The Fourth International Conference on Computational Systems Biology (ISB2010)
Suzhou, China, September 9–11, 2010
Copyright © 2010 ORSC & APORC, pp. 147–154

# Identification of Disease Locus Using Constrained Scaling Models[*]

Yiu-Fai Lee[*,†] Michael K. Ng[*,‡]

[*]Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong

**Abstract** In this paper, we develop an optimization approach to identify a position in a region where the corresponding SNP is associated with the disease locus optimally in the scaled distances among SNPs. The optimization model involves the construction of genetic mappings with the linkage disequilibrium among SNPs. A simulation study is given to illustrate the effectiveness of the model.

**Keywords** Disease locus, association, genetic mapping, linkage disequilibrium

## 1 Introduction

Linkage disequilibrium analysis offers the prospect of fine scale localization of genetic polymorphisms of medical importance, particularly when single nucleotide polynorphisms (SNPs) are densely appeared in a candidate region. The role of linkage disequilibrium (LD) is to identify and then narrow a candidate region. Because of the complex observed patterns, the modeling of the relationship between SNP markers and disease phenotypes is required [14]. Maniatis et al. [10] developed a metric LD map with additive distances in LD units based on the Malecot model. The application of LD maps to association mapping and positional cloning was studied in [11].

Because of their ubiquity there has been considerable interest in using single nucleotide polymorphisms (SNPs) to fine-map susceptibility loci [1, 13]. It is estimated that 90% of naturally occurring sequence variations are SNPs [3, 4] and these are sufficiently finely spaced that one may reasonably expect to find several within a defined chromosomal region which can be small enough to manifest detectable linkage disequilibrium in at least some human populations. Detecting association between SNPs and disease may provide useful evidence for the existence of a susceptibility locus within such a region, allowing one to proceed to more intensive investigations which can lead to identification of the gene and pathogenic polymorphisms.

Several strategies have been proposed by utilizing two-point methods to localize the position of a disease locus [5]. However, SNPs studied individually might be expected to provide relatively little information for detecting association between a disease and a

[*]Research supported in part by RGC 201508 and HKBU FRGs.

[†]ahyiu@graduate.hku.hk

[‡]mng@math.hkbu.edu.hk, the corresponding author

chromosomal region [14, 16], especially if more than one mutation is present. Potentially the amount of information available from SNPs could be increased dramatically by utilizing information from several marker loci simultaneously, with the aim of detecting association with a marker haplotype rather than just one biallelic marker. Composite likelihood methods combining disease associations with a series of linked markers from haplotypes have been proposed by Collins et al. [4], Lam et al. [8] and McPeek & Strahs [12]. With many of these methods the emphasis is to identify as closely as possible the probable position of the disease gene relative to the markers.

The main aim of this paper is to develop an optimization approach to identify a position in a region where the corresponding SNP is associated with the disease locus optimally in the scaled distances among SNPs. The optimization model involves the construction of genetic mappings with the linkage disequilibrium among SNPs.

The outline of this paper is as follows. In Section 2, we study our constrained scaling models. In Section 3, we present our optimization approach to identify a disease locus position. In Section 4, a simulation study is given to illustrate the effectiveness of the model. Finally, a concluding remark is given in Section 5.

## 2 Constrained Scaling Models

Let us consider two biallelic SNPs, where the rarest allele has frequency $p$, and is positively associated with an allele at the other SNP, which has frequency $q$. The haplotype frequencies of the 2 SNPs can then be represented in a 2-by-2 table as follows:

| | Allele B | Allele b | |
|---|---|---|---|
| Allele A | $pq+d$ | $p(1-q)-d$ | $p$ |
| Allele a | $(1-p)q-d$ | $(1-p)(1-q)+d$ | $1-p$ |
| | $q$ | $1-q$ | $1$ |

The parameter $d$ is defined as the linkage disequilibrium (LD) between the two SNPs. Because of the above allele assignment for $p$ and $q$, we have $p <= 1/2$, $p \leq q$, $p \leq 1-q$ and $d \geq 0$. The scaled measure of linkage disequilibrium between the two SNPs is defined as follows: $d' = \frac{d}{p(1-q)}$, Note that $d' = 1$ requires only $d = p(1-q)$ or $d = (1-p)q$. Since $d'$ decays by a factor of $1-\theta$ per generation where $\theta$ is the recombination fraction, the function $-\ln d'^2$ has the property that it is proportional to $-\ln(1-\theta)$. Note that for small values of $\theta$, $-\ln d'^2$ is approximately proportional to $\theta$, and therefore is also proportional to genetic map distance measured in units of Morgan, see for instance [14]. The LD distance between the $i$th SNP and the $j$th SNP can be given by $l_{ij} = -\ln d'^2_{ij}$. For a set of $n$ SNPs, their inter-marker LD distances can be represented in an $n$-by-$n$ matrix $[l_{ij}]_{i,j=1,2,\cdots,n}$.

We require a 1-dimensional representation of the SNPs, preserving the order of the SNPs on the chromosome, such that the distances between SNPs along this dimension are close to the distances in the $n$-by-$n$ LD distance matrix. The classical metric unidimensional scaling problem is to place $n$ objects on the real line, so that the interpoint distances best approximate the observed dissimilarities between pairs of objects. It is well-known that this problem is equivalent to an NP-hard combinatorial problem [9]. However, in the constrained unidimensional scaling problem, the objects are required to place in a given order. In our context, the order of the SNPs is already given as follows: 1st, 2nd, 3rd, $\cdots$, $n$th Therefore, the key issue is to determine the nonnegative interpoint distances among the ordered SNPs that best approximate the observed dissimilarities between pairs of SNPs. Mathematically, the problem is to minimize the objective function:

$$J(z_1, z_2, \cdots, z_{n-1}) = \sum_{i>j} w_{ij} \left( l_{ij} - \sum_{k=j}^{i-1} z_k \right)^2, \qquad (1)$$

subject to

$$z_k \geq 0, \quad k = 1, 2, \cdots, n-1,$$

where $z_k$ is the "genetic" (not physical) distance between the $k$th SNP and the $(k+1)$th SNP in the chromosome, and $w_{ij}$ is the positive weighting parameter in the approximation of the dissimilarity $l_{ij}$. Here we consider $\sum_{k=j}^{i-1} z_k$ is the scaled distance between the $j$th SNP and the $i$th SNP ($i > j$), and such distance should be close to the dissimilarity between the $j$th SNP and the $i$th SNP.

Here there are two remarks for the above constrained scaling model. (i) It is clear that the solution of (1) can be formulated as the solution of a quadratic programming problem: Such quadratic programming problems can be solved efficiently by interior point methods [7]. (ii) The above constrained unidimensional scaling model can also be applied to the case $l_{ij} = r_{ij}$ where $r_{ij}$ is the association between the $i$th SNP and the $j$th SNP.

# 3 The Optimized Identification Method

Our approach is to develop an optimization approach to identify a position in a region where the corresponding SNP is associated with the disease locus optimally in the scaled distances among SNPs. First, we define the objective function $\mathscr{J}(x)$ for each SNP $x$ for the chromosome region concerned where

$$\mathscr{J}(x) = \sum_{i=1}^{n} w_{ix}(l_{ix} - |x - \sum_{k=1}^{i-1} z_k|)^2 \qquad (2)$$

when $x$ is the order of SNPs and $z_k$ is the scaled distances obtained by (1). In the model, we would like to pick one SNP out as the disease locus and then try to apply the objective function $\mathscr{J}(x)$ to locate the disease locus. The minimum position of $\mathscr{J}(x)$ tells us the possible position of the missing locus. Biologically, we would like to determine a position in the chromosome where the disease locus is compromised with the scaled distances among the SNPs in the chromosome.

In order to use the model, we need to input the data $l_{ix}$ (the LD distance between the $i$th SNP and the disease locus). Case-control studies are one of the most useful and prevalent method in mapping disease loci. In the following discussion, we assume the hereditary disease to be a recessive disease. In other words, it is a single gene disorder that occurs when both copies of a gene must be malfunctioned. Here we make use of case-control data to infer the linkage disequilibrium between the SNP and the disease locus. This can be done by replacing Alleles $B$ and $b$ with Case and Control in the table of Section 2. After we obtain $d$, we can calculate $d'$ and the corresponding LD distance between $i$th SNP and the disease locus, i.e., $l_{ix}$.

# 4 Simulation Study

In this section, we perform several simulation examples to test the proposed model. We remark that the MATLAB program on CPU Intel 3.2c with 1G memory is used to solve the optimization problem in (1). The program takes a file of $d'$ values produced, for example, by HAPLOVIEW (Barrett et al., 2004). The weighting parameter $w_{ij}$ in (1) is one of the square of the length of the 95% confidence interval of $l_{ij}$ obtained by transforming the 95% confidence limits of $d'_{ij}$ provided in output of HAPLOVIEW. We note that if the length is large, the weighting parameter is small and therefore the importance of such $d'$ contributes to the scaled distance is small.

## 4.1 Experiment 1

In the first experiment, we download a data set of chromosome 9 ENCODE data of CHB people from Hapmap. It starts at the chromosome position 127063383 and ends at the position 127451913, there are totally 400 SNPs selected. As a common practice, we exclude those SNPs with minor allele frequency is smaller than 0.05. In Figure 1, we show the scaled distance LD map of this region. According to the scaled distance, we know that hot-spot regions are located around the big jump of the LD curve. Therefore, we select SNP positions 48 and 132 to be our testing subjects (missing locus) where the SNP position 48 is from the hot-spot region and the SNP position 132 is in the cold-spot region. We remark that the chance of the recombination is usually higher in the hot-spot region, while the chance is lower in the cold-spot region.

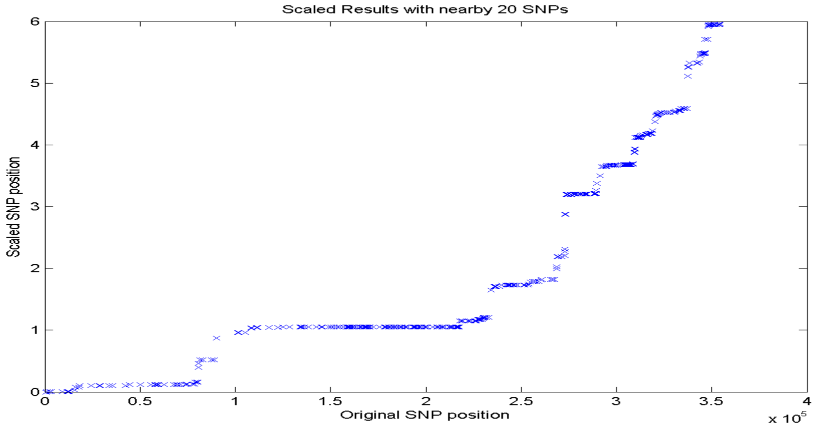
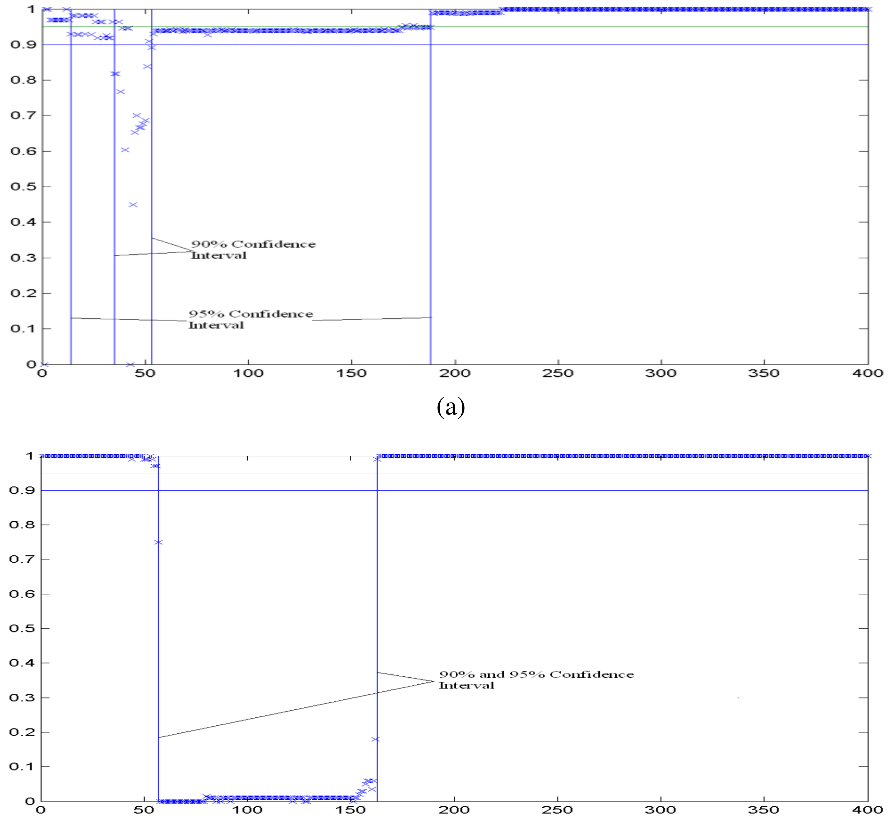


Figure 1: The scaled distance LD map of the testing region.

The simulation is carried out in the following way. The Hapmap data of CHB are the genotypes of unrelated individuals. Thus, we can prepare a bootstrap replicate of same size by simply drawing a random sample with replacement, see [6]. Therefore, we can perform a bootstrap to estimate a confidence interval to our point estimate. We compute the objective function $\mathscr{J}(x)$ at every SNP position and finally record the score of each SNP. After 100 bootstrapping, we compare the $\mathscr{J}(x)$ of all SNPs with the predicted position in the original data set. We calculate the relative frequency $f(x)$ that the score of a SNP is bigger than that of the predicted position in 100 bootstrap replicates. Assume $y$ to be the predicted position by the original data set. Then $x$ is in the $\alpha\%$ confidence interval if $f(x) < \alpha\%$. This is how we construct the confidence interval for the predicted position $y$.

In Figure 2, we find some numerical results on our described method. The confidence interval of SNP position 48 are quite narrow . For the SNP position 132, the confidence interval is wider. We suspect that this phenomena can be explained by their locations in hot-spot and cold-spot regions respectively.



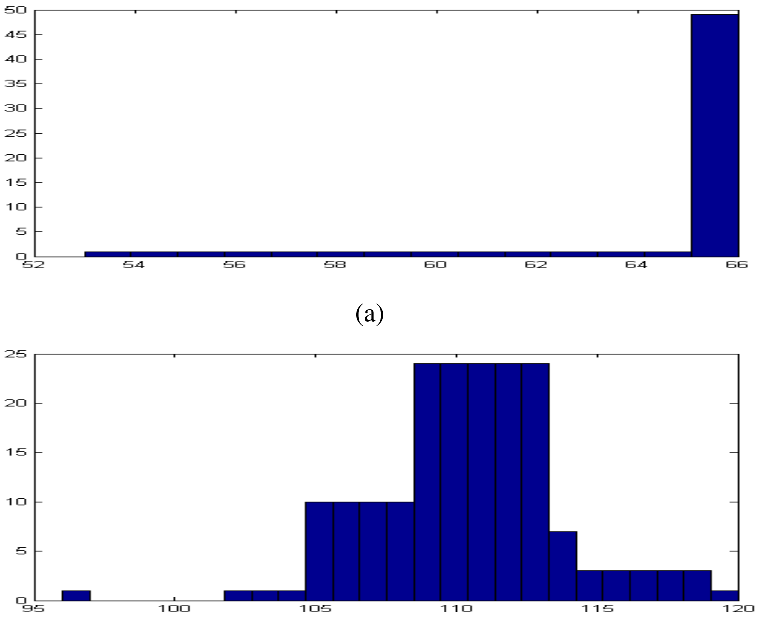
(a)

(b)

Figure 2: The predicted positions and their confidence intervals (a) SNP position 48 and (b) SNP position 132. Here $x$-axis is the SNP position and $y$-axis is the relative frequency.

## 4.2 Experiment 2

Suppose the disease is incomplete penetrance, that is to say, some carriers of the disease allele do not express a particular phenotype (disease).



(a)

(b)

Figure 3: The histogram for the predicted positions (a) SNP position 48 and (b) SNP position 132. Here $x$-axis is the SNP position and $y$-axis is the frequency.

In this subsection, we consider the following setting of the simulated data of case and control. Let $q_i$ and $p_i = 1 - q_i$ be the relative frequency of the major allele and minor allele of $i$-th locus. The parameter of mutation is set to be 0.03, that is, the probability of having a single mutation among 400 SNPs is 0.03 (here we use the same SNP data set in Experiment 1). The genotype probabilities are set as follows:

$$Prob(D|22) = 0.001; \quad Prob(D|12) = 0.01; \quad Prob(D|11) = 0.95;$$

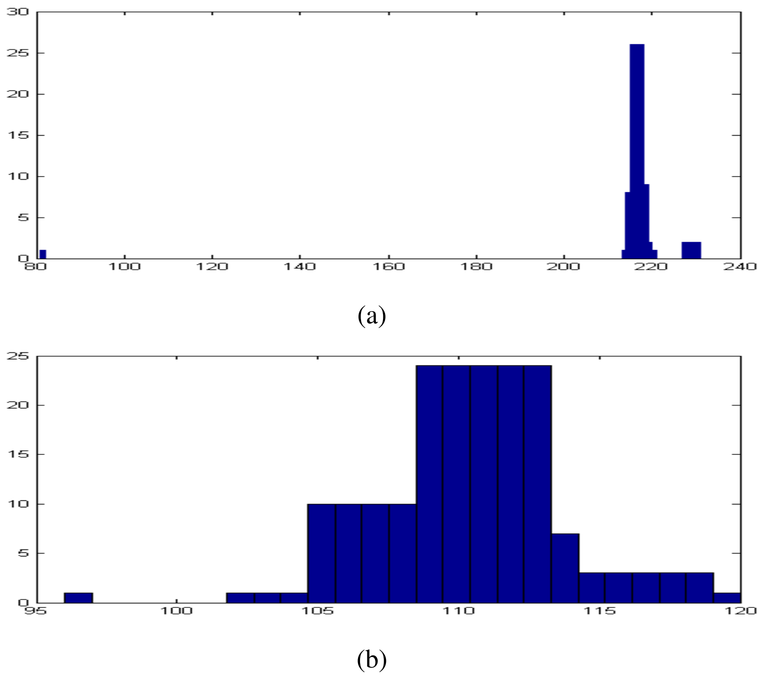
$$Prob(\overline{D}|22) = 0.999; \quad Prob(\overline{D}|12) = 0.99; \quad Prob(\overline{D}|11) = 0.05.$$

We can also set $K = {q_i}^2 P(D|11) + 2q_i p_i P(D|12) + {p_i}^2 P(D|22)$ to be the prevalence of the disease. By simple calculations, we know

$$Prob(11|D) = 0.7845; \quad Prob(12|D) = 0.1486; \quad Prob(22|D) = 0.06689;$$

and other probabilities related to $\overline{D}$, so we assign the genotype on the disease locus based on this probability. We build a pesudo-sample with $n$ controls and $N - n$ cases. Therefore, we pick $Prob(11|\overline{D}) * (n)$ people with genotype 11 and label as control, and we do similar things on the other genotypes and the clinical status. In the simulation, we also assume Hardy-Weinburg equilibrium is followed in the control group but it is not followed in the case group.

Let us test the SNP position 48 or 132 to be the disease locus position. In Figure 3, we show their results of the predicted positions for 50 trials. We remark that there may be more than one predicted positions in each trial, therefore the total count may be more than 50. We see from the figure that we cannot detect exactly disease locus position, but the proposed method still can detect nearby SNP positions. Here the total number of predicted positions are 62 and 187 for the testing SNP positions 48 and 132 respectively. In addition, we also test the other SNP positions (215 and 236). In Figure 4, we show their results of the predicted positions for 50 trials (the total numbers of predicted positions are 108 and 176 and for the SNP positions 215 and 236 respectively). We see from the figure the predicted positions are very close to the testing disease locus positions.



(a)

(b)

Figure 4: The histogram for the predicted positions (a) SNP position 48 and (b) SNP position 132. Here $x$-axis is the SNP position and $y$-axis is the frequency.

## 5 Concluding Remarks

In this paper, we have studied how to use an optimization approach to identify a position in a region where the corresponding SNP is associated with the disease locus optimally in the scaled distances among SNPs. The optimization model involves the construction of genetic mappings with the linkage disequilibrium among SNPs. Experimental results are also reported to show the effectiveness of the method. In the future, we study the following problems. Our aim is to understand the reason why the method can or cannot detect the disease locus position. We suspect that it may be related to genetic recombination of SNP regions and functional information of SNPs. On the other hand, an detailed biological analysis of some disease SNPs data sets will be studied using our method.

# References

[1] A. Brookes, The essence of SNPs, *Gene*, 234, 177-186, 1999.

[2] M. Cargill, D. Altshuler, J. Ireland and P. Sklar, Characterization of single nucleotide polymorphisms in coding regions of human genes, *Nature Genet.*, 22, 231-238, 1999.

[3] F. Collins, L. Brooks and A. Chakravarti, A DNA polymorphism discovery resource for research on human genetic variation, *Genome Research*, 8, 1229-1231, 1998.

[4] A. Collins and N. Morton, Mapping a disease locus by allelic association, *Proc. Natl. Acad. Sci. U.S.A.*, 95, 1741-1745, 1998.

[5] B. Devlin and N. Risch, A comparison of linkage disequilibrium measures for fine-scale mapping, *Genomics*, 29, 311-322, 1995.

[6] B. Efron and R. Tibshirani, *An introduction to the bootstrap*, Chapman & Hall, New York, 1993.

[7] S. Fang, and S. Puthenpura, *Linear Optimization and Extension*, Prentice-Hall, London, 1993.

[8] J. Lam, K. Roeder and B. Devlin, Haplotype fine mapping by evolutionary trees, *Am. J. Hum. Genet.*, vol. 66, 659-673, 2000.

[9] H. Liao, M. Ng, E. Fung, P. Sham P, Unidimensional nonnegative scaling for genome-wide linkage disequilibrium maps, *International Journal of Bioinformatics Research and Applications*, 4,

[10] N. Maniatis, A. Collins, C. Xu, L. McCarthy, D. Hewett, W. Tapper, S. Ennis, X. Ke and N. Morton, The First Linkage Disequilibrium (LD) Maps: DeLineation of Hot and Cold Blocks of Diplotype Analysis, *Proc. Natl. Acad. Sci.*, U.S.A., 99, 2228-2233, 2002.

[11] N. Maniatis, A. Collins, J. Gibson, W. Zhang, W. Tapper and N. Morton, Positional Cloning by Linkage Disequilibrium, *Am. J. Hum. Genet.*, 74, 846-855, 2004.

[12] M. McPeek and A. Strahs, Assessment of linkage disequilibrium by the decay of haplotype sharing with application to fine-scale genetic mapping, *Am. J. Hum. Genet.*, 65, 858-875, 1999.

[13] N. Risch and K. Merikangas, The future of genetic studies of complex human diseases, *Science*, 273, 1516-1517, 1996.

[14] P. Sham, *Statistics in Human Genetics*, Edward Amold, 1998.

[15] P. Sham, J. Zhao and D. Curtis, The effect of marker characteristics on the power to detect linkage disequilibrium due to single or multiple ancestral mutations, *Ann. Hum, Genet.*, 64, 161-169, 2000.

[16] M. Xiong and L. Jin, Comparison of the power and accuracy of biallelic and microsatellite markers in population-based gene-mapping methods, *Am. J. Hum. Genet.*, 64, 629-640, 1999.