Master thesis of *Hui-E Yang*

On solving population haplotype inference problems

Hui-E Yang
Institute of Information Management
National Cheng Kung University

Due to the completion of Human Genome Project, we have known more about the structure and sequence, but not yet the function of human DNA. The function of DNA may be a key for human disease. To analyze the function of DNA, researchers have to obtain each haplotype, the genetic constitution of an individual chromosome, of an individual for analysis. Nevertheless, considering the significant efforts required in collecting haplotypes, usually the descriptions of one conflated pair of haplotypes called genotypes are collected.

Since the genotype data contains insufficient information to identify the combination of DNA sequence in each copy of a chromosome, one has to solve the population haplotype inference (PHI) problem which infers haplotype data from genotype data for a population. Previous researchers use mathematical programming methods and heuristic algorithms to solve the population haplotype inference problem. This thesis surveys these methods and conducts computational experiments on the efficiency and effectiveness for these methods in solving a population haplotype inference problem based on pure parsimony criterion (HIPP) which seeks the minimum number of distinct haplotypes to infer a given genotype matrix.

We propose two heuristic algorithms to solve the HIPP problem with promising performance. The first heuristic algorithm exploits the compatible relations among genotypes to solve a reduced integer linear programming problem in a smaller solution space. The second heuristic algorithm selects popular haplotypes that can resolve more genotypes in a greedy fashion. Extensive computational experiments have been conducted for several PHI solution methods on both the biological and simulated data. The results show that our proposed algorithms are efficient and effective, especially for solving cases with larger recombination rates. Finally, we give a divide-and-concur technique to solve large-scale HIPP problems. We also improve a parsimonious tree growing heuristic to obtain all the multiple optimal solutions for an HIPP problem.

***Keywords*:** haplotype inference, genotype, pure parsimony, heuristic algorithm, integer programming

Thesis advisor: *I-Lin Wang,* ilinwang@mail.ncku.edu.tw
http://ilin.iim.ncku.edu.tw
***Hui-E Yang* entered IIM of NCKU in 2004, graduated in 2006.**