LinkageTracker: A Discriminative Pattern Tracking Approach to Linkage Disequilibrium Mapping

Li Lin[1], Limsoon Wong[2], Tzeyun Leong[3], and Pohsan Lai[4]

[1,3] School of Computing, National University of Singapore
[2] Institute for Infocomm Research, Singapore
[4] Dept of Pediatrics, National University Hospital, National University of Singapore
{linl, leongty}@comp.nus.edu.sg, limsoon@i2r.a-star.edu.sg,
paelaips@nus.edu.sg

Abstract. Linkage disequilibrium mapping is a process of inferring the disease gene location from observed associations of marker alleles in affected patients and normal controls. In reality, the presence of disease-associated chromosomes in affected population is relatively low (usually 10% or less). Hence, it is a challenge to locate these disease genes on the chromosomes. In this paper, we propose an algorithm known as LinkageTracker for linkage disequilibrium mapping. Comparing with some of the existing work, LinkageTracker is more robust and does not require any population ancestry information. Furthermore our algorithm is shown to find the disease locations more accurately than a closely related existing work, by reducing the average sum-square error by more than half (from 80.71 to 30.83) over one hundred trials. LinkageTracker was also applied to a real dataset of patients affected with haemophilia, and the disease gene locations found were consistent with several studies in genetic prediction.

1 Introduction

Linkage disequilibrium mapping has been used in the finding of disease gene locations in many recent studies [6][13]. The main idea of linkage disequilibrium mapping is to identify chromosomal regions with common molecular marker alleles[1] at a frequency significantly greater than chance. It is based on the assumption that there exists a common founding ancestor carrying the disease alleles, and is inherited by his descendents together with some other marker alleles that are very close to the disease alleles. The same set of marker alleles is detected many generations later in many unrelated individuals who are clinically affected by the same disease. In a realistic setting, the occurrence of such allele patterns is usually very low, and most often consist of errors or noise. For instance, the hereditary mutations of BRCA-1 and

[1] A molecular marker is an identifiable physical location on the genomic region that either tags a gene or tags a piece of DNA closely associated with the gene. An allele is any one of a series of two or more alternate forms of the marker. From the data mining aspect, we could represent markers as attributes, and alleles as attribute values that each attribute could take on.

L. Zhou, B.C. Ooi, and X. Meng (Eds.): DASFAA 2005, LNCS 3453, pp. 30 – 42, 2005.

BRCA-2 genes only account for about five to ten percent of all breast cancer patients[12]. Assuming that we know that BRCA-1 gene resides somewhere on chromosome 17, the finding of the exact location of BRCA-1 gene on chromosome 17 based on a set of sample sequence collected from breast cancer patients where at most ten percent of the sample sequence exhibit allelic association or linkage disequilibrium is a nontrivial task. To further complicate this task, the linkage disequilibrium patterns also consist of errors due to sample mishandling and contamination.

Due to errors and low occurrence of linkage disequilibrium patterns, existing data mining and artificial intelligence methods involving training and learning will not be applicable. In this paper, we propose a novel method known as *LinkageTracker* for the finding of linkage disequilibrium patterns and inference of disease gene locations. First of all, we identify the set of linkage disequilibrium patterns using a heuristic level-wise neighbourhood search and score each pattern by computing their *p*-values to ensure high discriminative powers of each pattern. After which, we infer the marker allele that is closest to the disease gene based on the *p*-value scores of the set of linkage disequilibrium patterns. *LinkageTracker* is a nonparametric method as it is not based on any assumptions about the population structure. The method is robust to cater for missing or erroneous data by allowing gaps in between marker patterns. Comparing our method with Haplotype Pattern Mining (*HPM*) which was reported by Tiovonen et. al. [16], *LinkageTracker* outperforms *HPM* by reducing the average sum-square error by more than half (from 80.71 to 30.83) over one hundred trials.

Organization of This Paper. In the next section, related work will be introduced, followed by a technical representation of the problem and a detailed description of the LinkageTracker algorithm. Next, the optimal number of gaps to set on LinkageTracker to achieve good accuracy will be discussed. We will then evaluate the performance of LinkageTracker with a recent work known as Haplotype Pattern Mining (HPM). Finally, we conclude our paper with a summary and the directions for future work.

2 Related Works

There are generally two methods used for detecting disease genes, namely, the direct and the indirect methods. Techniques used in the direct method include allele-specific oligonucleotide hybridization analysis, heteroduplex analysis, Southern blot analysis, multiplex polymerase chain reaction analysis, and direct sequencing. A detailed description of these techniques is beyond the scope of this paper but is available in [3] and [10]. Direct method requires that the gene responsible for the disease be identified and specific mutations within the gene characterized. As a result, direct method is frequently not feasible, and, the indirect method is used. The indirect methods such as [7], [14], and [16] involves the detection of marker alleles that are very close to or are within the disease gene, such that they are inherited together with the disease gene generation after generation. Such marker alleles are known as haplotypes. Alleles at these markers often display statistical dependency, a phenomenon known as linkage