Efficient Computation of Sequence Analysis in a Vector-Parallel Computer for the Study of Molecular Evolution

Takamasa Futatsuki [1] tfutatsu@genes.nig.ac.jp
Yuichi Kawanishi [1] ykawanis@genes.nig.ac.jp
Kimitoshi Naito [1] SGE00155@niftyserve.or.jp
Satoru Miyazaki [2] smiyazak@genes.nig.ac.jp
Hideaki Sugawara [2] hsugawar@genes.nig.ac.jp

[1] Computer Chemistry Systems Department
Science Systems Division, FUJITSU LIMITED
9-3, Nakase 1-Chome, Mihama-ku, Chiba City, Chiba 261, Japan

[2] DNA Data Bank of Japan, National Institute of Genetics
1111 Yata, Mishima, Shizuoka 411, Japan

Abstract

We analyzed and implemented Smith and Waterman algorithm and maximum likelihood method into the vector-parallel computer of Fujitsu VPP500. The programs optimized for the computer are ssearch, clustalw and fastDNAml. Our goal is to develop a total system which will cover all processes from database search to the construction of large scale phylogenetic trees on super-computer.

1 Introduction

The phylogenetic analysis is one of the fundamentals for the genome analysis. For the study, we require more and more efficient algorithm and programs to process the explosively increasing sequence data. Therefore, we optimized Smith and Waterman algorithm which are widely used for the homology search and multiple alignment for the vector-parallel supercomputer and could greatly shorten the computation time.

2 Material and methods

The vector-parallel supercomputer used is the Fujitsu VPP500 which has 40 processors with 40G byte main memory in total. The maximum computation speed is 64 Giga flops.

For the optimization of Smith and Waterman algorithm, we vectorized the calculation of the homology score between two sequences. Let us consider find out the best homology of sequence A of length n and sequence B of length m. Then dynamic programming algorithm will be implemented in following way;
1) take $n \times m$ matrix $H(H_{ij})$
2) calculate each element H_{ij} of H which represents to the best homology score between partial sequence A of length i and partial sequence B of length j successively
To calculate H_{ij}, existing program usually take double loop structure (we call these two loop external loop and internal loop, respectively). We vectorized the calculation of internal loop based on the idea of E. Lander et al. [4].

We also optimized a phylogenetic analysis program named fastDNAml [3] developed for the vectorization and parallelization of maximum likelihood methods.

3 Summary and conclusions

Tables 1 and 2 show the calculation time of VPP500 versions of ssearch and clustlaw comparing with original scalar versions. The result on fastDNAml will be introduced at the poster session.

	vector	scalar
real	2028.77	16388.45
user	2011.00	16315.27
sys	8.45	20.10
vu-user	1901.81	–

Table 1: Computation time of ssearch program on VPP500

	vector	scalar
real	1495.17	6631.75
user	1488.34	6606.80
sys	0.93	2.39
vu-user	539.00	–

Table 2: Computation time of clustalw program on VPP500

Based on the fast computation system, we will be able to analyze the phylogenetic relationships of a very large scale taxonomic units more than 1000 OTUs.

References

[1] T.F. Smith and M.S. Waterman, "Identification of Common Molecular Subsequences", *J. Mol. Biol.*, 147:195–197,1981.

[2] D. J. Lipman and W. R. Pearson, "Rapid and sensitive protein similarity searches", *Science* 227:1435–1441, 1985.

[3] G. J. Olsen, H. Matsuda, R. Hagstrom and R. Overbeek, "fastDNAml : a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood",*CABIOS*, 10:41–48, 1994.

[4] E.Lander and J.P. Mesirov, "Study of Protein Sequence Comparison Metrics on the Connection Machine CM-2", *Proceedings of the Supercomputing 88*, Vol. II, 1988.