Loci ordering via the Fiedler vector

CHOONGRAK KIM
Department of Statistics, Pusan National University, Pusan, 609-735,

SUMMARY

Locus ordering is the necessary step in constructing genetic map, and the construction of reliable and fine genetic map is one of the most important issue in genetic research area. Locus ordering searches for the best locus order among the possible orders and it amounts to evaluating the maximum likelihood for each order. With only 10 loci, for example, there are $1,814,000$ possible orders, and therefore, locus ordering entails a big computational problem. In this paper we suggest a useful algorithm for loci ordering via the Fiedler vector. The suggested algorithm is easy to compute and can handle many loci simultaneously. Furthermore, the required computation time is very short compared to others and the result of locus ordering is very accurate.

1. INTRODUCTION

Locus ordering is the necessary step in constructing genetic map, and the construction of reliable and fine genetic map is one of the most important issue in genetic research area. A locus is the chromosome location of a gene or any specific DNA sequence, and can be regarded as a point on a line. Locus ordering is a linear arrangement of genes or genetic markers in a linkage group which is a group of genes with their loci located on the same chromosome.

Let $l_1, l_2, \cdots, l_n$ denote n loci in a linkage group, then our interest is ordering n loci to construct genomic map based on two-point recombination fractions r_{ij}, $1 \leq i < j \leq n$ for a pair of loci i and j, which usually ranges from 0 to 1. If two loci i and j are closely located, i.e., tightly linked, then r_{ij} is close to 0, and if not it is away from 0. With n loci, there are $n!/2$ possible orderings if the orientation of the orders is ignored. Locus ordering searches for the best locus order among the possible orders and it amounts to evaluating the maximum likelihood for each order. With only 10 loci, for example, there are $1,814,000$ possible orders, and therefore, locus ordering entails a big computational problem. To avoid this problem, a locus ordering minimizing the number of crossovers was regarded as the best ordering, and this method has been shown to be the maximum likelihood ordering under the full penetrance assumption, i.e. lack of interference, by Thompson(1984). Also, several other approaches are closely related with this idea. Among them, Falk(1989) suggested the minimum sum of adjacent recombination fractions criterion, Weeks and Lange(1987) suggested the maximum sum of adjacent lod scores criterion, minimum sum of the probability of double recombinants (Knapp *et al.* 1989), maximum likelihood (Lander and Green 1987), minimum obligatory crossovers(Thompson 1989). Comparisons among those methods were done by Kammerer and MacCluer(1988) and Olson and Boehnke(1990) when the number of loci is 6 or 7. For details, see Weeks(1991) for an overview in locus ordering.

In this paper we suggest a useful algorithm for loci ordering via the Fiedler vector(Fiedler 1973, 1975). The suggested algorithm is easy to compute and can handle many loci simultaneously. Furthermore, the required computation time is very short compared to others and the result of locus ordering is very accurate. In §2, we introduce basic definitions of graph theory including the Laplacian matrix and the Fiedler vector, and give motivation on locus ordering. In §3, a locus ordering algorithm is suggested. An illustrative example based on the barley data is given in §4.

2. THE FIEDLER VECTOR

2.1 *Laplacian matrix and Fiedler vector*

A graph $G = G(V, E)$ consists of a set of vertices V and a set of edges E. Two vertices v_i and v_j of a graph G are said to be adjacent if there exists an edge e_{ij} connecting v_i and v_j. The degree of a vertex v_i is defined as the number of adjacent vertices to v_i. The adjacency matrix $A = A(G)$ of a graph G with n vertices is defined as an $n \times n$ symmetric matrix with components a_{ij}, where the diagonal elements a_{ii} are equal to zero for all $i = 1, 2, ..., n$. The Laplacian matrix of a graph G is defined as $L(G) = D(G) - A(G)$, where $D(G)$, called the degree matrix, is a diagonal matrix with the ith diagonal element $d_i = \sum_{j=1}^{n} a_{ij}$. Note that the Laplacian matrix is symmetric and positive semidefinite. If the graph is connected, then the rank of the Laplacian matrix is $n - 1$, so that the smallest eigenvalue of $L = L(G)$ is zero with constant eigenvector and all other eigenvalues are positive. Let $0 = \lambda_1 < \lambda_2 < \cdots < \lambda_n$ be n eigenvalues of L in an increasing order, then an eigenvector corresponding to λ_2, the nonzero smallest eigenvalue, is called the Fiedler vector and an eigenvector corresponding to λ_n is called the Frobenius vector.

2.2 *Motivation on locus ordering*

Here we describe how to use the Fiedler vector to locus ordering based on estimates of recombination fraction r_{ij} between two loci i and j, $1 \leq i < j \leq n$. To use the Fiedler vector to locus ordering let the adjacency between two loci i and j be $a_{ij} = 1 - r_{ij}$ because if two genes are closely located then a_{ij} will be large. Also, let $\mathbf{z} = (z_1, ..., z_n)'$, where z_i denotes the relative order of the ith gene among n genes. Therefore, estimating the order of n loci for each gene corresponds to finding $\mathbf{z}$. Then, this goal can be achieved by minimizing the weighted sum of squares

$$Q = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(z_i - z_j)^2 a_{ij}.$$

To avoid the trivial solution $z_i = 0$ for all i, the constraint $\mathbf{z}'\mathbf{z} = 1$ is imposed. Also, the constraint $\mathbf{z}'\mathbf{1} = 0$, where $\mathbf{1} = (1, \cdots, 1)'$, is imposed since the minimum is invariant under translations. Therefore the problem can be rewritten as

$$\arg\min_{\mathbf{z}} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(z_i - z_j)^2 a_{ij} \quad \text{subject to } \mathbf{z}'\mathbf{z} = 1 \text{ and } \mathbf{z}'\mathbf{1} = 0.$$

To solve the problem, note that

$$\begin{aligned} Q &= \tfrac{1}{2}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{n}(z_i^2 - 2z_iz_j + z_j^2)a_{ij} \\ &= \textstyle\sum_{i=1}^{n} z_i^2 d_i - \sum_{j=1}^{n}\sum_{i\neq j}^{n} z_iz_ja_{ij} \\ &= \mathbf{z}'\mathbf{L}\mathbf{z} \end{aligned}$$

To minimize Q subject to $\mathbf{z}'\mathbf{z} = 1$, use Lagrangian method, i.e., for a Lagrangian multiplier λ,

$$\begin{aligned} T &= \mathbf{z}'\mathbf{L}\mathbf{z} - \lambda(\mathbf{z}'\mathbf{z} - 1) \\ \frac{\partial T}{\partial \mathbf{z}} &= 2\mathbf{L}\mathbf{z} - 2\lambda\mathbf{z} = \mathbf{0} \\ \Rightarrow & \quad (\mathbf{L} - \lambda\mathbf{I})\mathbf{z} = \mathbf{0} \end{aligned}$$

which yields a nontrivial solution $\mathbf{z}$ if and only if λ is an eigenvalue of $\mathbf{L}$ and $\mathbf{z}$ is the corresponding eigenvector. By multiplying $\mathbf{z}'$ on both sides, we have

$$\mathbf{z}'\mathbf{L}\mathbf{z} = \lambda$$

Therefore, the nonzero smallest eigenvalue and the associated eigenvector, which is just the Fiedler vector, yields the optimal solution.

3. LOCUS ORDERING ALGORITHM

3.1 *Hard and soft-thresholding*

As is well known each observation is contaminated by noise such that the observed value can be regarded as a sum of the deterministic part and the noise part. Therefore, to obtain better trend for given data sets we need to remove the noise part, and this step is also necessary in locus ordering. As a method of shrinkage, we use the hard- and soft-thresholding functions given by $t_H(x) = xI(|x| > \delta)$ and $t_S(x) = sgn(x)(|x| - \delta)_+$, respectively. Note that $\delta > 0$ is a thresholding parameter to be estimated. There are many methods of estimating the thresholding parameter, and we will discuss them in §3.2.

Let $\mathbf{t} = (t(1), \cdots, t(n))$ and $\mathbf{e} = (e(1), \cdots, e(n))$, where $t(i)$ denote the true order of the ith gene and $e(i)$ denote the estimated order of the ith gene, respectively. The accuracy of the estimated order $\mathbf{e}$ can be measured by

$$NCL = \sum_{i=1}^{n} I(t(i) = e(i))$$

and

$$PIL = \sum_{i=1}^{n} |(t(i) - e(i))|,$$

where NCL(number of correct loci) denotes the number of correctly ordered loci and PIC(penalized incorrect loci) denotes the number of incorrectly ordered loci with the discrepancy as penalty. Therefore, good ordering should have both the large value of NCL and the small value of PIL.

3.2 *Estimation of thresholding parameter*

Let $\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n$ be corresponding eigenvectors of eigenvalues $0 = \lambda_1 < \lambda_2 < \cdots < \lambda_n$ of the Laplacian matrix L, respectively. Note that $\mathbf{z}_2$ is the Fiedler vector and it contains the ordering information on n loci. To choose the thresholding parameter δ, note that as δ increases λ_2 decreases, and we get more information as λ_2 becomes smaller. However, if δ is too large, then λ_2 will be zero so that we lose all the information. To solve this contradicting situation, we consider λ_3. Since the eigenvectors $\mathbf{z}_2$ and $\mathbf{z}_3$ are orthogonal,the information contained in $\mathbf{z}_3$ is independent on the information contained in $\mathbf{z}_2$. Therefore, it is desirable that λ_2 must be relatively small compared to λ_3. Then, most of locus ordering information is contained in the corresponding eigenvector $\mathbf{z}_2$. Hence, to estimate the thresholding parameter δ, we propose to choose δ which maximizes

$$\Lambda = \lambda_3/\lambda_2.$$

4. EXAMPLE

As an illustrative example, we consider 26 loci of barley chromosome IV generated by the North American Barley genome Mapping Project (NABGMP). Based on the several preliminary locus ordering method it has been known that the best ordering for the 26 loci is $1, 2, \cdots, 26$, where each number represents the relative locus of each gene. By applying the proposed method in §3.2, we obtained $\delta = 0.58$ as a thresholding value. Based on this value, we get the exactly the same order as the result given by the existing methods.

REFERENCES

Falk, C. T. (1989). A simple scheme for preliminary ordering of multiple loci: application to 45 CF families. In*Multipoint mapping and linkage based upon affected pedigree members* (Elston, Spence, Hodge and MacCluer eds), 17-22. Genetic Workshop 6. Liss, New York,

Fiedler, M. (1973). Algebraic connectivity of graphs, *Czech. Math. J.* **23**, 298-305.

Fiedler, M. (1975). A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory, *Czech. Math. J.* **23**, 298-305.

Kammerer, C. M. & MacCluer, J. W. (1988). Empirical power of three preliminary methods for ordering loci. *Am. J. Hum. Genet.* **43**, 964-970.

Olson, J. M. & Boehnke, M. (1990). Monte Carlo comparison of preliminary methods for ordering multiple genetic loci. *Am. J. Hum. Genet.* **47**, 470-482.

Weeks, D. E.(1991). Human linkage analysis: Strategies for locus ordering. In *Advanced Techniques in Chromosomes Research* (K.W. Adolph ed.), 297-330. Marcel Dekker, New York.

Weeks, D. E. & Lange, K. (1987). Preliminary ranking procedures for multilocus ordering. *Genomics* **1**, 236-242.