

Haplotype Pattern Mining & Classification for detecting disease associated Site

Takashi Kido¹, Masanori Baba¹, Hirohito Matsumine¹, Yoko Higashi²,
Hirotaka Higuchi², Masaaki Muramatsu^{1,3}

¹ *HuBit Genomix, Inc., JowareHanzomon 2F, 2-19, Hayabusa-cho, Chiyoda-ku, Tokyo, Japan*

² *NTT Data Co. Ltd., Kayabacho Tower, Shinkawa 1-21-2, Chuo-ku, Tokyo, Japan.*

³ *School of Biomedical Science, Tokyo Medical Dental University, 2-3-10, Kandasurugadai
Chiyoda-ku, Tokyo, Japan
tkido@hubitgenomix.com*

Abstract

Finding the causative genes for common diseases using SNP (Single Nucleotide Polymorphism) markers is now becoming a real challenge. Although traditional statistical SNP association tests exist, these tests could not explain the effects of SNP combinations or probable recombination histories from ancestral chromosomes. Haplotype analysis of disease associated site provides more powerful markers than individual SNP analysis, and can help identify probable causative mutations. In this paper, we introduce a new method for effective haplotype pattern mining to detect disease associated mutations. Using this procedure, we can discover some of the new disease associated SNPs, which can not be detected by traditional methods. We will introduce a powerful tool for implementing this procedure with some worked examples.

1. Introduction

Gene mapping of complex diseases has been a major challenge due to its etiological complexity, such as environmental factors, gene-environment interactions, as well as gene-gene interactions. These factors make it very difficult to detect the causative genes. Because the power of existing algorithm to detect gene with weak effect is low in explicit statistical modeling approaches, more powerful association methods should be explored [2] [3] [4]. This paper introduces a method of detecting the disease associated sites by fully utilizing haplotype information.

2. Method

The summary of our developed algorithm is as follows;
(1) We define the haplotype patterns as the possible SNP patterns denoted by the set of {A, B,*} Here, "*" represents A or B.
(2) We detect similar patterns which are frequently appeared in disease associated sites, which we call haplotype pattern mining. All possible haplotype combinations are examined with this purpose.
(3) We also compare all possible pairs of haplotypes which differ in only one SNP. If we detect the significant difference in association tests on these two haplotypes,

we make assumptions that this mutation may be causative SNP for the disease.

(4) We classify the haplotypes based on their phylogenetic similarities and generate haplotype trees (or graphs), where each node represents a haplotype, and edge represents a transition event (mutation or recombination). We calculate the disease association trend for each haplotype and map it into the haplotype trees.
(5) We can localize the position of disease associated site and make hypothesis on causative SNP based on above analysis.

3. Example

Our method uses haplotypes as input. In this example, we use the 8 haplotypes (H1, H2, ..., H8; each of H_i consists of 5 SNPs) with their case/control frequencies estimated by LDsupport [1]. Significant genetic heterogeneity among 8 haplotypes is observed in T5 permutation association test with $P=0.0043$. The most "major haplotype", meaning most frequently appeared, is H1 (AAABA) which frequency is 45% in case and 36% in control samples. In order to localize the disease-associated locus of the haplotype, we search for all $3^5 - 1 = 342$ haplotype patterns, where BAABA, BA**A etc, are some of these examples. (Step 1)

Table 1 shows an example of strongly disease-associated patterns discovered in our data by applying haplotype pattern mining, where we say that P is "strongly associated with the disease" if $\chi^2(P) \geq k$, given a (positive) association threshold k . (In this example, $k = 3.84$.) In the "+/-" sign of the right column of the table, "+" means haplotype pattern P is more frequent in cases than in controls and respectively. All patterns exceeding the threshold can be enumerated efficiently with data-mining algorithms. [3] In this example, A*A** is derived to be the most significant disease risk marker. (Step 2)

Table 2 shows the significant phenotypic changes in the pairwise comparisons among one-different mutant

haplotype patterns. It is reported that this kind of pairwise comparisons, such as cladistic analysis works more effectively than single SNP markers as risk predictors [2]. This example shows that the mutation on first locus (AA*** =>BA***) is the most significant. (Step 3)

Figure 1 shows the haplotype classification, where each node represents a haplotype, each arrow indicates one mutational event, and dotted line indicates one recombination event. The size of the node represents its frequency; color with darkness represents the disease associated trend. This example shows that H4 is the most strongly disease-associated and "H1 => H4" (first locus mutation of A=>B) is the most significant mutation event. (Step 4)

Figure 2 shows the statistics on 5 loci for localizing the causative position by 3 different measurements; the p value distribution of the single SNP analysis for each locus, histogram of the marker frequency of the strong disease associated haplotype patterns in step 2, and also the marker frequency of the patterns in step 3. These 3 results suggest that the left sided SNP "1" is strongly disease associated. By these observation, we can make an assumption that causative site may locate in the left side of the gene, which occurred through the historical recombination event shown in step 4. (Step 5)

4. Summary and Future Work

We introduced a method for effective haplotype pattern mining to detect disease associated SNPs with some examples. We believe that this data mining approach can complement some of the weaknesses inherited in statistical modeling approaches, especially in the case of complex disease analysis where genetic effects are occasionally weak. Further systematic evaluation of the method and comparison with other approaches such as BLADE algorithm [4] will solidify the validation of our method.

Reference

[1] Kitamura Y, et al., "Determination of probability distribution of diplotype configuration (diplotype distribution) for each subject from genotypic data using the EM algorithm", *Ann. Hum. Genet.*, 66, 183-193, 2002

[2] Alan R. Templeton, "A Cladistic Analysis of Phenotypic Associations with Haplotypes Inferred from Restriction Endonuclease Mapping or DNA Sequencing. V. Analysis of Case/Control Sampling Designs: Alzheimer's Disease and the Apoprotein E Locus", *Genetics* 140: 403-409 (May, 1995)

[3] Hannu T. T. Toivonen, et al., "Data Mining Applied to Linkage Disequilibrium Mapping", *Am.J.Hum.Genet.* 67:133-145, 2000

[4] Jun S. Liu, et al., "Bayesian Analysis of Haplotypes for Linkage Disequilibrium Mapping", *Genome Research*, 11:1716-1724, 2001

Table 1. Strongly disease associated patterns obtained by haplotype mining (Example)

Haplotype	Frequency in case (%)	Frequency in control (%)	P value	sign
A*A**	97.9	33.6	0.00037	+
A*AB*	97.6	33.6	0.000468	+
BA**A	18.2	13.5	0.00145	+
B***A	18.9	13.5	0.0027	+

Table 2. Pairwise comparisons of the mutant patterns (Example)

Pattern 1	Pattern 2	Locus	P value	sign
AA***	BA***	1	0.00031	+
AA**A	BA**A	1	0.00049	+
A**B*	B**B*	1	0.00054	+
AA***	AB***	2	0.016	+
A**AA	A**BA	4	0.022	-

Figure 1. Haplotype classification and association trend mapping (Example)

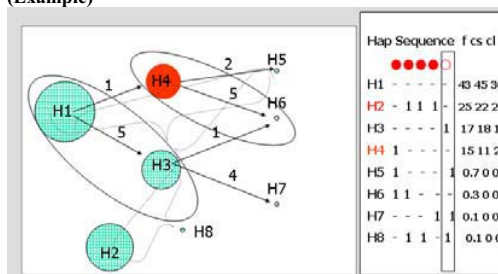


Figure 2. Localizing the disease associated locus (Example)

