An optimized web server for metagenomics data analysis

Paoletti D[1], D'Antonio M[1], D'Onorio De Meo P[1], Chillemi G[1], Desideri A[2], Castrignanò T[1] , Pesole G[3,4]

Motivation

The advent of next-generation sequencing (NGS) platforms has given an amazing burst to Metagenomics, a new rampant discipline addressing the analysis of the genetic complexity of environmental samples, allowing for the first time the identification and functional characterization of the huge amount of so far unknown microorganisms which cannot cultured in the lab. Indeed, metagenomic analyses make now possible the full exploitation of the products of the evolution of life in different environments and conditions with unprecedented impacts in several biotechnological and medical areas. However, the large size of NGS data and the complexity of their analyses involve computational workloads requiring high-performance computing systems.

Methods

A high-throughput pipeline has been developed to provide high-performance computing to automate the taxonomic and functional assignments of short reads obtained by the pyrosequencing technology through extensive similarity searches against both protein and nucleotide databases. The pipeline is implemented in PHP and involves three main open source components: NCBI BLAST [1], MySQL [2], and Apache [3]. The server takes a multi-fasta format as input and performs an optimized pipeline of customizable BLAST searches on daily updated databases of microbial and eukaryotic species. The BLAST searches incrementally add data to a "job directory" that contains all job-relevant data in XML files. PHP scripts parse the results files, classify hit reads based on a configurable threshold given by the user (e-value, identity, read overlapping, maximum residual read length) collecting results into a MySQL database. For what concerns the alignment tasks, the workflow dispatches several parallelized BLAST searches on different computing nodes in order to achieve an overall high-performance computing time.

Results

The web server offers a full view of all the analyzed data by querying the results database through some specific search forms. Specifically, the pipeline is

[1] Consorzio per le Applicazioni di Supercalcolo per Università e Ricerca, Rome, Italy [2] Department of Biology, University of Rome Tor Vergata, Rome, Italy [3] Istituto Biomembrane e Bioenergetica, Consigli Nazionale delle Ricerche, Bari, Italy [4] Dipartimento di Biochimica e Biologia Molecolare, University of Bari, Bari, Italy

structured in three different phases: 1) detection of host reads (this phase is necessary in the case of metagenomic analyses of clinical samples); 2) detection of reads unambiguously assignable to known species; 3) assignment of residual reads to higher taxomic ranks. This latter phase may be accomplished by available software like MEGAN [4] using as input the BLAST output. The web service provides: i) an alignment view of each identified read, including the organism description, taxonomy ID and relative taxonomic tree recognition; ii) a taxonomic map of the unidentified reads, showing a wide-range taxonomic tree and highlighting the lowest common ancestor for reads that have been aligned on multiple organisms; iii) global statistics of species and organisms distribution among the samples, including the identification of the host reads for samples extracted from animal environments or tissues.

Contact e-mail

graziano.pesole@biologia.uniba.it

Supplementary information

References

[1] Altschul SFl, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–3402. doi: 10.1093/nar /25.17.3389. [2] MySQL . http://www.mysql.com [3] Apache. http://www.apache.org [4] Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res 2007;17 (3) :377-86.