Enhanced Reference Guided Assembly

Vezzi F[1,2], Policriti A[1,2], Cattonaro F[2]

Motivation

The presence of several complete and draft genomes together with the advent of Next Generation Sequencing (NGS) technologies, allow analysis thought infeasible only a few years ago. Despite that, de-novo assembly remains a hard task and assembling new organism using only short reads is in practice extremely difficult. A reference sequence from a related organism, when available, can be used to assist the assembly of the new organism. In this case the sequences are first aligned against the reference and then a consensus sequence must be extrapolated. The only way to align the huge amount of sequences produced by NGS instruments is to use fast aligners like SOAP2. These aligners tend to be highly conservative: reads can be aligned only with a low number of errors and usually without gaps. While this allows to reconstruct the similarities between the two organisms we are unable to reconstruct the divergent parts. It is clear that a new strategy to assemble new organisms in presence of the sequence of a closely related species is necessary. This is especially true when using NGS data, as either reference guided or de-novo assembly in this case, present very peculiar problems. Here we propose a pipeline named Enhanced Reference Guided Assembly (e-RGA) that combines reference and de-novo assembly in order to obtain an improved assembly for a new sequenced organism in presence of a closely related reference.

Methods

Let R be the collection of short reads produced in the sequencing effort for a given higher organism B, and let A be the reference sequence belonging to a closely related organism of B. As shown in the attached figure there are essentially two possible ways to perform Reference Guided Assembly (RGA). The standard way, here named s-RGA, consists in aligning all R's reads on the reference A and then extracting the consensus sequence s-A. The other way, here named dn-RGA, consists in performing de-novo assembly on R, then aligning the resulting contigs on the reference sequence, and hence extracting the consensus sequence dn-A. This approach has the significant advantage of allowing the usage of BLAST-like tool to align, thereby permitting low similarity parameters and partial hits. The resulting assemblies are composed of a set of ordered and oriented contigs separated by gaps. As shown in Figure 1, once s-A and dn-A are available we

[1] Department of Mathematics and Informatics, University of Udine, Udine [2] IGA Istituto di Genomica Applicata, Udine

have to merge the two sequences. This situation is similar to the one already studied in the so-called Assembly Reconciliation (AR). The aim of AR is to merge the outputs of two different assemblers run on the same set of reads in order to obtain an improved final assembly. The merging step is the computationally most demanding. The hard task is the identification of the areas to be merged. Using an approach similar to AR and exploiting some specific constraints, we are able to solve this task in time linear on the length of A without performing global alignment between s-A and dn-A.

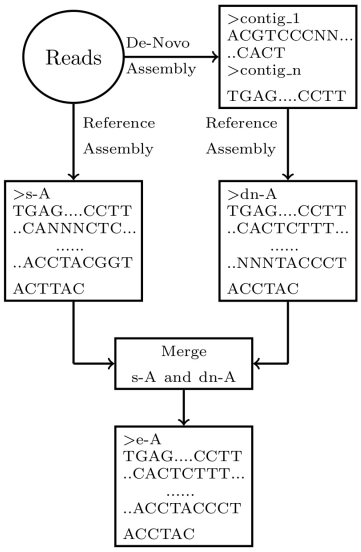
Results

The e-RGA pipeline has been implemented and tested on two different datasets. The first one consisted of four Conifer chloroplast genomes (Black, White, Red and Norway spruce) sequenced with an Illumina instrument GAII producing 45bp reads. A reference for pine chloroplast is available (Pinus thunbergii), while there is no reliable one for spruce chloroplast. Therefore, e-RGA was applied using the pine chloroplast as reference. The second data set consists of a set of 32bp reads from a microbial genome of length 2,7Mb, sequenced using an Illumina Instrument GAI. A reference sequence is present but the reads were sequenced from an enriched meta-population highly divergent from the reference. The results obtained show that e-RGA is able to produce an enhanced assembly: for all the five genomes both the number of reads aligned and the N50 contig size are improved on e-A with respect to dn-A and to s-A (see Figure 1 for details). The core of e-RGA is written in Perl and it is easily extensible to complex genomes.

Contact e-mail

francesco.vezzi@dimi.uniud.it

Image



Results for the four chloroplast genomes:

Experiment	Norway	White	Black	Red
Total reads	1911362	2245440	2175376	2571001
Al. on *A*	107338	278123	92454	50232
Al. *A* (%)	5.62%	12.39%	4.25%	1.95%
Al. on *s-A*	122764	313160	105076	55400
Al. *s-A* (%)	6.42%	13.95%	4.83%	2.15%
# Contig	326	307	323	339
N50	778	883	853	763
Al. on *dn-A*	135953	361189	120107	62256
Al. *dn-A* (%)	7.11%	16.09%	5.52%	2.42%
# Contig	106	59	106	159
N50	1477	2759	1288	1002
Al. on *e-A*	150663	382987	129649	67213
Al. *e-A* %	7.88%	17.06%	5.96%	2.61%
# Contig	129	113	130	156
N50	2614	3313	2465	1842

Results for the microbial genome:

Sequence	Aligned	Alingned (%)	# Contig	N50
A	863573	12.95%	1	
s-A	1160966	17.41%	7986	583
dn-A	1271461	19.07%	2630	710
e-A	1553050	23.29%	5600	1168

Figure 1: On the left a schematic representation of the e-RGA pipeline. On the right the two tables summarizing the results achieved with s-RGA, dn-RGA and, e-RGA on four chloroplast genomes and one microbial genome of length 2,7Mbp.