Higher reliable tag SNP Selection Strategies in Plants

Author: Tai-Chun Wang, twan5724@it.usyd.edu.au
Supervisor: Professor Albert Zomaya
School of Information Technologies

1. Aims of this Project

- Designing a better method for discovering Single Nucleotide Polymorphisms (SNPs) from largely redundant data sets.
- Selecting higher reliable tag SNPs for further experiments.

2. Introduction

Analysing DNA sequence variation is one of the major subjects of genetic studies. In this context, the availability of molecular biomarkers are fundamental within agricultural biology and plant breeding (such as identifying individuals and across generations). Compare with tandem repeat sequences, SNPs have lower mutation rate and are available more; therefore, their usage for genetic studies is massively increased nowadays. They are also used as biomarkers/tools to detect alleles associated with genetic diseases and/or to identify individual breeds.

Because in *vivo* SNP discovery process is very time consuming and also very expensive, in *silico* methods are usually used for this purpose. These methods are mainly used to discover SNPs from public data sets as well as to select a tag subset to capture unparsed SNPs. The latter problem has been proved to be NP-Hard.

In this project, two major issues in selecting SNPs are under focus:

How to efficiently eliminate/filter noise form public data sets.

How to effectively select higher reliable tag SNPs.

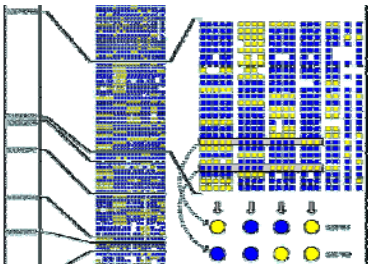

Figure 1: A part of SNPs from 20 samples of human chromosome 21.

3. Material

- Oryza sativa spp.Indica and Oryza sativa spp.Japonica EST data sets were downloaded from National Centre for Biotechnology Information dbEST (Mar - 2009). These datasets contain 106,532 and 833,672 ESTs, respectively. The Single Nucleotide Polymorphism database (dbSNP) was downloaded from NCBI (rice_4530) for data verification.
- Human chromosome 21 haplotype information was downloaded from HapMap (released 27-2009).

4. Discovering SNPs from EST

We design a pipeline method, namely SASNP, to deploy two external software packages to assemble all EST data sets before discovering SNPs (Figure 2). Discovering part of SASNP consists of three statistical strategies to eliminate/filter noise. To identify sequence errors, slide window strategy is the first steps deployed here. After that, based on the assumption that each contig has only one haplotype, normal distribution analysis in conjunction with modified neighbourhood quality standard method is used to discard SNPs with 'unusual' values in comparison with their neighbours.

We compared result of SASNP with two in *silico* methods: autoSNP and QualitySNP. autoSNP is a SNP discovering method employed by Australian Centre for Plant Functional Genomic for four different plant species . QualitySNP however showed better performance in potato ESTs experiment. Figure 3 showed the superiority of SASNP compared with these two methods for the downloaded datasets.

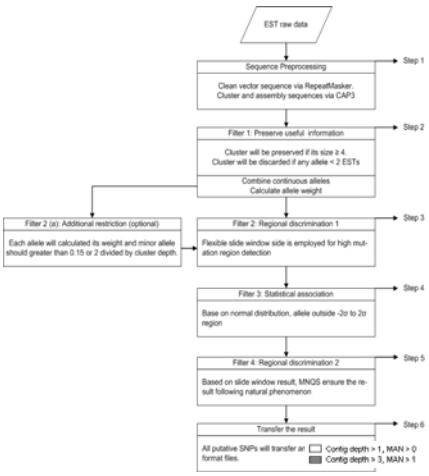

Figure 2: The pipeline structure of SASNP

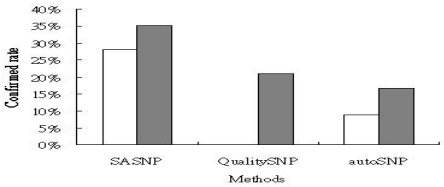

Figure 3: The performance comparison between SASNP and other two methods. Confirmed rate equals the percentage of how many hits we found in dbSNP for each results.

5. Tag SNP selection strategy

This problem can be defined from two perspectives: LD-based and block-based. In the Linkage disequilibrium (LD) based perspective, a statistical measurement is used to detect associations between paired alleles.

In the Block-based perspective, on the other hand, all SNPs on the same chromosome are partitions into several blocks to complete haplotypes. In this case, the aim is to find a minimum subset of SNPs to distinguish haplotypes in each block.

In the tag SNP selection, we try to combine advantages of these two perspectives to built a new tag SNP selection strategy that considers both inter- and intra-block LDs simultaneously. In our method, haplotypes are first partitioned into several fragments (block-based), then, LD values for each block is recorded to a decision making reference table. Using this table, pair of alleles with highest LD values in each block are used (as candidate tag SNPs) to measure their LD values with other pairs in other blocks.

6. Future Works

Tag SNP selection algorithm is an ongoing project. Currently, our method is using brute force method to find the optimal result in human chromosome 21 data set. However, because brute force method needs a large amount of memory to handle the experiment data and is also a very time-consuming process, an efficient algorithm for solving this issue is in next step in our work.

The other extension to our work is to use several potential methods like Bayesian networks to improve the performance of our algorithm. This is mainly because inter-block LD problem could be translated into finding maximum weight in a complete graph (figure 4).

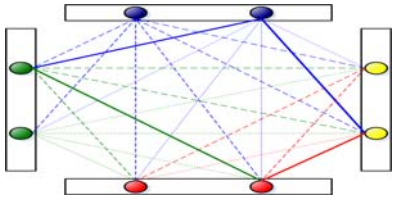

Figure 4: Selecting maximum association tag SNPs from a complete graph.

The University of Sydney