Complete and comparative analysis of algorithms for whole genome shotgun assembly

Finotello F[1], Peruzzo D[1], Lavezzo E[2], Di Camillo B[1], Toffolo GM[1], Cobelli C[1], Toppo S[3]

Motivation

Assembly of high-throughput data from new-generation DNA sequencing has a great impact on sequence reconstruction. However, the available assembly tools are not fully characterized and compared yet. Here, we present a comparative assessment of the most widely used assembly algorithms, in terms of reliability of the assembled sequences.

Methods

A benchmark was implemented using the genomes of two Neisseria meningitidis bacteria belonging to serogroup C: the FAM18 sequence, available in the NCBI database, and an unknown sequence of a Neisseria. The latter was sequenced with 454 technology using a shotgun approach, resulting in 257 909 reads (240 bp average length). The benchmark consists of 1 790 786 base pairs divided in 384 strings composed only of those regions that are mapped by the reads (82% of the FAM18 sequence). The corresponding mapped reads were used as input for the assembly algorithms CABOG, Newbler, PCAP, PCAP.REP and PHRAP, using the default parameter settings. Results were evaluated only on the large contigs, i.e. long more than 500 bp, referred as contigs in the following. Software performance was assessed using the most widely applied statistics, i.e. the number of contigs, the average length, the longest contig length, the N50 contig size and the total contig span. An hypothetical coverage index (HCI) was also defined as the ratio between the number of bases in the contigs and the number of bases in the benchmark. Moreover, the contigs generated by each assembly software were realigned to the target sequences using MegaBLAST and the accuracy in sequence reconstruction was evaluated. For this purpose, the contigs were divided into four different categories, corresponding to different accuracy degrees and/or to different errors in the assembly procedure: - Correct: contigs that correctly reconstruct a region of the original genome; - Overhang: contigs with only a portion that correctly reconstructs the original genome; - Larger: contigs longer than the corresponding string in the benchmark; - Local errors: contigs containing a substring with an error rate larger than 80%. Finally, a real coverage index (RCI)

[1] Department of Information Engineering, University of Padova [2] Department of Histology, Microbiology, and Medical Biotechnologies, University of Padova [3] Department of Biological Chemistry, University of Padova

was defined as the ratio between the number of bases in the correct contigs, and the number of bases contained in the benchmark. This value provides the percentage of correct reconstruction, whereas the hypothetical coverage is only a measure of the number of bases in the reconstructed contigs.

Results

Table 1 reports the results obtained by the different algorithms. PHRAP presents the greatest HCI; however, more than 10% of the assembled contigs include several errors if compared to the benchmark. In particular, they are often longer than the strings they should span because they join different sequences taken from different regions of the original genome (chimeric contigs). CABOG, on the contrary, is very conservative: it produces fewer and smaller contigs, with the effect of obtaining the lowest hypothetical coverage value. On the other hand, when only correct contigs are considered, PCAP and Newbler provide the best results. However, PCAP presents a great difference between hypothetical and real coverage, meaning that the resulting contigs contain a large number of errors. If we consider both the real coverage value and the difference between hypothetical and real coverage, then Newbler performs best. This study demonstrates that a complete and structured analysis of the assembly software is not only useful, but even necessary to obtain a reliable and accurate reconstruction of a DNA sequence. Future developments include a study on different genomes using both shotgun and paired ends protocol, to assess the robustness and the best parameters setting of each tool. Then, a new assembly procedure will be investigated, integrating the most reliable algorithm results to increase the confidence in the sequence reconstruction.

Contact e-mail

denis.peruzzo@dei.unipd.it

Image

	CABOG	Newbler	PCAP	PCAP.REP	PHRAP
Number of contigs	425	450	476	465	414
Average length	3 753	3 709	3 658	3 736	4 291
Maximum length	22 338	41 706	16 196	16 196	64 064
N50 contig size	5 270	8 165	8 268	8 265	10 813
Total span	1 595 324	1 669 264	1 741 515	1 737 448	1 776 617
HCI	89.08%	93.21%	97.24%	97.02%	99.20%
Correct	97.65%	96.00%	93.07%	92.69%	82.61%
Overhang	0.70%	0.22%	1.47%	1.29%	4.35%
Local errors	0.00%	0.00%	0.00%	0.00%	0.00%
Larger	1.65%	1.78%	3.78%	4.30%	10.63%
Others	0.00%	2.00%	1.68%	1.72%	2.41%
RCI	85.03%	87.63%	88.69%	86.80%	67.07%
HCI – RCI	4.05%	5.58%	8.55%	10.22%	32.13%

Table 1: Results for large-contigs (i.e., more than 500 base pairs) analysis. The label "others" indicates contigs discarded because of bad reconstruction of the original sequence.