A Mixed Integer Linear Programming Approach for Genome Haplotyping

Kerian Thuillier[1], Pierre Peterlongo[1], Rumen Andonov[1]

Univ Rennes, Inria, CNRS, IRISA, Rennes, France,
{ppeterlo,randonov}@irisa.fr, kerian.thuillier@ens-rennes.fr

Mots-clés : *flow network, maximum multi-commodity flow problem, genome haplotyping, variant phasing*

1 Introduction

Reading DNA has been made possible thanks to sequencers. A sequencer is a machine able to read short fragments of genomes, called the "*reads*". A read is a sequence composed of ≈ 100 to 200 characters. In this case the original localisation of each read in the originally sequenced genome is unknown. Reads can also be "*paired*", in this case, reads come by pairs, and in each pair the approximate distance on the genome between the two reads is known. The task of reconstructing the genome from the reads is called the "*assembly*".

An haplotype can be considered as one version of the genome. Diploid species, such as humans, contain two close versions of their genome, each called an haplotype. Polyploid species, as many plant species, contain $n > 2$ versions of their genomes (with n known). More generally, in metagenomics data (in which many species are sequenced together), n distinct genomes have to be reconstructed, with n unknown.

From sequencing data, haplotype-aware genome assembly consists in reconstructing all individual haplotypes contained in the sequenced sample. Thus, for diploid genomes, the goal is to reconstruct the sequence of each of the two haplotypes, and for polyploid and metagenomic data, the goal is to reconstruct the n sequences originally being sequenced. With actual sequencing technologies, there exist algorithmic solutions [1, 2], that are error prone and dedicated to small bacterial genomes.

In this context, we are working on a new method whose objective is to reconstruct a consequent fraction of each haplotype, with a nearly perfect precision. To do this we rely on the *DiscoSnp* [5] algorithm. *DiscoSnp* is a *de novo* variant detection tool. From one or several read set(s) it detects small variants called "*SNPs*" (Single Nucleotide Polymorphism). When two variants or more are detected at least once belonging to the same input read or pair of reads, one knows that those variants belong to the same molecule and thus to the same haplotype. Such variants are said "*phased*". However these fragments of haplotypes are of limited span, due to the short size of input reads.

In this work, our objective is thus the following : from a set of phased variants (fragments of haplotypes) and their abundances, we derive the total number of haplotypes we aim to reconstruct and we reconstruct highly reliable portions of their sequences. This is different from general assemblers which tend to collapse haplotypes. This is a mandatory step for downstream analyses such as *de novo* population genomics for instance. We propose a MILP approach for doing this task.

2 Our approach

We show that the above problem can be formulated in the framework of a directed graph $G = (V, E, w)$ where the vertices V correspond to the fragments of haplotypes while edges E

are associated with the overlaps between these fragments. The weight $w_v \geq 0$ for each vertex $v \in V$ corresponds to the abundance of the related fragment and is considered in our model as a lower bound. This bound corresponds to number of times that this vertex has been observed and can be seen as the minimal value of a flow passing through it. In addition, it is given a set of Paired Vertices (PV) indicating that the vertices u and v for any couple $(u, v) \in PV$ belong to the same haplotype. Haplotypes can be considered as commodities in our model. At this stage of the study their number is given by the user. We solve an MILP problem based on multi-commodity flow [4] where the commodities number is fixed, while the flow satisfies the given lower/upper bound constraints. The haploids' sequences are here found as paths associated to the commodities. Note that here the flow extremities (source/target) are unknown but are found by the model.

As we aim to reconstruct highly reliable portions of sequences and not the global sequences, loops could be ignored. Therefore, we search for elementary paths and Miller-Tucker-Zemlin (MTZ) technique [3, 4] is used to avoid cycles.

In contrast to [2], we seek to extract haplotypes from the maximum flow in a global way and not by greedy extraction. Our global approach avoids the need for heuristics, and optimizes the problem globally rather than locally. In addition, we use complementary information provided by sequencers : paired variants, *i.e.* variants present on the same haplotype, and contradictory variants, *i.e.* variants that cannot belong to the same haplotype. Note that our data contains noise induced by sequencers, this noise is also found in paired variants data. For this reason we maximize the number of satisfied couples instead of satisfying all of them.

Références

[1] Jasmijn A Baaijens and Alexander Scho. Sequence analysis Overlap graph-based generation of haplotigs for diploids and polyploids. (April) :1–9, 2019.

[2] Jasmijn A. Baaijens, Leen Stougie, and Alexander Schönhuth. Strain-aware assembly of genomes from mixed samples using variation graphs. *bioRxiv*, 2019.

[3] C. E. Miller, A. W. Tucker and R. A. Zemlin. Integer programming formulation of the traveling salesman problems. *Journal of the ACM*, 7(4) :326–329, 1960.

[4] Leonardo Taccari. Integer programming formulations for the elementary shortest path problem. *European Journal of Operational Research*, 252(1) :122–130, 2016.

[5] Raluca Uricaru, Guillaume Rizk, Vincent Lacroix, Elsa Quillery, Olivier Plantard, Rayan Chikhi, Claire Lemaitre, and Pierre Peterlongo. Reference-free detection of isolated snps. *Nucleic acids research*, 43(2) :e11–e11, 2014.