GenHap: Haplotype Assembly Using Genetic Algorithms

Tangherloni A(1)[†], Spolaor S(1), Rundo L(1,5), Nobile MS(1,6), Cazzaniga P(2,6), Liò P(3), Merelli I(4), Besozzi D(1), Mauri G(1,6)

(1) Department of Informatics, Systems and Communication, University of Milano-Bicocca, Viale Sarca 336, 20126 Milano, Italy
(2) Department of Human and Social Sciences, University of Bergamo, Piazzale Sant'Agostino 2, 24129 Bergamo, Italy
(3) Computer Laboratory, University of Cambridge, 15 JJ Thomson Avenue, CB3 0FD Cambridge, UK
(4) Institute of Biomedical Technologies, Italian National Research Council, Via Fratelli Cervi 93, 20090 Segrate (MI), Italy
(5) Institute of Molecular Bioimaging and Physiology, Italian National Research Council, Contrada Pietrapollastra-Pisciotto, 90015 Cefalù (PA), Italy
(6) SYSBIO.IT Centre of Systems Biology, Piazza della Scienza 2, 20126 Milano, Italy

[†] Email: andrea.tangherloni@disco.unimib.it

Motivation
Every somatic human cell is diploid and contains 22 pairs of homologous chromosomes and a pair of sex chromosomes, one copy inherited from each parent. The reconstruction of these two distinct copies, called haplotypes, of each chromosome allows us to fully characterize the genome of an individual [Levy et al., PLoS Biol 5(10), 2007]. The haplotyping process, which consists in assigning all heterozygous Single Nucleotide Polymorphisms (SNPs) to exactly one of the two chromosome copies, was shown to be one of the most prominent approaches to infer the full haplotype information related to a cell [Patterson et al., J Comput Biol 22(6), 2015]. SNPs are one of the most studied genetic variations as they play a fundamental role in many medical applications (e.g., drug-design, disease susceptibility studies, the expression of phenotypic traits) [Hirschhorn et al., Nat Rev Genet 6(2), 2005]. The knowledge of the complete set of all heterozygous SNPs of an individual provides more information than analyzing single SNPs. Given the huge amount of sequencing data available to date, computational approaches gained ground to solve this problem by dividing the entire sequencing dataset into k partitions, corresponding to the k different haplotypes. In the case of diploid organisms, 2^n possible haplotypes for n heterozygous SNP positions exist; in this context, the Minimum Error Correction (MEC) is one of the most used and promising methods to compute the two haplotypes by partitioning the sequencing reads into two disjoint sets with the least number of corrections to the SNP values [Wang et al., Bioinformatics 21(10), 2005].
Methods
GenHap [Tangherloni et al., submitted, 2018] is a novel computational method for haplotype assembly that can efficiently solve large instances of the weighted MEC

(wMEC) problem without relying on any a priori knowledge about the probability distributions of the sequencing errors in the reads. GenHap is based on Genetic Algorithms (GAs), which are suitable to deal with the combinatorial nature of the haplotype assembly problem. In particular, a population of randomly created solutions of GAs iteratively adapts to the user-defined fitness function, by means of selection, mutation and crossover operators. We propose a novel fitness function that relies on the extended Hamming distance between two ternary strings r and h, which represent a read and an estimated haplotype, respectively. Both strings are codified over the alphabet 0, 1, -, where 0 denotes a position that is equal to the reference sequence, 1 represents a SNP and - is used when a position is not covered by the read. In particular, the proposed fitness function computes the total number of positions in which both characters of r and h belong to 0,1, but at the same time they are different from -. So doing, a solution to the wMEC problem is obtained by minimizing the fitness function; the total number of errors is calculated by considering both the reads belonging to the first partition and the first inferred haplotype, and the reads belonging to the second partition and the second estimated haplotype. The computational complexity of the problem is addressed by exploiting a divide-et-impera approach in which the entire problem is partitioned into smaller and manageable sub-problems, and all calculations are distributed using a Master-Slave computing paradigm. Each sub-problem is tackled by means of a GA executed by each Slave process, which produces optimal sub-haplotypes from the point of view of corrections to the SNP values. When all the sub-haplotypes are calculated, they are combined by the Master process to provide the two final haplotypes.

Results

In order to evaluate the effectiveness of our approach, we compared the performance of GenHap against HapCol [Pirola et al., Bioinformatics 32(11), 2015], an efficient state-of-the-art algorithm for haplotype phasing, on two synthetic (yet realistic) datasets generated relying on the Roche/454 and PacBio RS II sequencing technologies. We tested GenHap's performances by assessing the average haplotype error rate (HE) and the average running time. The former represents the amount of erroneously assigned SNPs with respect to the ground truth, while the latter is the time required to reconstruct the haplotypes of a given instance. In the performed tests, GenHap reconstructed accurate haplotypes, independently of the number, frequency and coverage of SNPs in the input instances. To be more precise, in the case of the Roche/454 dataset, GenHap inferred the haplotypes with an average HE lower than 0.1%. For what concerns the PacBio RS II sequencing dataset, all the haplotypes were reconstructed with an average HE lower than 2.2%. Our results also show that GenHap is 2 times faster than HapCol in the case of Roche/454 reads, and up to 7 times faster in the case of data from PacBio RS II sequencing systems.