Master thesis of *Chia-Yi Ma*

TagSNP selection problems based on linkage disequilibrium and Lagrangian Relaxation

Chia-Yi Ma

Department of Industrial and Information Management

National Cheng Kung University

Among all possible DNA sequence variations, Single Nucleotide Polymorphism (SNP) is the most common genetic variation. Sequence of closely linked SNPs composes the haplotype. Changes in the sequence can have significant influence on disease occurrences and the phenotype of human traits. The changes can be applied for identifying diseases and other medical research. At present, there are voluminous data of discovered SNPs. To make SNP database more cost-effective, many researchers propose tagging SNPs, a minimal subset of SNPs called tagSNP, to capture the full information of the original SNP sequence (haplotype).

Various SNP tagging methods have been proposed in the literature, based on different purposes of applications. This paper focuses on the most commonly cited definition of tagSNP and proposes methods to select them. In particular, we first seek the smallest SNP subset to identify all haplotype patterns. The tagSNP Selection Problem can be modeled as a 0-1 binary integer programming problem with multiple optimal solutions. Previously, scholars focused on improving the efficiency of their solution methods, and ignored the differences among multiple optimal solutions. To this end, we proposes a Heuristic algorithm based on graph theory that first solves for all the multiple optimal solutions and then recommends a more informative set of optimal solution based on the concept of Linkage Disequilibrium (LD). Our method calculates the relevancy between a selected tagSNP and the other SNPs, which later serves as an indicator for assessing the amount of information it carries. Among all the multiple optimal solutions, we recommend the one with the largest sum of LD values. In addition, we also propose a bi-objective integer programming model which tries to minimize the number of selected tagSNPs while maximize the sum of their LD values.

To deal with large-scale tagSNP selection problems, we propose a heuristic called LRH, based on the theory of Lagrangian Relaxation. In particular, a modified subgradient method is proposed to update the Lagrangian multiplier which in turn approaches to the optimal solution step by step. In LRH, we incorporate the concept of Greedy Algorithm which selects some good SNP column, gradually reduces the problem size, and thus greatly improves the efficiency and quality. The computational results show that LRH has good efficiency and effectiveness, since it can quickly converges to a better solution.

Thesis advisor: *I-Lin Wang*, ilinwang@mail.ncku.edu.tw
http://ilin.iim.ncku.edu.tw
***Chia-Yi Ma* entered IIM of NCKU in 2006, graduated in 2008.**

Moreover, we also give a two-stage solution method (MIX) which first uses LRH to select good candidate SNP columns and then solve a reduced tagSNP selection problem of much smaller size based on the selected candidate SNPs. The computational results show that the proposed two-stage approach converges to a better solution in a much shorter time.

Finally, in a suggested future research topic, we consider the case where the selected tagSNPs is to be included in a biochip, while the capacity for the biochip is sufficiently large. In this case, we no longer have to select the SNPs of minimum size. Instead, we focus on selecting those tagSNPs that can be used to differentiate haplotypes as many as possible so that the selected SNPs are more robust or reliable.

***Keywords*:** tagSNP, linkage disequilibrium， haplotype， algorithm， graph， Lagrangian Relaxation

Thesis advisor: *I-Lin Wang,* ilinwang@mail.ncku.edu.tw
http://ilin.iim.ncku.edu.tw
***Chia-Yi Ma* entered IIM of NCKU in 2006, graduated in 2008.**