ESTCLASS: a pipeline for EST data analysis using parallel computing

D'Agostino N. , Aversano M. , Chiusano M.L.

Department of Structural and Functional Biology, University 'Federico II', Naples

Motivation

The vast amount of expression sequence tags (EST) data in the public databases provides an important resource for comparative and functional genomic studies and, moreover, represents a useful information for a reliable annotation of genomic sequences. ESTs are short and error-prone sequences often including artifactual contamination, nevertheless they represent the framework from which fundamental information for both bioinformatic and experimental analysis is derived. Therefore, EST data need suitable preprocessing to provide high quality information. Because of the advances in biotechnologies, large EST datasets are daily released to the scientific community and fast and efficient bioinformatic approaches are the first step to support these data management and analysis. Therefore the requirements for this kind of analysis are suitable computational tools based on advanced technologies. We describe here a pipeline analysis for ESTs clustering, assembling and annotation by parallel computing that optimizes execution time for the processing of large data sets.

Methods

The pipeline has been developed on a 'Beowulf class' cluster, using Linux (Red Hat Fedora Core 2) as default operating system and the PBS batch system(TORQUE release from the OSCAR 4.0 distribution) for the management and the distribution of the tasks across the cluster nodes. The hardware is made up of 8 computational nodes, single processor AMD ATHLON 64 3.0 Ghz , connected through a dedicated Ethernet network of the gigabit class and controlled by a bi-processor (INTEL XEON 3.0 Ghz) master machine. Clusterwide are available 20 GB of RAM and 700 GB of total hard disk space. The main process of the pipeline is implemented in Perl. Its tasks are to serialize and control the parallel execution of the work-flow and to parse and integrate the results obtained from the different steps of the pipeline.

Results

Input datasets are pre-processed in two steps to remove contaminating sequences in order to avoid misclustering and/or misassembling. The first step requires RepeatMasker [1] and the NCBI's VECTOR [2] database for checking vector contaminations. In the second step, RepeatMasker and RepBase [3] are used for filtering and masking low complexity sequences and interspersed repeats. To accomplish sequence pre-processing, a specific utility has been designed to distribute the tasks across the nodes. Job assignments is managed by PBS [4]. Job control at each step and output files integration is managed by the main process. PaCE [5] is the software we selected for the clustering step. It requires the MPI implementation and the PBS job assignment for parallel execution. Once the whole pre-processed sequences (cleaned from vector and masked) are clustered, they are assembled in contigs (consensus sequences) using CAP3 [6]. To exploit the efficiency of CAP3 and avoid the overhead time consuming of PBS, the main process we implemented has been designed to bundle groups of commands to be executed sequentially by each processor. MPI-Blast [7] versus non redundant protein databases for

functional annotation is integrated in the pipeline. The presented pipeline has been developed to perform an exhaustive analysis on ESTs data. Moreover, it is designed to reduce execution time of the specific steps required for a complete analysis by means of distributed processing and parallelized software. It is conceived to run on low requiring hardware components, to fulfill increasing demand, typical of the data used, and scalability at affordable costs.

Contact email: chiusano@unina.it

Supplementary Information:

1.http://www.repeatmasker.org/ 2.ftp://ftp.ncbi.nih.gov/blast/db/FASTA/vector.gz 3.Jurka J. Repbase Update: a database and an electronic journal of repetitive elements. Trends Genet. 9:418-420 (2000) 4.http://www.openpbs.org/ 5.Kalyanaraman, A., Aluru, S., Kothari, S. & Brendel, V. (2003) Efficient clustering of large EST data sets on parallel computers. Nucl. Acids Res. 31, 2963-2974. 6.Huang X., Madan A. CAP3: A DNA sequence assembly program. Genome Res. 1999 Sep;9(9):868-77. 7.http://mpiblast.lanl.gov/