2

Kullback-Leibler Information for Ordering Genes Using Sperm Typing and Radiation Hybrid Mapping

Herman Chernoff

Harvard University
Mathematical Sciences Research Institute

Technical Report No. ONR-C-9

October, 1991

Kullback-Leibler Information for Ordering Genes Using Sperm Typing and Radiation Hybrid Mapping

Herman Chernoff

Harvard University
Mathematical Sciences Research Institute

ABSTRACT

Two technologies applicable to gene mapping are those of sperm typing and radiation hybrid mapping. Sperm typing makes use of the polymerase chain reaction, a biochemical technique which alows enormous amplification (production of multiple copies) of small, selected DNA fragments from a single chromosome. A sample of sperm from a single donor is analyzed to see which alleles (distinct forms of the various genes) are present in the individual sperms. The frequencies with which the various possibilities occur can be used to supply estimates of the ordering and of the recombination probabilities among the genes for which that donor is heterozygous (having different alleles of the same gene.) Radiation hybrid mapping employs a different technology where hybrid mouse cells containing a human chromosome are subjected to a dose of radiation, which leads to breaking the chromosome into segments, a fraction of which are retained in succeeding generations. The simultaneous presence or absence of various genes provides indirect information on how close together these genes are, and also on the ordering of these genes.

For each of these methods, the analysis grows in complexity as the number of genes being considered increases. At the same time the accuracy of the probabiliistic models used in the analysis becomes more questionable. On the other hand the ability to determine the order of three genes may be enhanced by the inclusion, in the analysis, of the data on nearby genes. For both of these methods, Kullback-Leibler information numbers are derived to test hypotheses involving the order of m genes. These information numbers are computed for testing hypotheses concerning the ordering of three genes with and without considering the presence of data involving other nearby genes.

AMS 1980 *subject classifications.* Primary 62B10, 92D20.
Key Words and Phrases. Kullback-Leibler Information, Sperm Typing, Radiation Hybrid Mapping, Gene Ordering.

Kullback-Leibler Information for Ordering Genes Using Sperm Typing and Radiation Hybrid Mapping

Herman Chernoff

Harvard University
Mathematical Sciences Research Institute

1 Introduction

Two technologies applicable to gene mapping are those of sperm typing and radiation hybrid mapping. Sperm typing makes use of the polymerase chain reaction, a biochemical technique which alows enormous amplification (production of multiple copies) of small, selected DNA fragments from a single chromosome. A sample of sperm from a single donor is analyzed to see which alleles (distinct forms of the various genes) are present in the individual sperms. The frequencies with which the various possibilities occur can be used to supply estimates of the ordering and of the recombination probabilities among the genes for which that donor is heterozygous (having different alleles of the same gene.) Radiation hybrid mapping employs a different technology where hybrid mouse cells containing a human chromosome are subjected to a dose of radiation, which leads to breaking the chromosome into segments, a fraction of which are retained in succeeding generations. The simultaneous presence or absence of various genes provides indirect information on how close together these genes are, and also on the ordering of these genes.

For each of these methods, the analysis grows in complexity as the number of genes being considered increases. At the same time the accuracy of the probabiliistic models used in the analysis becomes more questionable. On the other hand the ability to determine the order of three genes may be enhanced by the inclusion, in the analysis, of the data on nearby genes. For both of these methods, we shall examine the relevant Kullback-Leibler information numbers for hypotheses concerning the ordering of three genes with and without considering the presence of data involving other nearby genes.

In Section 2 we introduce the model for sperm typing and discuss the maximum likelihood estimates of the recombination probabilities.

In Section 3 we derive expressions for the relevant Kullback-Leibler informations for sperm typing. In Section 4 we describe the model for radiation hybrids and derive the corresponding information numbers. The outcome of the calculations is described in Section 5. We terminate this introduction with a brief discussion of the Kullback-Leibler (KL) information.

Given two simple hypotheses concerning the (density) distribution $f(x)$ of the data X, $H_0 : f(x) = f_0(x)$ and $H_1 : f(x) = f_1(x)$ the KL information for discriminating between H_0 and H_1 is

$$K(f_0, f_1) = E_{f_0}\{\log[f_0(X)/f_1(X)]\}. \tag{1}$$

The subscript f_0 refers to the fact that the expectation is calculated for the case where the distribution of X is governed by f_0. The information K measures the exponential rate at which the posterior probability of H_1 approaches zero when H_0 is true, as independent observations on X are obtained. It is particularly relevant in the design of sequential experiments, such as were discussed by Goradia and Lange (1990). Suppose now that under our model the density of X can be described in terms of a parameter θ, i.e. $f(x) = f(x, \theta)$, and the underlying probability distribution is governed by $\theta = \theta_0$, and we are interested in a composite alternative $H_1 : \theta \in \Omega_1$ to the true hypothesis $H_0 : \theta = \theta_0$. Then the appropriate measure is

$$K(H_0, H_1) = \inf_{\theta_1 \in \Omega_1} E_{\theta_0}\{\log[f(X, \theta_0)/f(X, \theta_1)]\} \tag{2}$$

which can be decomposed into the following difference if either term is finite

$$K(H_0, H_1) = E_{\theta_0}\{\log[f(X, \theta_0))\} - \sup_{\theta_1 \in \Omega_1} E_{\theta_0}\{\log f(X, \theta_1))\}.$$

We shall suppress the subscript θ_0 when there is no danger of ambiguity.

2 The sperm typing model and maximum likelihood

Consider first the case of three genes for which the donor is heterozygous, and his two chromosomes have genes ABC and abc respectively.

A sperm will have a chromosome providing one of the 8 following observations, $ABC, ABc, AbC, Abc, aBC, aBc, abC, abc$ with probabilities depending on the recombination probabilities and the ordering of the three genes on the chromosome. Suppose that the genes appeared in the order ABC rather than ACB or BAC. Suppose also that the recombination probabilities (indicating the probabilities that in the reproduction process, the chromosomes would seperate and recombine) between A and B is ϕ_{ab} and between B and C is ϕ_{bc}. Finally suppose that the recombination events are independent. Then the probabilities associated with ABC, abc, and AbC, would be $(1-\phi_{ab})(1-\phi_{bc})/2$, $(1-\phi_{ab})(1-\phi_{bc})/2$, and $\phi_{ab}\phi_{bc}/2$ respectively. The probabilities associated with the other 5 events can be calculated similarly.

While the estimation of ϕ_{ab} and ϕ_{bc} are of interest and relevant, our main focus in the next section will be on deciding which is the correct one of the three possible orderings ABC, ACB, BAC. Note that without reference to other parts of the chromosome the orderings ABC and CBA are equivalent and we need consider only three, or half of the six possible permutations of ABC. It is also evident that the relevant information in the observed categories ABC and abc are equivalent, and thus we may combine these two observations into one equivalent one, $\bar{A}\bar{B}\bar{C}$ with probability $(1-\phi_{ab})(1-\phi_{bc})$ under the ordering ABC, and probability $(1-\phi^{*}_{ac})(1-\phi^{*}_{bc})$ under the ordering ACB, and probability $(1-\phi^{**}_{ab})(1-\phi^{**}_{ac})$ under the ordering BAC. Thus we need only consider 4 possible observations, e.g. $\bar{A}\bar{B}\bar{C}$, $\bar{A}\bar{B}\bar{c}$, $\bar{A}\bar{b}\bar{C}$, and $\bar{A}\bar{b}\bar{c}$, each representing a pair of the original 8 categories.

In our analysis it would seem important to bear in mind that the statistician does not know which alleles appear on the original chromosomes. Thus, even with the order ABC, it might be that the original chromosomes of the donor have AbC and aBc. For our problem involving relatively small recombination probabilities, the data would quickly and easily determine the form of the chromosome, for an original chromosome with ABC would lead to a great preponderance of the $\bar{A}\bar{B}\bar{C}$ observations independent of the order. Nevertheless it turns out that symmetry aspects of the analysis make it unimportant to hypothesize or estimate which alleles appear on each chromosome.

Goradia and Lange (1990) analyze two sequential methods of selecting the correct order. They do not analyze the sequential probability ratio method, since the two approaches that they use are much easier for them to analyze. One may wonder whether there is a substantial loss of efficiency in using their methods. The related question that we address is whether there would be an increase in the efficiency of deciding the order of ABC if the analysis were extended to include 4 or 5 genes. Several complications arise in the use of KL numbers to address this question. One is that in ordering 4 (or 5) genes, there are $12 = 4!/2$ (or $60 = 5!/2$) possible orderings of concern. Another issue is that it is more difficult to find the donor who is heterozygous on four, rather than three, specified genes. Finally, technical problems in the technology may make the simple extension of the above probability model less reliable in the application to four or more genes.

In any case, when the KL numbers indicate that there is little to be gained by introducing 4 or 5 genes, then it makes sense to confine attention to three at a time. In case there is a potential gain of a great amount of information, then one ought to consider the relative merit of doing the possibly more complicated analysis required to deal with more than 3 genes.

Assuming the order ABC, the likelihood, based on n_{ABC},n_{ABc},n_{AbC}, and n_{Abc} observations $\bar{A}\bar{B}\bar{C}$,$\bar{A}\bar{B}\bar{c}$, $\bar{A}\bar{b}\bar{C}$,$\bar{A}\bar{b}\bar{c}$ respectively, is

$$\begin{aligned} L &= [(1-\phi_{ab})(1-\phi_{bc})]^{n_{ABC}}[(1-\phi_{ab})\phi_{bc}]^{n_{ABc}}[\phi_{ab}\phi_{bc}]^{n_{AbC}} \\ &\qquad \cdot[\phi_{ab}(1-\phi_{bc})]^{n_{Abc}} \\ &= \phi_{ab}^{n_{Ab}}(1-\phi_{ab})^{n_{AB}}\phi_{bc}^{n_{Bc}}(1-\phi_{bc})^{n_{BC}} \end{aligned}$$

where

$$n_{AB} = n_{ABC} + n_{ABc} = n - n_{Ab}$$

$$n_{BC} = n_{ABC} + n_{Abc} = n - n_{Bc}$$

and

$$n = n_{ABC} + n_{ABc} + n_{AbC} + n_{Abc}$$

is the total number of observations. The corresponding maximum likelihood estimates are

$$\hat{\phi}_{ab} = n_{Ab}/n$$

and

$$\hat{\phi}_{bc} = n_{Bc}/n$$

yielding the likelihood

$$L(ABC) = \{\hat{\phi}_{ab}^{\hat{\phi}_{ab}}(1-\hat{\phi}_{ab})^{(1-\hat{\phi}_{ab})}\hat{\phi}_{bc}^{\hat{\phi}_{bc}}(1-\hat{\phi}_{bc})^{(1-\hat{\phi}_{bc})}\}^{n}$$

with logarithm

$$\log L(ABC) = -n\{V(\hat{\phi}_{ab}) + V(\hat{\phi}_{bc})\} \tag{3}$$

where, for $0 < x < 1$,

$$V(x) = -\{x \log x + (1-x)\log(1-x)\} \tag{4}$$

is an entropy.

The likelihood corresponds to that calculated from observing the two sets of independent binomials corresponding to the recombinations from A to B and from B to C. Notice that if the original chromosomes had AbC and aBc, the estimates of $\hat{\phi}_{ab}$ and $\hat{\phi}_{bc}$ would be replaced by the complements $1-\hat{\phi}_{ab}$ and $1-\hat{\phi}_{bc}$ and $\log L$ would be unaltered.

These results help to understand the derivation of the KL numbers in the following section where we deal with expected log likelihoods for $n = 1$.

In generalizing to m genes, we could extend the alphabetic notation, but it seems more convenient to change the notation slightly. We label the genes 1 to m and consider those permutations, $\pi = (\pi_1, \pi_2, \ldots, \pi_m)$, for which 1 appears in the first half, or, in the case where m is odd, may appear in the center, but 2 appears in the first half. Thus, for $m = 3$, we have the permutations $(123, 132, 213)$ representing the 3 possible orderings.

A possible parametric point θ is described by a permutation π and a vector ϕ with components $\phi_{\pi_i,\pi_{i+1}}$ for $1 \leq i \leq m-1$, representing recombination probabilities. For the time being this notation seems mildly ambiguous since ϕ_{12} associated with $\pi^0 = (1,2,3,4,5)$ and ϕ_{12} associated with $\pi^1 = (1,2,5,4,3)$ should be designated separately, possibly with superscripts. Our observations will consist of n independent vectors of the form $X = (X_1, X_2, \ldots, X_m)$, where the i-th component of X is zero or one depending on which allele of the i-th gene is observed in the given sperm observation.

Supposing that the true ordering is π^0, and given the associated values of $\phi = (\phi_{12}, \phi_{23}, \ldots, \phi_{m-1,m})$, the likelihood is easily seen to be

$$L = \prod_{i=1}^{m-1} [\phi_{i,i+1}^{n_{i,i+1}} (1 - \phi_{i,i+1})^{n - n_{i,i+1}}] \tag{5}$$

where n_{ij} is the number of times that $X_i \neq X_j$ in the sample of n observations. Then the maximum likelihood estimates are

$$\hat{\phi}_{i,i+1} = n_{i,i+1}/n, \qquad 1 \leq i \leq m - 1 \tag{6}$$

and the maximum likelihood under the ordering π^0 satisfies

$$\log L(\pi^0) = - \sum_{i=1}^{m-1} V(\hat{\phi}_{i,i+1}) \tag{7}$$

Given an alternate permutation π, it is clear that the corresponding MLE of the related recombination probabilities are given by

$$\hat{\phi}^*_{\pi_i \pi_{i+1}} = n_{\pi_i \pi_{i+1}}/n, \qquad 1 \leq i \leq m - 1 \tag{8}$$

and the maximum likelihood satisfies

$$\log L(\pi) = - \sum_{i=1}^{m-1} V(\hat{\phi}^*_{\pi_i \pi_{i+1}}) \tag{9}$$

Note that if $\pi = (1, 2, 5, 4, 3)$, the MLE of ϕ^*_{12} under the ordering π is exactly the same as that of ϕ_{12} under π^0. Also, under the hypothesis $H_0 : \theta = \theta_0 = (\pi^0, \phi)$ where $\phi = (\phi_{12}, \phi_{23}, \ldots, \phi_{m-1,m})$ is specified, the variables n_{ij} are binomial random variables associated with probabilities

$$\phi_{ij} = P\{X_i \neq X_j\}, \qquad 1 \leq i, j \leq m \tag{10}$$

Here, and later, we assume that H_0 applies and suppress the subscript θ_0 for P and E. Then ϕ_{ij} is the probability of an odd number of recombinations between the i-th and j-th genes. Thus $\phi_{ii} = 0$, $\phi_{i,i+1} = \phi_{i+1,i}$, and, for $1 \leq i < j \leq m - 1$,

$$\phi_{i,j+1} = \phi_{j+1,i} = \phi_{ij}(1 - \phi_{j,j+1}) + (1 - \phi_{ij})\phi_{j,j+1} \tag{11}$$

3 Kullback-Leibler information for sperm typing

To calculate the KL numbers, consider the case $n = 1$. Then $En_{ij} = \phi_{ij}$,

$$\begin{aligned} E \log f(X, \theta_0) &= E \sum_{i=1}^{m-1} \{n_{i,i+1} \log \phi_{i,i+1} + (1 - n_{i,i+1}) \log(1 - \phi_{i,i+1})\} \\ &= -\sum_{i=1}^{m-1} V(\phi_{i,i+1}). \end{aligned}$$

For specified $\theta = (\pi, \phi^*)$,

$$\begin{aligned} E \log f(X, \theta) &= E \sum_{i=1}^{m-1} \{n_{\pi_i\pi_{i+1}} \log(\phi^*_{\pi_i\pi_{i+1}}) \\ &\qquad + (1 - n_{\pi_i\pi_{i+1}}) \log(1 - \phi^*_{\pi_i\pi_{i+1}})\} \\ &= \sum_{i=1}^{m-1} \{\phi_{\pi_i\pi_{i+1}} \log \phi^*_{\pi_i\pi_{i+1}} \\ &\qquad + (1 - \phi_{\pi_i\pi_{i+1}}) \log(1 - \phi^*_{\pi_i\pi_{i+1}})\} \end{aligned}$$

which is maximized with respect to ϕ^* by

$$\phi^*_{\pi_i\pi_{i+1}} = \phi_{\pi_i\pi_{i+1}}$$

Thus if H_1 corresponds to the composite hypothesis of the ordering π, we would have

$$K(H_0, H_1) = \sum_{i=1}^{m-1} [V(\phi_{\pi_i\pi_{i+1}}) - V(\phi_{i,i+1})] \tag{12}$$

In particular,suppose that we are dealing with 3 genes and H_1 corresponds to the order $(1,3,2)$. Then

$$K(H_0, H_1) = V(\phi_{13}) + V(\phi_{23}) - V(\phi_{12}) - V(\phi_{23}) = V(\phi_{13}) - V(\phi_{12})$$

whereas for H_2 corresponding to $(2,1,3)$,

$$K(H_0, H_2) = V(\phi_{13}) - V(\phi_{23}).$$

Finally

$$K(H_0, H_1 \cup H_2) = V(\phi_{13}) - \max[V(\phi_{12}), V(\phi_{23})] \qquad (13)$$

More generally for m genes, let H_0 correspond to $\pi_0 = (1, 2, \ldots, m)$ and $\phi = \{\phi_{i,i+1} : 1 \leq i \leq m-1\}$, and let $\mathcal{A}$ be a subset of the $m!/2-1$ other permutations corresponding to alternate orders. Then

$$K(H_0, H_{\mathcal{A}}) = \min_{\pi \in \mathcal{A}}\{\sum_{i=1}^{m-1} V(\phi_{\pi_i \pi_{i+1}})\} - \sum_{i=1}^{m-1} V(\phi_{i,i+1}) \qquad (14)$$

We are mainly concerned with 3 cases. Given genes 1 to 5 with $\{\phi_{i,i+1} : 1 \leq i \leq 4\}$, we have

case 1: $m = 3$, $\phi = (\phi_{23}, \phi_{34})$, $\mathcal{A}$ is the set of 2 orderings of $(1, 2, 3)$ other than $(1, 2, 3)$

case 2: $m = 4$, $\phi = (\phi_{23}, \phi_{34}, \phi_{45})$, $\mathcal{A}$ is the set of 8 orderings of $(1, 2, 3, 4)$ inconsistent with the ordering $(1, 2, 3)$ or its equivalent $(3, 2, 1)$

case 3: $m = 5$, $\phi = (\phi_{12}, \phi_{23}, \phi_{34}, \phi_{45})$, $\mathcal{A}$ is the set of 40 orderings inconsistent with the ordering $(2, 3, 4)$ or $(4, 3, 2)$.

These three cases give us the relevant KL numbers for the ordering of genes 2, 3, and 4 when considering data involving (1) the three genes $(2, 3, 4)$, (2) the four genes $(2, 3, 4, 5)$, and (3) the five genes $(1, 2, 3, 4, 5)$.

4 Radiation hybrid model

Another technology for estimating distances along the chromosome and for ordering genes is that of radiation hybrid mapping. Here again we introduce the model via cases involving few genes. This model was analyzed by Boehnke *et al.* (1991) and Lange and Boehnke (1991). The technology consists of radiating a hybrid mouse cell which carries a human chromosome, thereby breaking the chromosome into several fragments, a proportion $r = 1 - \bar{r}$ of which are retained. The higher the rate of radiation, λ, the more fragments are made.

We will assume that r is known, that the distance between two genes A and B is δ, unknown. Then the probability that the two genes will be on separate fragments is

$$\phi = 1 - \exp(-\lambda\delta) \qquad (15)$$

assuming that breaks occur like a Poisson process with rate λ. The probabilities of observing, among the retained fragments, both A and B, A alone, B alone, and neither are

$$\begin{aligned} p_{11} &= r(1-\bar{r}\phi) \\ p_{10} &= r\bar{r}\phi \\ p_{01} &= p_{10} \\ p_{00} &= \bar{r}(1-r\phi) \end{aligned} \tag{16}$$

respectively. Here we have assumed that breaks and retention events are independent, and that r is constant.

It is relatively easy to calculate the Fisher Information for estimating ϕ. That is

$$\begin{aligned} J &= E\left\{\left[\frac{\partial(\log likelihood)}{\partial\phi}\right]^2\right\} \\ &= \frac{r\bar{r}(2-\phi)}{\phi(1-r\phi)(1-\bar{r}\phi)} \end{aligned} \tag{17}$$

It follows that the Fisher information with respect to the distance δ is

$$J^* = J\left(\frac{d\phi}{d\delta}\right)^2 = \frac{r\bar{r}\lambda^2(1-\phi)^2)(2-\phi)}{\phi(1-r\phi)(1-\bar{r}\phi)} \tag{18}$$

For small $\lambda\delta$, $\phi \approx \lambda\delta$ and $J^* \approx 2r\bar{r}\lambda/\delta$. Insofar as $1/(nJ^*\delta^2)$ is the asymptotic *relative variance* of the large sample estimate of δ, it gives us a clue about what values of λ would be useful for ordering the genes. Uncertainty in the knowledge of r complicates matters somewhat. In that case the information matrix for δ and r should be evaluated and inverted.

To proceed with the ordering problem, suppose that the genes are arranged in order $\pi^0 = (1, 2, \ldots, m)$ and that the distances between successive genes $\delta_{i,i+1}$, give rise to separation probabilities $\phi_{i,i+1}$. Then let the observation be a vector $X = (X_1, X_2, \ldots, X_m)$ to indicate which genes are retained. That is $X_i = 1$ indicates retention of the i-th gene and otherwise $X_i = 0$. Then X is a Markov Process where

$$f_X(x) = r^{x_1}\bar{r}^{(1-x_1)} \prod_{i=1}^{m-1} g(x_i, x_{i+1}; \phi_{i,i+1}) \tag{19}$$

and

$$\begin{aligned} g(1,1;\tau) &= 1-\bar{r}\tau \\ g(1,0;\tau) &= \bar{r}\tau \\ g(0,1;\tau) &= r\tau \\ g(0,0;\tau) &= 1-r\tau. \end{aligned} \tag{20}$$

Further, for $i < j$,

$$P\{X_j = x_j | X_i = x_i\} = g(x_i, x_j; \phi_{ij}) \tag{21}$$

where

$$\phi_{ij} = 1 - \exp(-\lambda\delta_{ij}) = \phi_{ji} \tag{22}$$

and

$$\delta_{ij} = \sum_{k=i}^{j-1} \delta_{k,k+1} \tag{23}$$

is the distance between the i-th and j-th genes.

We shall be interested in maximizing

$$w_{ij}(\tau) = E \log g(X_i, X_j; \tau)$$

with respect to τ. Then

$$\begin{aligned} w_{ij}(\tau) &= r(1-\bar{r}\phi_{ij})\log(1-\bar{r}\tau) + r\bar{r}\phi_{ij}\log(\bar{r}\tau) + \bar{r}r\phi_{ij}\log(r\tau) \\ &\quad + \bar{r}(1-r\phi_{ij})\log(1-r\tau) \end{aligned}$$

and

$$w'_{ij} = \frac{r\bar{r}(\phi_{ij}-\tau)(2-\tau)}{\tau(1-r\tau)(1-\bar{r}\tau)}$$

vanishes only at $\tau = \phi_{ij}$ in the interval $(0,1)$, and indeed, $w_{ij}(\tau)$ attains its maximum value

$$\begin{aligned} W(\phi_{ij}) &= r(1-\bar{r}\phi_{ij})\log(1-\bar{r}\phi_{ij}) + r\bar{r}\phi_{ij}\log(r\bar{r}\phi_{ij}^2) \\ &\quad + \bar{r}(1-r\phi_{ij})\log(1-r\phi_{ij}) \end{aligned} \tag{24}$$

at $\tau = \phi_{ij}$. Incidentally, this result could also be derived without calculating the derivative, by noting the relationship between w_{ij} and a Kullback-Leibler number and that a KL number is always nonnegative, and hence an expression of the form $E_{\theta_0}[\log f(x,\theta)]$ attains its maximum value when $\theta = \theta_0$.

We are now in position to calculate the KL information. Let

$H_0 : \theta = \theta_0$ correspond to the permutation π^0 and
$\phi = (\phi_{12}, \phi_{23}, \ldots, \phi_{m-1,m})$, and
$H_1 : \theta = \theta_1$ correspond to the permutation π^* and
$\phi^* = (\phi^*_{12}, \ldots, \phi^*_{m-1,m})$.

Then, with E_{θ_0} represented by E, we have

$$\begin{aligned} K(H_0, H_1) &= E \log f(X, \theta_0) - E \log f(X, \theta_1) \\ &= E \log \left[r^{X_1} \bar{r}^{(1-X_1)} \prod_{i=1}^{m-1} g(X_i, X_{i+1}; \phi_{i,i+1}) \right] \\ &\quad - E \log \left[r^{X_{\pi_1}} \bar{r}^{(1-X_{\pi_1})} \prod_{i=1}^{m-1} g(X_{\pi_i}, X_{\pi_{i+1}}; \phi^*_{\pi_i \pi_{i+1}}) \right] \\ &= -V(r) + \sum_{i=1}^{m-1} W(\phi_{i,i+1}) + V(r) - \sum_{i=1}^{m-1} w_{\pi_i \pi_{i+1}}(\phi^*_{\pi_i \pi_{i+1}}) \end{aligned}$$

which is minimized with respect to ϕ^* by $\phi^*_{\pi_i \pi_{i+1}} = \phi_{\pi_i \pi_{i+1}}$. Thus for H_1 corresponding to the ordering π,

$$K(H_0, H_1) = \sum_{i=1}^{m-1} W(\phi_{i,i+1}) - \sum_{i=1}^{m-1} W(\phi_{\pi_i \pi_{i+1}}).$$

Further when $\mathcal{A}$ is an arbitrary set of permutations,

$$K(H_0, H_{\mathcal{A}}) = \sum_{i=1}^{m-1} W(\phi_{i,i+1}) - \max_{\pi \in \mathcal{A}} \sum_{i=1}^{m-1} W(\phi_{\pi_i \pi_{i+1}}) \qquad (25)$$

Thus we can evaluate the effect of considering neighboring genes for the case of radiation hybrids just as we did in the case of sperm typing with W playing the role of $-V$.

5 Calculations

The Kullback-Leibler numbers for ordering the three genes (2,3,4) using sperm typing were calculated for various values of ϕ, yielding

$S_3 = S_3(\phi_{23}, \phi_{34})$,$S_4 = S_4(\phi_{23}, \phi_{34}, \phi_{45})$ and $S_5 = S_5(\phi)$, when considering 3,4,and 5 genes respectively.

The values of $S_3(a,a)$ peak at $S_3(0.12, 0.12) = 0.14$, but this peak is rather broad, since $S_3(a,a)$ is 0.103 at $a = 0.04$ and 0.099 at $a = 0.25$. The function $S_3(a,b)$ drops rapidly as a and b separate. It seems that $S_4(\phi_{23}, \phi_{34}, \phi_{45})$ is no improvement over $S_3(\phi_{23}, \phi_{34})$ when $\phi_{23} \leq \phi_{34}$. When $\phi_{23} > \phi_{34}$, there is room for substantial improvement by including gene 5. However, in those cases, including gene 1 also, rarely gives additional gain.

If $\phi_{23} = \phi_{34}$ is kept fixed, then S_5, regarded as a function of ϕ_{12} and ϕ_{45} is constant along squares for which the diagonal is along $\phi_{12} = \phi_{45}$. The function $S_5(a,a,a,a)$ attains a maximum value of 0.231 at $a = 0.10$. For fixed a, $S_5(b,a,a,b)$ peaks at $b = \tilde{b}(a)$ where $\tilde{b}(a) \approx 1.4a$. This value in turn has a peak of 0.258 at $a = 0.10$ and $b = 0.14$.

If ϕ_{23} is substantially larger than ϕ_{34}, then $S_5(\phi) = S_4(\phi_{23}, \phi_{34}, \phi_{45})$. If ϕ_{23} is not much larger than ϕ_{34}, the introduction of gene 1 begins to have some effect if ϕ_{45} is rather close to optimal for S_4 and ϕ_{12} is neither very small nor very large. There is another way to look at this phenomenon. If ϕ_{23} and $\phi_{34} < \phi_{23}$ are kept fixed, then S_5 is constant along rectangles in the (ϕ_{12}, ϕ_{45}) space. As ϕ_{34} decreases, these rectangles become elongated along the ϕ_{12} direction, and some of these rectangles degenerate to lines for small and large values of ϕ_{45}. When ϕ_{34} decreases enough, all the rectangles degenerate, and the level lines become parallel lines and S_5 is independent of ϕ_{12}.

In summary, if $\phi_{23} \approx \phi_{34}$, consideration of five genes is required to get improvement over that of three genes. If ϕ_{23} is considerably different than ϕ_{34}, one extra gene on the side of the two adjacent genes can give improvement, but the gene on the other side will not help. Table 1 presents the results for some cases and illustrates these comments.

The qualitative results for the use of radiation hybrid mapping are similar to those for sperm typing. Table 2 presents some results. The KL information depends on $\lambda\delta$ and r. In our table we take $\lambda = 1.0$ and $r = 0.4$, and we present the KL numbers $R_3 = R_3(\delta_{23}, \delta_{34})$,$R_4 = R_4(\delta_{23}, \delta_{34}, \delta_{45})$ and $R_5 = R_5(\delta)$. We note that the peak of $R_3(a,a)$ is 0.144 at $a = 0.28$ while values of a at 0.10 and 0.70 give 0.109 and

0.096. The peak value of $R_5(a,a,a,a)$ is 0.224 at $a = 0.22$, while the peak value of $R_6(b,a,a,b)$ is 0.250 at $a = 0.23$, $b = 0.34$.

These results have obvious potential application in selecting appropriate doses of radiation to increase the information content. Of course, the broad peak of R_3 indicates that KL values are not very sensitive to the choice of λ. The tables indicate that there are circumstances where considering 4 or 5 genes may double the information content, but also suggest that often there is little to gain by considering five or more genes simultaneously. The tables can easily be supplemented, since the calculations of the KL numbers are easily implemented.

REFERENCES

[1] Boehnke M, Lange K, Cox, DR (1991) Statistical Methods for Multipoint Radiation Hybrid Mapping. Submitted to *American Journal of Human Genetics* 1-47

[2] Goradia TM, Lange K (1990) Multilocus ordering strategies based on sperm typing, *Annals of Human Genetics*, **54** 49-77

[3] Lange K, Boehnke M (1991) Bayesian Methods and Optimal Experimental Design for Gene Mapping by Radiation Hybrids. *Technical Report*, 1-50

Table 1. KL information for ordering genes (2, 3, 4) using sperm typing considering 3,4, and 5 genes

ϕ_{12}	ϕ_{23}	ϕ_{34}	ϕ_{45}	S_3	S_4	S_5
.01	.01	.01	.01	.041	.041	.077
.02	.02	.02	.02	.067	.067	.122
.04	.04	.04	.04	.103	.103	.180
.10	.10	.10	.10	.146	.146	.231
.14	.14	.14	.14	.147	.147	.217
.20	.20	.20	.20	.127	.127	.169
.25	.25	.25	.25	.099	.099	.123
.14	.10	.10	.10	.146	.146	.258
.20	.14	.20	.14	.147	.147	.239
.10	.01	.01	.01	.041	.041	.059
.10	.02	.02	.01	.067	.067	.096
.10	.02	.02	.10	.067	.067	.101
x	.02	.01	.01	.035	.067	.067
x	.04	.02	.02	.055	.101	.101
x	.04	.02	.04	.055	.109	.109
x	.04	.02	.10	.055	.088	.088
x	.10	.04	.02	.065	.092	.092
x	.10	.04	.04	.065	.117	.117
x	.10	.04	.10	.065	.130	.130
x	.10	.08	.04	.121	.162	.162
y	.10	.08	.08	.121	.168	.199
z	.10	.08	.10	.121	.168	.227
y	.10	.08	.15	.121	.168	.207
x	.10	.08	.25	.121	.161	.161

x represents any value in the interval $(0, \infty)$
y represents any value in an interval containing $(0.04, 0.25)$
z represents any value in an interval containing $(0.08, 0.20)$

Table 2. KL information for ordering genes (2, 3, 4) using radiation hybrid mapping considering 3,4, and 5 genes

λ_{12}	λ_{23}	λ_{34}	λ_{45}	R_3	R_4	R_5
.01	.01	.01	.01	.023	.023	.044
.02	.02	.02	.02	.040	.040	.073
.10	.10	.10	.10	.109	.109	.180
.20	.20	.20	.20	.139	.139	.223
.30	.30	.30	.30	.144	.144	.217
.50	.50	.50	.50	.125	.125	.168
1.50	1.50	1.50	1.50	.024	.024	.025
.34	.23	.23	.23	.143	.143	.250
.20	.04	.04	.04	.064	.064	.099
.20	.10	.10	.04	.109	.109	.143
.20	.10	.10	.20	.109	.109	.188
x	.10	.04	.01	.048	.059	.059
x	.20	.10	.04	.079	.105	.105
x	.20	.10	.10	.079	.139	.139
x	.20	.10	.20	.079	.158	.158
x	.40	.20	.10	.084	.114	.114
x	.40	.20	.20	.084	.137	.137
x	.40	.20	.40	.084	.168	.168
x	.20	.15	.04	.111	.134	.134
y	.20	.15	.10	.111	.161	.164
z	.20	.15	.20	.111	.161	.206
w	.20	.15	.40	.111	.161	.178
x	.20	.15	.60	.111	.153	.153

x represents any value in the interval $(0, \infty)$

y represents any value in an interval containing $(0.01, 1.50)$

z represents any value in an interval containing $(0.10, 0.64)$

w represents any vlaue in an interval containing $(0.04, 1.10)$

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION
Unclassified

1b RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE

3 DISTRIBUTION / AVAILABILITY OF REPORT
Unlimited

4 PERFORMING ORGANIZATION REPORT NUMBER(S)
TR No. ONR-C-9

5. MONITORING ORGANIZATION REPORT NUMBER(S)

6a. NAME OF PERFORMING ORGANIZATION
Dept. of Statistics
Harvard University

6b. OFFICE SYMBOL *(If applicable)*

7a. NAME OF MONITORING ORGANIZATION

6c. ADDRESS *(City, State, and ZIP Code)*
Department of Statistics SC713
Harvard University
Cambridge, MA 02138

7b. ADDRESS *(City, State, and ZIP Code)*

8a. NAME OF FUNDING / SPONSORING ORGANIZATION

8b. OFFICE SYMBOL *(If applicable)*
Code 1111

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
N00014-91-J1005

8c. ADDRESS *(City, State, and ZIP Code)*
Office of Naval Research
Arlington, VA 22217-5000

10 SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO

11. TITLE *(Include Security Classification)*
Kullback-Leibler Information for Ordering Genes Using Sperm Typing and Radiation Hybrid Mapping

12. PERSONAL AUTHOR(S)
Herman Chernoff

13a. TYPE OF REPORT
Technical

13b. TIME COVERED
FROM ________ TO ________

14. DATE OF REPORT *(Year, Month, Day)*
October 1991

15. PAGE COUNT
15

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

See Reverse Side

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS

21 ABSTRACT SECURITY CLASSIFICATION

22a NAME OF RESPONSIBLE INDIVIDUAL
Herman Chernoff

22b TELEPHONE *(Include Area Code)*
617-495-5462

22c OFFICE SYMBOL

83 APR edition may be used until exhausted
All other editions are obsolete

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED