Parallelization of BLAST on Cluster Systems and Its Performance Study

Hong-Soog Kim[1*], Hae-Jin Kim[2], DongSoo Han[3]

[1,2,3,] School of Engineering, Information and Communications University, Daejeon, Korea
[*]To whom correspondence should be addressed. E-mail: kimkk@icu.ac.kr

Abstract

Basic Local Alignment Search Tool (BLAST) is one of the most widely used similarity search tools available to computational biologist. It has been used to find biologically similar sequences to the given query sequence from the database of the annotated sequences. It rapidly identifies statistically significant matches by comparing newly sequenced segments of genetic material or proteins with already annotated nucleotide or amino acid sequences in the database. This kind of search allows biologists to make inferences on the structure and function of the unknown gene or to screen new sequences for further investigation using more sensitive and computationally expensive methods. The information that BLAST provides in a few hours otherwise would take months of laboratory work.

For high throughput processing of huge number of query sequences, there have been many studies on parallel batch processing of sequence similarity search using BLAST. Although NCBI (National Center for Biotechnology Information) has developed a parallel BLAST using the thread on SMP (Symmetric MultiProcessors) machines for the speedup of BLAST, the speedup is still limited because the SMP machine has restricted the number of processors due to its architecture. Hence, the speedup improvement of BLAST on the SMP machine is not sufficient to cope with the current situation where enormous sequences are newly added in the database at exponential rate. In order to use more processors for more speedup of BLAST, we consider PC cluster system as an alternative to SMP machine.

In this paper, we present a parallel BLAST (Hyper-BLAST) on cluster systems for further speedup. In Hyper-BLAST, we extend application of the intra-search parallelism, which is exploited in NCBI BLAST, from one SMP node to multiple nodes in cluster system by logically partitioning of the sequence database.

The extension of intra-search parallelism enables us to use more processors for similarity search using BLAST and gives more speedup of individual query sequence search. The performance evaluation result shows that Hyper-BLAST gives scalable speedup in terms of response time as more processors are used. Compared to the cost of SMP machines, cluster system with same computation capability can be built with fewer budgets and Hyper-BLAST on cluster system provides fast similarity search with moderate cost. From the comprehensive performance evaluation with different configurations of cluster systems, we observed scalable speedup on various configurations of cluster systems.