Setup and benchmarking of a new scalable sequence alignment service for UHTS data.

Ivan Topolsky, Robin Engler, Marco Pagni and Heinz Stockinger.
SIB Swiss Institute for Bioinformatics

Introduction and project objectives

Aligning translated DNA sequence reads against a protein reference database is a highly demanding task in terms of computational resources. The software of reference for performing that type of sequence alignment, which is currently installed and used on Vital-IT's HPC cluster, is NCBI's BLAST (Basic Local Alignment Search Tool; *https://blast.ncbi.nlm.nih.gov/Blast.cgi*). While BLAST performs well to align a limited number of reads, scaling up to 10 or 100 of thousands of reads becomes extremely time consuming.

A number of alternative software packages has been developed with the promise to perform DNA sequence alignments to protein databases much faster. However, these are not installed yet on Vital-IT's HPC cluster and, most importantly, they have not yet been independently benchmarked to see if they really deliver the claimed increase in performance.

This project aims to install and evaluate – using benchmarking based on real-use cases – two new sequence alignment software: "SANSparallel" (*http://ekhidna2.biocenter.helsinki.fi/sans*) and "DIAMOND" (*http://ab.inf.uni-tuebingen.de/software/diamond*). Both of these claim to perform sequence alignments to protein databases much faster than the current aligner (NCBI BLAST).

Project outcomes

Software packages:

Software packages (.rpm files) have been built for SANSparallel version 2.0.0 (SANSparallel-2.0.0-2.el6.x86_64.rpm) and DIAMOND version 0.8.5 (diamond-0.8.5-2.el6.x86_64.rpm).

Benchmarking pipeline:

A pipeline of bash and R scripts that allow to evaluate the performance of sequence alignment software (time needed for alignment and quality of matches that are found) has been produced. By default, the UHTSbenchmark pipeline allows users to assess the performances of the "SANSparallel" and "DIAMOND" software against the reference software BLAST (Basic Local Alignment Search Tool). However, the UHTSbenchmark pipeline is written so that new "modules" can be added in order to benchmark new software (these "module" files are essentially a short text file containing the commands to run the software to be evaluated).

Benchmarking results of SANSparallel vs BLAST:

SANSparallel v. 2.0.0 was benchmarked against BLASTX v. 2.2.29. See details in the next section.

Benchmarking procedure

The benchmarking procedure was developed with the idea that it should (i) reflect as much as possible how actual users would use the software, and (ii) that is should be reproducible by anyone, and should thus be based on publicly available data.

The benchmarking scenario that was selected simulates the case were a user has an unknown DNA sample and wishes to know from which organisms it likely originates. This requires the sample to be aligned against the largest possible number of databases to identify which reference organisms has the most matches in the sample.

Reference database:

We used the publicly available OMA database that contains protein sequences for 1706 species (omabrowser.org/oma/home). The specificity of this database is that it is very well curated and that all (or almost all) ortholog genes that are contained within the database are correctly identified.

This enabled us to extract, for each species, a list of all ortholog genes (proteins), that could then be used as a reference to decide whether a match detected by the sequence alignment software (BALSTX or SANSparallel) was correct or not. It also allows compute, for each sequence aligned against the database, what is the percentage of matches that was found by the alignment software.

Definition box

Ortholog genes.
Ortholog definition: Orthologs are genes in different species that evolved from a common ancestral gene by speciation. Normally, orthologs retain the same function in the course of evolution. Identification of orthologs is critical for reliable prediction of gene function in newly sequenced genomes. (See also Paralogs).

Paralog genes.
Paralogs are genes related by duplication within a genome. In general, paralogs evolve new functions, even if these are related to the original one.

Test methodology:

Two different tests were implemented

Small scale test

1. 50 proteins are randomly selected from the OMA database. These are chosen while ensuring that none of them are orthologs. This protein subset is here-after be referred to as the "random sample".
2. All orthologs of the protein sequences from the "random sample" are extracted from the OMA database. These form the "ortholog sample".
3. A number of non-ortholog sequences equal to the number of sequences in the "ortholog sample" are randomly selected from the OMA database. This is the "non-ortholog sample". This sampling ensures that the number of protein sequences in the "ortholog sample" and the "non-ortholog sample" are equal.
4. A combined sample of the "ortholog sample" and "non-ortholog sample" is also created. This is the

"combined sample".

5. The following sequence alignments are performed for each sequence alignment software that is tested: "random sample" against "ortholog sample", "random sample" against "non-ortholog sample" and "random sample" against "combined sample". In all alignments, only the best match is kept.
6. The results of the different sequence alignments are then analyzed with the following expectations:
   - when the protein sequences from the "random sample" are aligned against the sequences from the "ortholog sample" we expect that the large majority of sequences are matched to their correct ortholog sequence.
   - in the "random sample" against "non-ortholog sample" sequence alignment, we expect to find few matches and that these matches are of poor quality (i.e. little overlap between sequences).
   - in the "random sample" against "combined sample", we expect that the alignment software should be able to select the correct ortholog sequence.

Large scale test

The large scale test is follows the same procedure as the small scale test, with the difference that the "random sample", "ortholog sample" and "non-ortholog sample" are now defined as follows:

- "random sample": all protein sequences for a given species randomly selected in the OMA database. The number of sequences in the "random sample" thus depends on the selected species.
- "ortholog sample": all protein sequences from the OMA database that are orthologs of sequences from the "random sample".
- "non-ortholog sample": randomly selected protein sequences from the OMA database that are non-orthologs of the sequences from "random sample". The number of non-ortholog sequences that are selected is equal to the number of sequences found in the "ortholog sample", so that both the "ortholog sample" and the "non-ortholog sample" have the same size.
- "leave-one out sample": equivalent of the "combined sample" of the small scale test, but this time using all OMA sequences except those that belong to the randomly selected species.

Benchmarking results.

The wall time needed by BLASTX to complete the sequence alignments varied between 7 and 25.5 hours depending on the number of threads that were used (Figure 1). In contrast, SANSparallel completed the alignment in less than 15 minutes. However, BLASTX found on average much more orthologs for each sequence than SANSparallel did (76% for BLASTX vs 40% for SANSparallel, Figure 2).

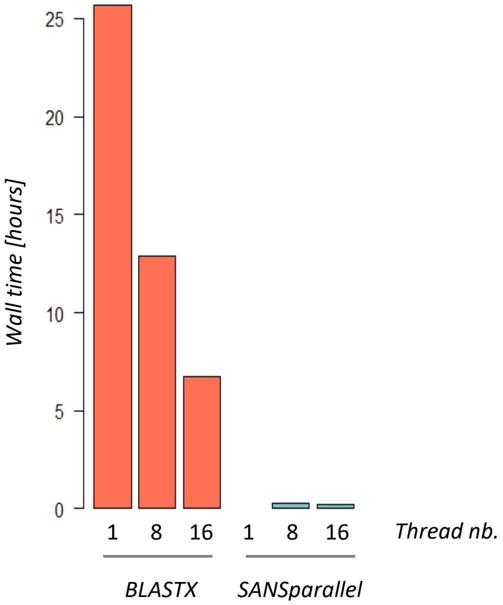

Figure 1. Run time (wall time) of BLASTX and SANSparallel runs when using respectively 1, 8 and 16 threads. SANSparallel needs at least 2 threads to run, explaining the lack of data in the "1 thread" column.

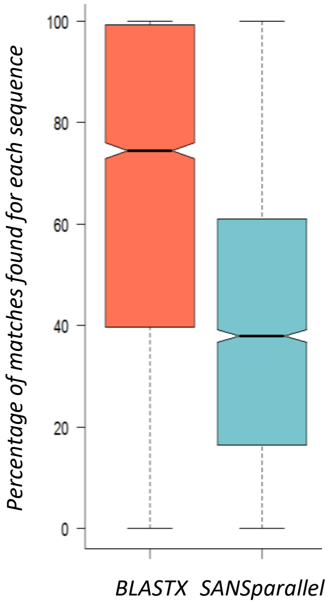

Figure 2. Percentage of true orthologs (according to the OMA database reference) found for each aligned sequence by BLASTX and SANSparallel.

Project outcome

Assessment of SANSparallel against BLAST shows that while SANSparallel is much faster it performs poorly at finding an exhaustive list of matches (it does however find some of the top matches also found by BLAST). The use of SANSparallel is thus recommended for the cases when the user wishes to quickly identify a number of top matches between sample sequences and a reference database, but not when obtaining an exhaustive list is the objective. The performance of DIAMOND was not assessed.

Reusability

Packages (.rpm files) to install SANSparallel v. 2.0.0 (SANSparallel-2.0.0-2.el6.x86_64.rpm) and DIAMOND v. 0.8.5 (diamond-0.8.5-2.el6.x86_64.rpm) are available for download (https://zenodo.org/; DOI "10.5281/zenodo.269799"). Note however that these rpm packages require that the Vital-IT "/software" environment must be present.

The benchmarking pipeline allows users to benchmark sequence alignment software in their own environment. The benchmark pipeline requires writing a "module" file (a small text files that contains the command needed to run the software) for each software that should be tested. Module files already exists for SANSparallel and DIAMOND but they will have to be adapted to a users' environment. The benchmarking pipeline can be downloaded from the zenodo archive (https://zenodo.org; DOI "10.5281/zenodo.269799").