Guided contig assembly in de novo high-throughput sequencing of bacterial genomes

Petrillo M[(1)], Cozzuto L[(1,2)], Paolella G[(1,3)]

[(1)] CEINGE Biotecnologie Avanzate scarl, Via Comunale Margherita 482, 80145 Napoli, Italy
[(2)] S.E.M.M. -European School of Molecular Medicine - Naples site, Italy
[(3)] DBBM - Dipartimento di Biochimica e Biotecnologie Mediche,
Universita' di Napoli FEDERICO II, Via S. Pansini 5, 80131 Napoli, Italy

Motivation

High-throughput sequencing platforms such as 454 GS FLX and Illumina genome analyzer, have been recently used in genome sequencing projects. They produce very large amounts of data in a single one-day experiment, consisting of hundred thousands of sequence reads, but these reads are shorter than those obtained from Sanger sequencing and are assembled by ad hoc developed programs, such as EULER and Newbler, which are very effective in combining millions of reads into large contigs. Unfortunately the following step of fusing contigs into a unique genomic sequence still represents a major obstacle to the final assembly of these contigs, mainly because of the presence of repeated genomic regions, and the final assembly depends on providing additional experimental data, such as long-range connectivity information. Scaffolder is a new tool, developed to address these problems in de novo sequencing projects, which is able to detect links between contigs and solve most ambiguities deriving from repeated sequences. Scaffolder guides the overall assembly process by linking contigs into a multi-connected net, separating repeated sequences by a computational approach based on sequence microheterogeneities and selecting primer pairs to experimentally verify predicted links and untangle zones that cannot be computationally solved.

Methods

Scaffolder is a command-line tool written in PHP, which relies on a RDBMS for storing both the initial data and the subsequent results. Connections between contigs are identified by a BLAST-based search, based on the assumption that when a high sequence coverage has been reached, every base is likely to be covered by many reads, and that contigs fail to to be connected due to excess rather than lack of links. All contig bordering sequences are compared by BLAST with the complete set of reads, to identify possible links between contiguous contigs. i.e. reads spanning contig borders. The method results in connecting most contigs into complex linked networks, characterized by multiple weighted links. Such networks can be displayed as a multiple-connected graph by using visualization software, such as GRAPHVIZ. Multiply connected contigs derive from the fusion of reads from different repeats into a single multiple coverage contig. Analysis of micro-heterogeneities in read alignments is done by identifying all the original reads fused in a multiple contig by BLAST and by analyzing their alignment to the assembled sequence. Detecting single base variations and following them along the alignment, it is possible to assign most reads to the two or more variants of the contig and to separate the hidden components of a multiple contig. PCR experiment design is done by using eprimer3 and simulating PCR amplification. A user friendly web interface was built by using PHP objects.

Results

Scaffolder is able to detect, store and analyze connections between contigs, displaying the results as a graphical map. It highlights the presence of ambiguous links and automatically selects primer pairs to verify the connections and untangle multiply connected contigs by PCR experiments. It also uses experimental results to test for errors in the initial assembly, such as wrong bases or repeated sequences assembled together in the same contig. The web interface allows to review the assembly, to interactively duplicate multiple contigs and, once ambiguities are resolved, to join flanking contigs into larger scaffolds. Intermediate assembly versions are stored into a relational database and maps for each step may be quickly produced. Being organized in modules, it is very easy to add new modules for detections of new types of links. The tool has been applied to sequencing projects aimed to the de novo sequencing of large bacterial genomes by using a 454 GS FLX platform, where it has been successful in achieving a rapid and substantial reduction of a large number of contigs into a few very large scaffolds.

Contact : petrillo@ceinge.unina.it