required only 0.438 backtrack checks per energy evaluation. Preliminary investigation on significantly longer sequences indicates that backtracking scales at roughly the same ratio.

6. REFERENCES

[1] Dill, K. A. Theory for the folding and stability of globular proteins. *Biochemistry*, 24, 6 (March 12, 1985), 1501-1509.

[2] Patton, A. L., Punch, W. F., and Goodman, E. D. A standard GA approach to native protein conformation prediction. In *Proceedings of the sixth international conference on genetic algorithms (ICGA '95)* (Pittsburgh, PA, July 15-19, 1995). Morgan Kaufmann, San Francisco, CA, 1995, 574-581.

[3] Unger, R., and Moult, J. A genetic algorithm for three dimensional protein folding simulations. In *Proceedings of the fifth international conference on genetic algorithms (ICGA '93)* (Urbana-Champaign, IL, 17-21 July, 1993). Morgan Kaufmann, San Francisco, CA, 1993, 581-588.

Table 1: Comparison of Results to Unger and Moult and Patton *et al.*

Sequence	Unger & Moult		Patton *et al.*		GA with Backtracking		
	Lowest Energy	Energy Evaluations	Lowest Energy	Energy Evaluations	Lowest Energy	Energy Evaluations	Backtrack Count
273d.1	-9	1,227, 964	-9	27,786	-9	15,854	6,578
273d.2	-9	1,225, 281	-10	81,900	-10	19,965	7,794
273d.3	-8	1,247, 208	-8	16,757	-8	7,991	3,547
273d.4	-15	1,207, 686	-15	85,447	-15	23,525	11,736
273d.5	-8	1,118, 202	-8	8,524	-8	3,561	1,422
273d.6	-11	1,226, 090	-11	44,053	-11	14,733	5,885
273d.7	-12	1,239, 519	-13	85,424	-13	23,112	10,538
273d.8	-4	1,248, 118	-4	3,603	-4	889	352
273d.9	-7	1,198, 945	-7	10,610	-7	5,418	2,424
273d.10	-11	1,174, 297	-11	16,282	-11	5,592	2,613
Total		**11,113,310**		**380,386**		**120,640**	**52,889**