論文内容の要旨

Dissertation Abstract

Title: Computational approaches for high-throughput sequencing and assembly of clone sequences

(クローン配列のハイスループット・シークエンシングとアセンブリのための計算機的アプローチ)

氏名：クロス　レジナルド　マサノブ

Introduction

Molecular cloning is an established and reliable way to isolate and to sequence specific fragments of DNA sequences. We can mention important applications of cloning such as the construction of full-length cDNA libraries and fosmid clone end-sequencing for whole genome sequencing and structural variation discovery. With the increase of throughput provided by second-generation sequencing technologies recently, application of these technologies in clone sequencing approaches can be considered to reduce the cost and the time consumed when traditional capillary sequencing is employed. However, computational problems arise mainly because of the short length of reads and the parallel property of the method. Here, I propose and develop computational approaches to tackle these limitations in the assembly of full-length cDNA clones and in the pooling of non-overlapping clones for high-throughput sequencing.

Methods

The replacement of capillary sequencers with Illumina GA is a cost-efficient extension to the existing approach of multiclone-shotgun sequencing of cDNA clones. As the high accuracy of the sequence that is required in these projects is fundamental to provide reliable information about the complete coding sequences, I propose a new *de novo*-reference hybrid assembly approach that generates sequences to fulfill that quality requirement. This new method, MuSICA 2, assembles full-length cDNA sequences of hundreds of clones from several short reads sequenced by Illumina GA, requiring Sanger reads from either or both ends of the clones to identify individual clone sequences in the assembly.

The assembly strategy used was a hybrid of reference assembly and *de novo* assembly, taking advantage of the benefits of both different approaches. Initial contigs are generated by *de novo* assembly of short reads that are generally fragmented because of the conservative aspect of this approach. Aligning disjoint contigs with the reference genome generates many overlapping

alignments. Because we assume that clone sequences are non-overlapping on the genome, then we can confidently merge overlapping contigs with the help of the reference genome. These alignments are also used to identify exon sequences as well as sequences across exon-intron junctions that might be missing in the assembly. Alignments of short reads with exon filling candidate regions and spliced-alignments of short reads with intron filling candidate regions were used to close gaps in the assembly. The final contigs are finally identified and associated with individual clones by using Sanger reads from clone ends.

In our assembly approach, I assume that simultaneously sequenced clones do not overlap on the genome. As clone end-sequences are often obtained before sequencing the full insert of clones, such as in cDNA clone sequencing and in genomic structural variation detection with fosmid libraries, then we can use their location information to define an optimization problem to select non-overlapping clones for high-throughput sequencing. The objective is to arrange clones that overlap on the genome in different pools, minimizing the number of sequencing runs. Overlaps on the genome are defined with an interval graph $G = (V, E)$, in which two overlapping clones $(v_1,v_2)\in E$ are assigned to different pools. If $E=\emptyset$, the specific problem corresponds to the classical bin-packing problem, or if the sequence sizes are ignored, then it is equivalent to the graph coloring problem. The non-overlapping clone pooling problem is a generalization of both specific problems and therefore is also computationally intractable. To provide near-optimal solutions for this problem, approximation algorithms based on bin-packing heuristics and optimal coloring are proposed and discussed.

Results

To evaluate the accuracy of the assembly of full-length cDNA sequences, I compared the output of our approach for human and toxoplasma cDNA clones with sequences finished by the traditional Sanger method. The assembly steps were successful in increasing the sequence contiguity. The exon-intron structure of the coding sequence was correct for more than 95% of the clones with coding sequence annotation and the nucleotide-level accuracy of coding sequences of those clones was over 99.99%. These results show that high sequence quality can be achieved for the sequencing of full-length cDNA clones using Illumina GA, reducing the cost and time by one order of magnitude when compared with traditional Sanger method.

To assess the performance of the approximation algorithms proposed for the non-overlapping clone pooling problem, real data from full-length cDNA clones and simulated end sequences from real fosmid clones were used. Optimal solutions could be found by a combination of optimal coloring and bin-packing heuristics, showing the applicability of the approach with real data.