Abstract

Aligning DNA sequences is a fundamental problem in bioinformatics. The exponential growth of protein and DNA databases makes this problem pose a great amount of challenge. Exact methods, which produce optimum sequence alignment according to a scoring function, have quadratic time and space complexity. Therefore, most of the current solutions are based on
heuristic methods, which do not guarantee an optimum solution. Recently, many parallel solutions were proposed in order to accelerate the exact methods. However, most of these solutions restrict the sequence’s sizes to be in kilobytes, in such a way that megabyte-scale genome comparison cannot be achieved. In addition, these solutions calculate only the alignment similarity score without finding the actual alignment. This paper presents an efficient solution to find the optimal alignment of the huge DNA sequences. This solution releases the condition of the sequence size to be in megabyte-scale instead of few kilobytes. The fundamental innovation in this work is developing efficient, linear space complexity, parallel solution to achieve the optimum alignment with relatively good performance. The shared memory parallel architecture is the focus of this work and therefore we have considered off-the-shelf systems like multi-core CPUs as well as advanced shared memory platforms. Experimental results show that, the proposed solution achieved high records compared to other solutions that targeted the same goal with less hardware requirements.

Reference

- Rognes Torbjørn, "Faster Smith-Waterman database searches with inter-sequence SIMD
- Xiandong Meng, "A High-Performance Heterogeneous Computing Platform for Biological
- Talal Bonny, Khaled N. Salama M. Affan Zidan, "High performance technique for
database applications using a hybrid GPU/CPU platform," in Proceedings of the 21st edition of
the great lakes symposium, New York, 2011, pp. 879 - 899.
- Sean R. Eddy, Anders Krogh, , Graeme Mitchison Richard Durbin, Biological sequence
- Dan Gusfield, Algorithms on Strings, Trees, and Sequences: Computer Science and
- Gregory F. Pfister, In search of clusters: the coming battle in lowly parallel computing.
- Christian Ledergerber, Philipp Krähenbühl, Christophe Dessimoz Adam Szalkowski,
"SWPS3 – fast multi-threaded vectorized Smith-Waterman for IBM Cell/B.E. and ×86/SSE2,"
- A Jankowski, A Modzelewski, A Piotrowski, A Zadrożny W Rudnicki, "The new SIMD
Implementation of the Smith-Waterman Algorithm on Cell Microprocessor," Fundamenta

Index Terms

Computer Science
Bioinformatics

Key words
sequence alignment megabase DNA sequence
space-efficient
parallel computing
multi-core