Title of the Project : **AN EFFICIENT SCHEDULING STRATEGY FOR PROTEIN SEQUENCE ANALYSIS ON THE GRID**

Funding Agency : University Grants Commission (UGC) – MRP, New Delhi

Approval No. : F.No. 36-354/2008 (SR), Dated 28.03.2009

Period of Execution: 2009-2011

Principal Investigator: Dr.D.Ramyachitra
Assistant Professor
Department of Computer Science
Bharathiar University
Coimbatore - 641 046

EXECUTIVE SUMMARY

The main purpose of this project is to carry out sequence analysis in the protein database. To find out sequence similarities, search has to be carried out in the primary database. If the primary database cannot answer the questions, then secondary database can be searched looking for matches to the patterns they contain. As the secondary databases are derived from multiple alignments, the searches on them finds better relationships compared to search in the primary databases. Basic Local Alignment Search (BLAST) tool is used to compare a novel sequence with those contained in nucleotide and protein databases by aligning the novel sequence with previously characterized genes. Through this search, we will be able to find, regions of sequence similarity, its evolutionary history and homology with other sequences in the databases. The BLAST programs first looks for similar segments between the query sequence and a database

sequence, and evaluates the statistical significance of any matches that were found, and finally reports only those matches that satisfy a user selectable threshold of significance.

Bacterial Foraging Optimization algorithm is proposed in this research work to schedule the sequences to appropriate databases. The performance of the proposed method is evaluated in an intra grid environment connected by 10 systems using globus toolkit. In the 10 systems, FASTA database was stored upto the maximum of 5 machines and scheduler ran in one machine. Maximum of 100 queries were submitted by the users from 10 machines and the queries were sent to the scheduler. The scheduler in turn redirects the query to the appropriate machines based upon the scheduling algorithm. The experiment was conducted for various combinations of user queries and the resources. Resources are added one by one to find out the performance of the algorithms. For 2 resources, the existing algorithm does not show a decrease in makespan, whereas BFO gives an improvement in performance. The grid environment was scaled upto 5 systems for executing the queries and the proposed algorithm shows a decrease in the makespan compared to the existing algorithm. When, the proposed algorithm alone is considered, there is an improvement in performance as the number of resource increases. The reason for the existing scheduling algorithm FCFS to take more time for executing the user queries are the resources are not executed in a balanced way. In the case of BFO, number of jobs executed by the resources is almost balanced. Sometimes, one machine may be executing more number of jobs and the reason for that is the speed of the processor. It is also inferred that the resources are utilized in a balanced way and there is no need for rescheduling the jobs to some other machine which is a very tedious job. Different users from different machines submitted totally 100 tasks to the scheduler. It is seen that number of tasks executed decreases as the system is scaled one by one

and this may be due to network failure. But the numbers of jobs that are not executed are very less which can be executed on resubmission.

Thus the proposed method reduces the search time compared to existing algorithms. Also, the number of user requests executed by the proposed method is more compared to the existing methods. Thus protein sequence analysis can be performed on a grid environment with the scheduler being designed using Bacteria Foraging Optimization algorithm so that all the resources in the grid can be utilized to the maximum.

Publications:

- D.Ramyachitra, "Scheduling approaches for protein sequence analysis on the grid", Indian Streams research Journal, vol. 1, issue 4, pages 224-226, July 2011, ISSN: 2230-7850.
- K.Vivekanandan, D.Ramyachitra, "Bacteria Foraging Optimization for protein sequence analysis on the grid", Future Generation Computer Systems, **Vol.** 28, issue 4, pages 647-656, April 2012, ISSN: 0167- 739X, doi: 10.1016/j.future.2011.10.009, (SCI/ Impact Factor: 2.365).