Design of a Hardware Board for Sequence Alignment

T. Kato [1]
kato@genta.c.u-tokyo.ac.jp
M. Taiji [2]
taiji@kyohou.c.u-tokyo.ac.jp
T. Ebisuzaki [2]
ebisu@chianti.c.u-tokyo.ac.jp
A. Suyama [1]
suyama@dna.c.u-tokyo.ac.jp

[1] Department of Life Sciences
and
[2] Department of Earth Science and Astronomy,
Graduate School of Arts and Sciences, The University of Tokyo
3-8-1 Komaba, Meguro-ku, Tokyo 153, Japan

Abstract

We have designed a special hardware board to calculate optimal alignments of two sequences based on the Myers-Miller dynamic programming algorithm. The board was designed to be able to calculate each similarity or distance matrix element in parallel in one system clock pulse. The present version of the board had four pipelines and thus can calculate 120 million matrix elements per one second.

Homology search and multiple-sequence alignment are frequently used analyses in molecular biology. The most reliable methods for homology search and sequence alignment are based on the dynamic programming algorithm [1]. The methods, however, take a long computation time for long sequences, which are now available by rapid progress in the genome projects. We have thus developed a hardware board to accelerate sequence alignment calculation on workstations based on the dynamic programming algorithms.

Execution of operations on a digital circuit board specially designed for them is much faster than that by a software program running on a standard computer. For performing complicated operations, however, constructing special digital circuit boards is more difficult and expensive than writing software programs. Only simple and highly repeated operations should therefore have much benefit of special digital circuit boards.

Sequence alignment based on dynamic programming algorithms is divided into two processes. The first one is similarity or distance matrix calculation, consisting of simple operations repeated many times. It requires $O(N^2)$ operations, where N is the length of sequences aligned, and occupies most of the total computation time. In contrast, the second process, reconstructing optimal alignments by tracing the matrix, consists of complicated operations mainly due to bifurcations on alignment paths. It requires $O(N)$ operations, thus occupying a small portion

[1]加藤 剛、陶山 明：東京大学大学院生命環境科学系物理学、〒153 東京都目黒区駒場 3-8-1
[2]泰地 真弘人、戎崎 俊一：東京大学大学院広域システム科学系宇宙地球科学、〒153 東京都目黒区駒場 3-8-1

of the total computation time. Therefore, only the first process is suitable for execution on special digital circuit boards.

The sequence alignment computer we have designed thus consists of two parts, a special digital circuit board and a host workstation, connected by the VMEbus. The board takes care of the matrix calculation, and the host, the reconstruction of optimal alignments.

The Smith-Waterman algorithm [2], which is the most popular sequence alignment algorithm based on dynamic programming method, was concluded to be unsuitable for a sequence alignment computer. Since the method uses the full matrix for reconstructing optimal alignments, a digital circuit board of the alignment computer must transfer the whole $O(N^2)$ matrix elements back to the host for the reconstruction. The actual data transfer rate of buses currently used on workstation falls around one MB/s. For long sequences, therefore, the bus transfer rate limits the speed of calculation.

The Myers-Miller algorithm [3], in contrast, fits well for a sequence alignment computer. The algorithm allows not only calculation of optimal similarity scores between two sequences in $O(N)$ space but also reconstruction of optimal alignments in $O(N)$ space. Therefore, $O(N)$ data transfer from a board to a host suffices for the reconstruction of alignments, giving a solution to the bus speed limitation problem. We have thus adopted the Myers-Miller algorithm for our sequence alignment computer.

A digital circuit board of the sequence alignment computer was composed of RAMs to store sequences (up to one kb for the present version) and boundary conditions, matrix calculation pipelines (four pipelines for the present version), a pipeline controller, and a similarity path midpoint calculator. Each pipeline has a serial arrangement of a similarity score RAM, a matrix element calculator, and a RAM storing terminal points of similarity paths. The score and the terminal point storage RAM were included in each pipeline to calculate every matrix element in parallel. The pipelines were designed to proceed with calculation in both a forward and a backward direction to reconstruct alignment paths based on the Myers-Miller algorithm. Digital circuits of the pipelines were made to be able to calculate each matrix element in one system clock pulse. The circuits implemented in logic cell arrays can operate at up to 30 MHz system clock frequency. Each pipeline can thus calculate 30 million matrix elements per one second.

Acknowledgments

This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas (Genome Informatics) from The Ministry of Education, Science, and Culture Japan.

References

(1) W. R. Pearson and W. Miller, *Methods Enzymol.*, **210**, 575, 1992.

(2) T. F. Smith and M. S. Waterman, *J. Mol. Biol.*, **147**, 195, 1981.

(3) E. W. Myers and W. Miller, *Comput. Appl. Biosci.*, **4**, 11, 1988.