Computing the Minimal Tiling Path from a Physical Map by Integer Linear Programming

Serdar Bozdag[1], Timothy J Close[2], and Stefano Lonardi[1]

[1] Dept. of Computer Science and Eng., University of California, Riverside, CA
[2] Dept. of Botany and Plant Sciences, University of California, Riverside, CA
{sbozdag,stelo}@cs.ucr.edu, timothy.close@ucr.edu

Abstract. We study the problem of selecting the minimal tiling path (MTP) from a set of clones arranged in a physical map. We formulate the constraints of the MTP problem in a graph theoretical framework, and we derive an optimization problem that is solved via integer linear programming. Experimental results show that when we compare our algorithm to the commonly used software FPC, the MTP produced by our method covers a higher portion of the genome, even using a smaller number of MTP clones. These results suggest that if one would employ the MTP produced by our method instead of FPC's in a clone-by-clone sequencing project, one would reduce by about 12% the sequencing cost.

1 Introduction

A physical map is a linear ordering of a set of clones encompassing one or more chromosomes. Physical maps can be generated by first digesting clones with restriction enzymes and then detecting the clone overlaps by matching the lengths of the fragments produced by digestion. A minimum-cardinality set of overlapping clones that spans the region represented by the physical map is called *minimal tiling path* (MTP).

The problem of determining a good set of MTP clones is crucial step of several genome sequencing projects. For instance, in the sequencing protocol called *clone-by-clone*, first a physical map is constructed, then the MTP is computed and finally, the clones in the MTP are sequenced one by one [9]. The clone-by-clone sequencing method has been used to sequence several genomes including *A. thaliana* [15] and *H. sapiens* [12] among others. Also, in several recent whole-genome shotgun sequencing projects, the MTP obtained from a physical map has been employed to validate and improve the quality of sequence assembly [23]. This validation step has been used, for example in the assembly of *M. musculus* [10], *R. norvegicus* [13], and *G. gallus* [18].

With the introduction of next-generation sequencing machines (454, Solexa/Illumina, and ABI SOLiD) we expect the MTP computation to become an essential step in *de novo* sequencing projects of eukaryotic genomes. Next-gen sequencing technology produces massive amount of very short reads (about 250bps for 454, 35bps for Illumina and SOLiD) [4] and therefore the *de novo* assembly of the whole eukaryotic genomes is extremely challenging [17]. Arguably, the only feasible method at this time is a clone-by-clone approach, where each clone in the MTP is sequenced using next-gen technology, and the assembly is resolved separately on each clone (see [17,21,25] and references therein).

K.A. Crandall and J. Lagergren (Eds.): WABI 2008, LNBI 5251, pp. 148–161, 2008.

If the exact locations of all clones in the physical map were known, computing its MTP would be straightforward; simply select the set of clones in the shortest path from the leftmost clone to the rightmost clone in the interval graph representing all the clones. This, however, is not a realistic solution. The noise in the fingerprinting data makes it impossible to build a perfect map. As a consequence, determining the minimal tiling path becomes a challenging computational problem. On one hand, a method that tends to select more clones as MTP might include many *redundant clones* (i.e., clones that do not provide additional coverage) and therefore it would waste time and money later in sequencing. On the other hand, an approach that tries to reduce the number of MTP clones may introduce gaps between the clones in some of the contigs, and thus reduce the coverage.

Although the problem of computing MTP has been studied extensively in the literature (see e.g. [22,15]), in practice there is only commonly used software tool, namely FingerPrinted Contigs (FPC) [7]. FPC provides three methods to compute an MTP, but only one uses solely restriction fingerprint data (hereafter called FPC-MTP). FPC-MTP computes the approximate overlap between clones in the contig and validates each overlap by using three extra clones, a *spanner* that verifies the shared fragments of the pair and two flanking clones that extend to the left and right of the pair and confirm fragments in the pair that are not confirmed by the spanner clone. Once the verification of the overlapping fragments between each clone pair is completed, a shortest path algorithm is used to find the minimal tiling path [2].

FPC-MTP's algorithm is quite good, but can be improved; our experimental results show that FPC-MTP is significantly distant from the optimal[1] MTP. In general, FPC-MTP selects fewer clones than necessary which in turns reduces the overall coverage. By changing parameters one can increase the coverage, but this comes at the cost of introducing many redundant clones. This limitation of FPC-MTP can be attributed to the fact that it checks the clone positions as an overall constraint when computing the MTP [2]. However, the positions of clones in contigs are known to be not very reliable [16].

Our contribution. We propose a new algorithm, called FMTP, that computes the MTP of a physical map based purely on restriction fingerprint data (and the contigs). In other words, our algorithm completely ignores the ordering of clones obtained by the physical map algorithm.

FMTP first computes a *preliminary* MTP by selecting the smallest set of clones that covers the genomic region that is covered by all clones in the contig. The problem of computing the preliminary MTP set is formulated in a combinatorial optimization framework as an Integer Linear Program (ILP). The preliminary MTP set may contain redundant clones. In the second phase, FMTP orders the clones in the preliminary MTP and computes the final MTP by using a shortest path algorithm.

We carried out an extensive set of experiments on the physical map of rice and barley. For the former dataset, the actual coordinates for the clones are known and therefore we could measure the accuracy of our algorithm. The experimental results show that the set of MTP clones computed by FMTP on the physical map for rice has higher coverage than the one produced by FPC (using approximately the same number of clones overall). This suggests that a larger portion of the genome could be obtained at the same cost when

[1] The optimal MTP is the one that we could compute if we knew the coordinates of all the clones.