An Approach to Amino Acid Sequence Alignment Using a Genetic Algorithm

Masato Wayama
wayama@cc.hirosaki-u.ac.jp

Katsutoshi Takahashi
sltaka@si.hirosaki-u.ac.jp

Toshio Shimizu
slsimi@si.hirosaki-u.ac.jp

Department of Information Science, Faculty of Science,
Hirosaki University,
Bunkyo-cho 3, Hirosaki 036, JAPAN.

1 Introduction

Every year an increasing number of amino acid sequences of proteins are being solved by genome sequencing techniques. As a result, it is becoming increasingly important to rapidly compare a newly determined amino acid sequence with a huge number of other sequences. The dynamic programming algorithm(Needlman & Wunsch [8]) is available for comparing two sequences automatically. In this algorithm, sequences are aligned by inserting gaps to maximize a similarity score which is calculated by means of a similarity matrix such as the Dayhoff matrix (Dayhoff et al.[2]). Although pairwise alignment has been widely used in sequence comparison, distantly related proteins often display ambiguous relationships that cannot be detected by automatic pairwise alignment algorithms. Such relationships can be detected by using simultaneous comparison of more than two sequences.

A number of automated algorithms to align multiple sequences have been proposed so far. These algorithms might be classified into two categories: tree-based algorithms with the use of dynamic programing including iterative progress methods (for example, Ishikawa et al.[4]) and simulated annealing methods (for example, Kim et al.[6]). Simulated annealing is a good heuristic method to solve combinatorial optimization problems (Kirkpatrick et al.[7]). Recently, a genetic algorithm has been used in computational molecular biology as a powerful combinatorial optimizer. Ishikawa et al. ([5]) applied a genetic algorithm to the multiple sequence alignment problem in combination with dynamic programing.

A simple genetic algorithm (Goldberg [3]) employs the "chromosomal" representation of possible solutions to a given problem - a population of random strings of 1's and 0's. Genetic operations, such as selection, crossover and mutation, are applied to the population of bit strings to create a new population of bit strings.

In this paper, we propose an approach to sequence alignment problems employing the simple genetic algorithm. We will demonstrate the applicability of the implementation to the pairwise sequence alignment. Our implementation is quite simple and is naturally extended to multiple alignment problems.

2 Method

We employed Goldberg's ([3]) simple genetic algorithm in our implementation. A population of possible alignments is described with a set of bit matrices of which the elements are strictly 1 or 0. A sequence, including gaps, in an alignment is represented as a bit string. In this bit string, '1' indicates the position of a gap, with the total number of '0's being exactly the length of the sequence. Any alignment is described by a matrix, which is a vertical arrangement of the bit strings.

和山正人、高橋勝利、清水俊夫：弘前大学理学部情報科学科、〒０３６　弘前市文京町３

A random population of bit matrices is initialized at the beginning of the genetic algorithm run. A next population is obtained by applying the three genetic operations: selection, crossover and mutation. The selection operator is applied in order to choose matrices for the next generation from the matrices in the current population, where the selection probability of each individual is proportional to the similarity score. Next, the crossover operator is used to obtain a partial exchange of the information between two parent matrices. Then, the mutation operator is applied to each bit matrix in the new population.

The crossover operator is a key operator in the genetic algorithm. We applied the standard one-point crossover operator (Goldberg [3]), except for gap treatment in each bit string in the matrix. This is done by generating a random crossover point in the range of values from 1 to the shortest length of the sequences, instead of the bit string length. The crossover point in each sequence is counted from the beginning of each component bit string excluding any gaps.

By means of genetic operators, a new population of bit matrices is generated from the previous population. This process is repeated many times to obtain optimized alignment.

3 Evaluations

At first, we generated artifical sequences randomly. The amino acid composition of these random sequences was adapted to the composition for the complete databank of SWISS-PROT (Bairoch & Boeckmann [1]) release 30.

In order to evaluate the efficiency of the genetic algorithm and to obtain optimal genetic algorithm parameters, we aligned a random sequence against itself, with the set of parameters being varied systematically. When a random sequence, which consists of 50 residues, was self-aligned, we succeeded to find optimal values of the genetic algorithm parameters, such as population size, maximum iteration number and mutation rate.

Our current research directions are, (i) to optimize genetic algorithm parameters for longer sequence alignment, (ii) to speed up genetic algorithm runs and (iii) to extend our implementation toward multiple sequence alignment.

Acknowledgement

All computations have been performed at the Hirosaki University Center for Computer and Communications.

References

[1] Bairoch,A. and Boeckmann,B.; Nucleic Acids Res. (1992), **20**.

[2] Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C.; Atlas of Protein Sequence and Structure (1978), **5**, Nat. Biomed. Res. Found., Wachington DC.

[3] Goldberg,D.E.; Genetic Algorithms in Search, Optimization and Machine Learning (1989), Addison Wesley, Reading, MA.

[4] Ishikawa,M., Hoshida,M., Hirosawa,M., Toya,T., Onizuka,K. and Nitta,K.; Proc. Fifth Generation Comput. Syst. '92 (1992), 294-299.

[5] Ishikawa,M., Toya,T., Totoki,Y. and Konagaya,A.; Proc. AI and Genome Workshop in 13th Int'l. Joint Conf. Artifi. Intelli. (1993), 13-22.

[6] Kim,J., Pramanik,S. and Chung,M.J.; CABIOS (1994), **10**, 419-426.

[7] Kirkpatrick,S., Gelatt,C.D. and Vecchi,M.P.; Science (1983), **220**, 671-680.

[8] Needlman,S.B. and Wunsch,C.D.; J. Mol. Biol. (1970), **48**,443-453.