

A Revised Li-Sacks Formula For Calculating the Probability of Identity-by-Descent Proportion

Wentian Li

*Laboratory of Statistical Genetics
Rockefeller University
wli@linkage.rockefeller.edu*

Poster presented at the
Annual Meeting of American Society of Human
Genetics (Oct 28, 1998, 4-6pm, Denver , CO)

Motivation

Comparing the identity-by-descent (IBD) proportion among affected relatives in a pedigree is usually used for linkage analysis (reject or accept the null hypothesis of no linkage) instead of segregation analysis (reject or accept a particular disease model). For the latter, one needs to calculate the expected probability of IBD proportion under a given disease model. For complex diseases, such calculation can be complicated. This poster presents an easy and flexible procedure to carry out this calculation.

What's New?

There are several ways to calculate the probability of IBD given a disease model. One method is to list all possible mating types, determining the IBD in each type, and taking the average. Another method, borrowed from the classical quantitative genetics, is to calculate the co-variance of a quantitative trait between two relatives, and convert this co-variance to the probability of IBD proportion.

The third method, to be extended here, first developed by C.C.Li and L. Sacks in 1954, is Bayesian-based, that uses the conditional probabilities (in matrix form) of the second relative to have certain genotype, given the first relative's genotype, *and* the IBD between the two.

Original Li-Sacks Formula

$$\begin{aligned}
 P(k|AA) &= \frac{P(k, AA)}{P(AA)} \\
 &= \frac{P(AA|k) \cdot P(k)}{\sum_k P(AA|k)P(k)} \quad \leftarrow \text{Bayes' theorem} \\
 &= \frac{[\sum_{G_1, G_2} P(A_2|G_2)P(G_2|G_1, k)P(A_1|G_1)P(G_1)] \cdot P(k)}{\text{sum of numerator over } k}
 \end{aligned}$$

where

k : number of genotype IBD between two relatives (k=0,1,2)

AA : the event that two relatives are affected

$G_1(G_2)$: the first (second) relative's genotype

$P(A|G)$: penetrance (given by a disease model)

$P(k)$: prior probability of IBD=k (e.g., 1/4, 1/2, 1/4 for sib pairs)

$P(G)$: genotype frequency (e.g. $p^2, 2pq, q^2$)

$P(G_2|G_1, k)$: conditional probability of the second relative to have G_2 given the first relative has G_1 , and given the IBD is k (Li-Sacks' matrices)

The Li-Sacks' matrices ($P(G_2|G_1, k)$ or $\{T_{ij}(k)\}$) are:

$$\{T_{ij}(2)\} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \{T_{ij}(1)\} = \begin{pmatrix} p & q & 0 \\ p/2 & 1/2 & q/2 \\ 0 & p & q \end{pmatrix}, \{T_{ij}(0)\} = \begin{pmatrix} p^2 & 2pq & q^2 \\ p^2 & 2pq & q^2 \\ p^2 & 2pq & q^2 \end{pmatrix}$$

with

i : row index, or index of the first relative

j : column index, or index of the second relative

One might consider $\{T_{ij}(2)\}$ as being related to the situation of identical twins, $\{T_{ij}(1)\}$ with that of a parent-child pair, and $\{T_{ij}(0)\}$ with that of two unrelated individuals in a population.

Revised Li-Sacks Formula

In the revised Li-Sacks formula (the main result of this poster), the summation \sum_{G_1, G_2} over two genotypes now becomes $\sum_{i_m, j_m, i_p, j_p}$ over four alleles (paternal and maternal alleles of the first and the second relative); and the 3-by-3 Li-Sacks matrices become 2-by-2 matrices. More specifically,

$$P(k_m, k_p | AA) = \frac{[\sum_{i_m, j_m, i_p, j_p=1,2} f_{j_m j_p} \cdot t_{i_m j_m}(k_m) t_{i_p j_p}(k_p) \cdot f_{i_m i_p} \cdot p_{i_m} p_{i_p}] p(k_m) p(k_p)}{\text{sum of numerators over } k_m \text{ and } k_p},$$

where

- $k_m(k_p)$: maternal (paternal) allele IBD (=0,1)
- AA : the event that two relatives are affected
- $i_m, i_p(j_m, j_p)$: the index for the 1st (2nd) relative's maternal & paternal allele
- $f_{i_m i_p}, f_{j_m j_p}$: penetrance
- $p(k_m)(p(k_p))$: prior prob of maternal (paternal) allele IBD
(e.g. 1/2, 1/2 for sib pairs)
- p_{i_m}, p_{i_p} : allele frequencies (e.g. p,q)
- $t_{i_m j_m}(k_m)(t_{i_p j_p}(k_p))$: conditional probability of the second relative's maternal (paternal) allele to be j given the first relative's maternal (paternal) allele being i , and given the allele IBD= $k_m(k_p)$ (revised Li-Sacks matrices)

The revised Li-Sacks matrices $\{t_{ij}(k)\}$ are

$$\{t_{ij}(1)\} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \{t_{ij}(0)\} = \begin{pmatrix} p & q \\ p & q \end{pmatrix}$$

with

- i : row index, or index for the allele of the first relative
- j : row index, or index for the allele of the second relative

The derivation of $\{t_{ij}(k)\}$ is very simple (see next page).

From Li-Sacks Matrices to Revised Li-Sacks Matrices

This revision is to focus on allele IBD instead of genotype IBD, and thus reducing the 3-by-3 Li-Sacks matrices to 2-by-2 matrices. The meaning of the revised Li-Sacks matrices is extremely simple: if the allele IBD is 1, the two relatives' alleles are the same; if the allele IBD is 0, the probability of the second relative to have an allele is equal to the population allele frequency.

This revision is a further extension of the paper by Campbell and Elston (1971), in which ordered genotypes were considered instead of genotype, and the Li-Sacks matrices there were 4-by-4. In fact, these 4-by-4 matrices are external tensor products of the 2-by-2 matrices used here.

Note that $P(k|AA) = \sum_{k_m+k_p=k} P(k_m, k_p|AA)$.

Extension 1: Unilineal Relative Pairs

For unilineal relative pairs (e.g. a parent and a child), the genotype IBD cannot be 2: one of the allele IBD between the two relatives is always 0. We can set the prior probability $p(k_p = 1) = 0$. By using a property of the revised Li-Sacks matrix $t_{ij}(0) = p_j$, we have

$$\begin{aligned} P(k_m|AA) &\equiv P(k_m, k_p = 0|AA) \\ &= \frac{\sum_{i_m j_m} (\sum_{j_p} f_{j_m j_p} p_{j_p}) t_{i_m j_m}(k_m) (\sum_{i_p} f_{i_m i_p} p_{i_p}) p(k_m)}{\text{sum of numerators over } k_m} \\ &= \frac{\sum_{i_m j_m} \overline{f_{j_m}} \cdot t_{i_m j_m}(k_m) \cdot \overline{f_{i_m}} \cdot p(k_m)}{\text{sum of numerators over } k_m}, \end{aligned}$$

where $\overline{f_{j_m}}$ and $\overline{f_{i_m}}$ are averaged penetrances.

Extension 2: Three Or More Alleles

Suppose a disease model is defined under the assumption of three alleles (wild type, severe mutant, and mild mutant), how will the Li-Sacks formula be modified?

Simple! First, the upper limit of the summation for i_m, i_p, j_m, j_p is 3 instead of 2. Second, the Li-Sacks matrices become:

$$\{t_{ij}(1)\} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \{t_{ij}(0)\} = \begin{pmatrix} p_1 & p_2 & p_3 \\ p_1 & p_2 & p_3 \\ p_1 & p_2 & p_3 \end{pmatrix}.$$

where p_1, p_2, p_3 are the three allele frequencies.

Extension 3: Unaffected-Unaffected and Unaffected-Affected pairs

The extension is very simple: replace the penetrance f by the $1 - f$. For example, for the unaffected-unaffected pairs:

$$P(k_m, k_p|UU) = \frac{[\sum_{i_m, j_m, i_p, j_p} (1 - f_{i_m i_p}) \cdot t_{i_m j_m}(k_m) t_{i_p j_p}(k_p) \cdot (1 - f_{j_m j_p}) \cdot p_{i_m} p_{i_p}] p(k_m) p(k_p)}{\text{sum of numerators over } k_m \text{ and } k_p}$$

where

UU : the event that two relatives are unaffected

For unaffected-affected pairs, the IBD probability is:

$$P(k_m, k_p|UA) = \frac{[\sum_{i_m, j_m, i_p, j_p} (1 - f_{i_m i_p}) \cdot t_{i_m j_m}(k_m) t_{i_p j_p}(k_p) \cdot f_{j_m j_p} \cdot p_{i_m} p_{i_p}] p(k_m) p(k_p)}{\text{sum of numerators over } k_m \text{ and } k_p}$$

where

UA : the event that one relative is affected and another is not

Extension 4: Identity-by-State

Interestingly, Li-Sacks formula can be easily extended to the situation of identity-by-state. We replace the revised Li-Sacks matrices $t_{ij}(k)$ by the similar matrices $s_{ij}(IBS = k)$:

$$\{s_{ij}(1)\} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \{s_{ij}(0)\} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

The new formula looks like this:

$$P^{IBS}(k_m, k_p|AA) = \frac{[\sum_{i_m, j_m, i_p, j_p=1,2} f_{j_m j_p} \cdot s_{i_m j_m}(k_m) s_{i_p j_p}(k_p) \cdot f_{i_m i_p} \cdot p_{i_m} p_{i_p}] p^{IBS}(k_m) p^{IBS}(k_p)}{\text{sum of numerators over } k_m \text{ and } k_p}$$

Extension 5: Two Unlinked Disease Genes (Two-Locus Models)

There are 8 indices for alleles at two locations of two relatives: $i_m, i_p, j_m, j_p, i'_m, i'_p, j'_m, j'_p$. The joint probability of IBD of four alleles (maternal and paternal alleles at first and second locus) is:

$$\begin{aligned}
 & P(k_m, k_p, k'_m k'_p | AA) \propto \\
 & \sum_{i_m, i_p, j_m, j_p, i'_m, i'_p, j'_m, j'_p} f_{j_m j_p j'_m j'_p} \cdot t_{i_m j_m}(k_m) t_{i_p j_p}(k_p) t_{i'_m j'_m}(k'_m) t_{i'_p j'_p}(k'_p) \cdot f_{i_m i_p i'_m i'_p} \\
 & \cdot p_{i_m} p_{i_p} p_{i'_m} p_{i'_p} \cdot p(k_m) p(k_p) p(k'_m) p(k'_p)
 \end{aligned}$$

where

- $i(j)$: index for the first (second) relative
 - $i, j(i', j')$: index for the first (second) locus
 - $i_m, j_m, i'_m, j'_m(i_p, \dots)$: index for the maternal (paternal) allele
 - $f_{i_m i_p i'_m i'_p}, f_{j_m j_p j'_m j'_p}$: penetrances (given by a two-locus model)
 - $p(k_m), p(k_p), p(k'_m), p(k'_p)$: prior probability of allele IBD
 - $p_{i_m}, p_{i_p}, p_{i'_m}, p_{i'_p}$: allele frequencies
 - $t_{ij}(k)$: revised Li-Sacks matrices (k is the allele IBD)
- (1)

The calculation of IBD probability for two-locus models or multiple-locus models is tedious but straightforward. Note $P(k, k' | AA) = \sum_{k, k'} P(k, k' | AA)$

Extension 6: IBD at a Marker Linked to a Disease Gene

This is another calculation, though not new, but becomes easier to derive using the revised Li-Sacks formula. We start with this formula:

$$P(k_m^M, k_p^M | AA) = \sum_{k_m k_p} P(k_m^M | k_m) \cdot P(k_p^M | k_p) \cdot P(k_m k_p | AA)$$

where

$P(k_m, k_p | AA)$: as defined before, probability of allele IBD at the disease gene

k_m^M, k_p^M : allele IBD at the marker

$P(k_m^M | k_m), P(k_p^M | k_p)$: conditional probability of an allele IBD at the marker given the allele IBD at the disease gene locus

There are four $P(k_m^M | k_m)$'s, which are:

$$\begin{aligned} P(k^M = 1 | k = 1) &= P(k^M = 0 | k = 0) = \theta^2 + (1 - \theta)^2 \\ P(k^M = 1 | k = 0) &= P(k^M = 0 | k = 1) = 1 - \theta^2 - (1 - \theta)^2 \end{aligned}$$

where θ is the recombination fraction between the marker and the disease gene locus.

The calculation of $P(k_m^M, k_p^M | AA)$ by the revised Li-Sacks formula is simpler than that by counting mating types as in Appendix B of Haseman & Elston, 1972.

Extension 7: IBD at Two Markers Linked to Two Disease Genes

This is the situation when the two disease genes are not linked (e.g. on two different chromosomes), but two markers are linked to the two genes. The probability for IBD proportion at four alleles is:

$$\begin{aligned}
 P(k_m^M, k_p^M, k_m'^M, k_p'^M | AA) &= \sum_{k_m, k_p, k_m', k_p'} P(k_m^M, k_p^M, k_m'^M, k_p'^M | k_m \cdot k_p \cdot k_m' \cdot k_p') \\
 &\times P(k_m, k_p, k_m', k_p' | AA) \\
 &= \sum_{k_m, k_p, k_m', k_p'} P(k_m^M, k_p^M | k_m \cdot k_p) P(k_m'^M, k_p'^M | k_m' \cdot k_p') \\
 &\times P(k_m, k_p, k_m', k_p' | AA) \tag{2}
 \end{aligned}$$

where

$$\begin{aligned}
 P(k_m, k_p, k_m', k_p' | AA) &: \text{ same as before (see Extension 5)} \\
 P(k_m^M, k_p^M | k_m, k_p) &: \text{ same as before (see Extension 6)}
 \end{aligned}$$

Note that for $P(k_m^M, k_p^M | k_m, k_p)$ at the first locus, the recombination θ is used. For $P(k_m'^M, k_p'^M | k_m', k_p')$ at the second locus, however, a different recombination θ' is used.

To make the notation less confusing, the following Drawing shows the meaning of the indices (i and j for the first and the second relative; subscripts m and p for the maternal and paternal allele; superscript $'$ for the second locus; and superscript M for the marker).

References

- CC Li, L Sacks (1954), “The derivation of joint distribution and correlation between relatives by the use of stochastic matrices”, *Biometrics*, 10:347-360.
- W Li (1998), “A revised Li-Sacks formula for calculating the probability of identical-by-descent proportion”, preprint.
- MA Campbell, RC Elston (1971), “Relatives of probands: models for preliminary genetic analysis”, *Annals of Human Genetics*, 35:225-236.
- W Li (1995-1998), “Bibliography on allele-sharing linkage analysis”, online resource. URL: <http://linkage.rockefeller.edu/bib/ibd/>
- JK Haseman, RC Elston (1972), “The investigation of linkage between a quantitative trait and a marker locus”, *Behavior Genetics*, 2:3-19.