Am. J. Hum. Genet. 64:1464–1472, 1999

Relationship Estimation by Markov-Process Models in a Sib-Pair Linkage Study

Jane M. Olson

Department of Epidemiology and Biostatistics, Rammelkamp Center for Education and Research, MetroHealth Campus, Case Western Reserve University, Cleveland

Summary

The results of sib-pair linkage studies may be compromised if a substantial number of putative sib pairs are not actually sib pairs. For classification of pairs in a sib-pair genome scan, I propose multipoint methods that are based on a Markov-process model of allele sharing along the chromosome. These methods can be implemented by standard algorithms that compute multipoint marker allele–sharing probabilities for sib pairs. When marker data from at least half the genome are used, misclassification rates are small. The methods will be implemented in an upcoming version of the computer software package S.A.G.E.

Introduction

Linkage studies, including those using affected relative pairs, rely heavily on the assumption that the type of relationship identified during pedigree collection is the true genetic relationship between the two members of the relative pair (Boehnke and Cox 1997; Goring and Ott 1997). Genotyping of intervening relatives, such as both parents of a sib pair, is an effective means of identification of incorrect relationships. In many linkage studies, particularly those of late-onset diseases, parents and other intervening relatives are not available for genotyping, and probabilistic methods must be relied on. Methods based on the observed numbers of marker alleles identical by state (IBS) have been proposed by several authors (Chakraborty and Jin 1993*a,* 1993*b;* Ehm and Wagner 1996; Stivers et al. 1996). More recently, other authors have proposed likelihood-based methods of relationship estimation for affected-sib-pair studies (Boehnke and Cox 1997; Goring and Ott 1997). These methods are based on computation of the probability of sets of marker data conditional on a given relationship. Goring and Ott (1997) have focused on elimination of false sib pairs from a sib-pair analysis, using a Bayesian approach that incorporates prior probabilities for each type of relationship. Boehnke and Cox (1997) have computed a multipoint likelihood for each possible type of relationship and then have formed likelihood ratios that are used to classify pairs on the basis of type of relationship.

In this study, I use Markov-process models to calculate pair-specific statistics that estimate average genomewide allele sharing, an idea that I first suggested in a earlier report (Olson 1998). The primary focus is on classification of putative sib pairs in a sib-pair linkage study when parental marker genotypes are unavailable. Since only one multipoint likelihood calculation is required, the statistics used for classification are constructed from multipoint sib-pair allele-sharing probabilities currently available from standard algorithms (e.g., see Kruglyak and Lander 1995; Idury and Elston 1997). As a result, they can be incorporated easily into existing sib-pair multipoint-linkage programs so that relationship estimation can be made an automatic component of a genomewide sib-pair linkage study, with little or no additional effort. Visual display of the relationship statistics in the form of histograms provides the researcher an immediate sense of the types and numbers of relationships present in a set of putative sib pairs. Using simulations, I show that a simple classification method is extremely accurate when relationship estimation is based on more than half the genome.

Methods and Results

I focus on the problem of classification of sib pairs, half-sib pairs, unrelated pairs, MZ twin pairs, and parent/offspring (P/O) pairs in a sample of putative sib pairs, in the context of a genome scan. Let $\hat{f}_{jis}$ be the estimated probability that sib pair j shares i marker alleles identical by descent (IBD) at a location s on a chromosome of length L cM. I assume, throughout this report, than these

Received June 10, 1998; accepted for publication February 19, 1999; electronically published March 26, 1999.

Address for correspondence and reprints: Dr. Jane M. Olson, Department of Epidemiology and Biostatistics, Case Western Reserve University, MetroHealth Medical Center R-255, 2500 MetroHealth Drive, Cleveland, OH 44109. E-mail: olson@darwin.cwru.edu

0002-9297/99/6405-0028$02.00

allele-sharing probabilities are obtained by multipoint methods. At location s, the estimated number of alleles shared IBD by a sample of N sib pairs may be obtained by

$$X_s = \sum_{j=1}^{N} (\hat{f}_{j1s} + 2\hat{f}_{j2s}) \ .$$

If the marker is fully informative, then X_s is the total number of alleles shared IBD in the sample, at location s.

Feingold et al. (1993) proposed a Gaussian-process model to describe the ideal (i.e., infinitely dense and fully informative) process X_s along the chromosome. For the ideal process, consider the mean-sharing statistic $Z_s = (X_s - N)/(N/2)^{1/2}$. For a large sample of randomly sampled sib pairs, the statistic Z_s has a mean of 0, a variance of 1, and an approximating Gaussian-process covariance function of $\exp(-\beta|t|)$, where $\beta = .04$ for sib pairs (Feingold et al. 1993). The parameter β is a function of the recombination process and assumes that crossovers are independent—that is, that there is no crossover interference.

Now consider a single random sib pair j and let Z_{js} be the mean-sharing statistic for a single pair ($N = 1$). We wish to obtain a measure of the average number of shared alleles over the entire genome. Let $k = 1,2,\ldots,22$ index the human autosomes and let L_k be the length, in centimorgans, of chromosome k. The statistic

$$Y_{jk} = \frac{1}{L_k}\int_0^{L_k} Z_{js}ds$$

has expectation

$$E(Y_{jk}) = \frac{1}{L_k}\int_0^{L_k} E(Z_{js})ds = 0$$

and variance

$$\mathrm{Var}(Y_{jk}) = \frac{1}{L_k^2}\int_0^{L_k}\int_0^{L_k} \mathrm{Cov}(Z_{js},Z_{jr})drds$$

$$= \frac{2}{\beta L_k} - \frac{2}{(\beta L_k)^2}(1 - e^{-\beta L_k}) \qquad (1)$$

(Parzen 1962, pp. 78–86; Olson 1998). In the ideal case of fully informative, infinitely dense markers, the statistic Y_{jk} is the difference between the proportions of the chromosome sharing two and zero alleles IBD. More generally, it is the difference between the absolute areas above and below the null mean (sharing of one allele IBD), divided by the length of the chromosome.

If putative sib pair j is a true sib pair, then $Y_{jk}/[\mathrm{Var}(Y_{jk})]^{1/2}$ has a standard normal distribution as $L_k \to \infty$. In practice, the normal approximation is somewhat inadequate for single chromosomes of modest length (see simulations below). A genomewide measure is given by

$$Y_j = \left(\sum_{k=1}^{22} Y_{jk}\right) \Bigg/ \left[\sum_{k=1}^{22} \mathrm{Var}(Y_{jk})\right]^{1/2} \ , \qquad (2)$$

which is well approximated by a standard normal distribution (see simulations below). Similar measures may be computed for any number of chromosomes. Since the focus is on relationship estimation, I propose the estimation of genomewide Y_j for each of $j = 1,\ldots,N$ sib pairs in the sample. The statistics $\hat{Y}_j$ can be obtained, in practice, by a standard algorithm (e.g., see Kruglyak and Lander 1995; Idury and Elston 1997), to calculate multipoint allele sharing at equally spaced points throughout the genome. For each chromosome, the absolute areas above and below the estimated mean-corrected mean allele-sharing curve can be approximated by rectangles (fig. 1), then divided by the length of the chromosome, which is equivalent to computation of

$$\hat{Y}_{jk} = \left[c\sqrt{2}\sum_{p=1}^{P}(X_p - 1)\right] \Big/ P \ , \qquad (3)$$

where P is the number of points at which allele sharing is computed and c is the distance between points. For example, if marker allele–sharing estimates are available at 1-cM intervals over a 150-cM chromosome, then $P = 151$ and $c = 1$.

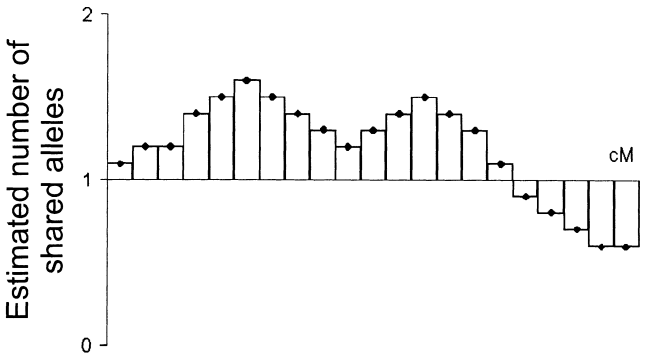

Figure 1 Example sample path. Dots indicate estimated multipoint number of alleles shared IBD; rectangles indicate approximated area.

Table 1

Empirical Mean, SD, and Two-Sided–Tail Probabilities for $\hat{Y}_j$, for True Full Sibs

Intermarker Distance and No. of Chromosomes	$\hat{Y}_j$				
			For Two-Sided–Tail Probability =		
	Mean	SD	.05	.01	.001
Ideal (AMIC = 1.000):[a]					
1	−.0018	1.0004	.0450	.0032	.0000
5	−.0039	1.0071	.0515	.0101	.0008
10	−.0056	1.0086	.0512	.0101	.0015
22	−.0083	1.0022	.0501	.0098	.0011
10:					
AMIC = .633; 5 alleles:					
1	.0093	.9136	.0254	.0000	.0000
5	.0102	.9153	.0317	.0045	.0003
10	.0157	.9114	.0300	.0044	.0002
22	.0229	.9028	.0280	.0042	.0000
AMIC = .731; 10 alleles:					
1	.0060	.9529	.0336	.0014	.0000
5	.0107	.9500	.0384	.0057	.0004
10	.0135	.9469	.0377	.0057	.0004
22	.0201	.9553	.0374	.0068	.0008
25:					
AMIC = .417; 5 alleles:					
1	.0037	.7647	.0076	.0000	.0000
5	.0083	.7625	.0088	.0004	.0000
10	.0117	.7615	.0104	.0006	.0000
22	.0174	.7615	.0088	.0004	.0000
AMIC = .526; 10 alleles:					
1	.0037	.8451	.0154	.0000	.0000
5	.0083	.8478	.0199	.0013	.0001
10	.0117	.8514	.0217	.0020	.0000
22	.0174	.8553	.0228	.0024	.0000

[a] Infinitely dense and fully informative.

Classification of Sibs and Half-Sibs

I simulated genomic marker data for two types of relative pairs: sibs and half-sibs. Each marker locus had 5 or 10 equally frequent alleles, and no crossover interference was assumed when the data were generated. Two intermarker distances, 10 cM and 25 cM, were studied. Each chromosome had length 150 cM, giving a total of 3,300 cM for 22 chromosomes. Each simulation consisted of 220,000 chromosomes, which were then grouped into sets of 1, 5, 10, or 22 chromosomes, so that results were based on ⩾10,000 replicates. Statistics $\hat{Y}_j$ were computed by equations (2) and (3). During multipoint calculation of allele-sharing, the Kosambi (1944) map function was used to convert map distance to recombination fraction, so that robustness to incorrect interference assumption could be partially assessed. One simulation assuming an infinitely dense, fully informative map was also performed, to ensure performance of the statistic in the ideal setting. For this simulation, recombinations along each chromosome were assumed to follow a Poisson process, with a mean of 1 Morgan. Allele sharing at the starting point was determined randomly for the pair of chromosomes from the same parent, and the lengths of segments shared and not shared were determined on the basis of the locations of the recombinations.

Before evaluating the ability of the method to correctly classify pairs, I examined the accuracy of the normal approximation. Table 1 gives the mean, SD, and two-sided–tail probabilities of $\hat{Y}_j$ for simulations in which the true relationship is full sib. For the ideal marker map, the standard normal distribution is a good approximation, except when only one chromosome is included. For less informative marker maps, tail probabilities are too small, especially for the sparsest, least informative marker maps. The parameter β, which I take to be fixed at a value of .04, actually depends on map density and marker informativity, as well as on degree of interference, the validity of the fixed marker map, and assumed marker-allele frequencies. Teng and Siegmund (1998) have shown that the covariance of the Gaussian process increases as marker informativity decreases, so that the variance of $\hat{Y}_j$ decreases. This effect can be observed in the present simulations, since the empirical SDs of $\hat{Y}_j$ are considerably <1. The number of chromosomes, which is not related to marker informativity, does not affect

the variance. Use of the incorrect interference assumption gives a negligible positive bias in the mean of the test statistic.

I then evaluated the ability of the statistics to accurately classify sibs and half-sibs. Table 2 gives the mean and SD of $\hat{Y}_j$ when the true relationship is half-sib. The mean of the test distribution increases with decreasing marker informativity and decreasing numbers of chromosomes, so that overlap between the true-sib and half-sib distributions increases as the amount of information decreases. To provide guidelines for using $\hat{Y}_j$ in practice, I calculated the optimal classification value (OCV) for each simulation, using the empirical means and SDs of the two distributions and assuming normality and equal proportions of sibs and half-sibs. I defined the OCV as the value that minimizes the total misclassification; it is also the value at which the two density functions are equal. OCVs and misclassification probabilities are also given in table 2. When the entire genome is used, misclassification is rare, if OCVs are used. The probability of misclassification of a sib as a half-sib is higher than that of misclassification of a half-sib as a sib, because the sib distribution has a larger SD.

The OCVs vary considerably and increase with decreasing marker informativity. Using the results from these simulations, I fitted a linear-regression model to the logarithm of minus the OCV as a function of the length, in centimorgans, of the genotyped genome, divided by 150 (T) and average marker information content (AMIC) for true sib pairs. To compute AMIC, I computed marker information content (Kruglyak and Lander 1995) at 1-cM intervals throughout the genome and averaged those values. I then averaged these AMIC values over 10 replicate genomes; the SD of AMIC values was ~.01, indicating that there is little variability in AMIC over replicates. The resulting AMIC values are shown in table 1. The best-fit regression model,

$$\log_{10}(-\mathrm{OCV}) = -.141 + .524 \log_{10}\mathrm{T} + .237 \log_{10}\mathrm{AMIC} - .861 (\log_{10}\mathrm{AMIC})^2 , \quad (4)$$

accounted for 99.9% of the variance of $\log_{10}(-\mathrm{OCV})$. When this regression equation is used to obtain classification values a priori, it is important to use the Kruglyak and Lander (1995) algorithm to obtain AMIC. (Note that T actually equals the number of 150-cM chromosomes used in the simulations and was convenient

Table 2

Characteristics of $\hat{Y}_j$ for Half-Sibs

Intermarker Distance and No. of Chromosomes	$\hat{Y}_j$			Minimum Probability of Misclassification as		
	Mean	SD	OCV	Half-Sibs	Sibs	Total
Ideal:[a]						
1	−1.34	.71	−.61	.2708	.1507	.4215
5	−3.00	.71	−1.68	.0477	.0309	.0786
10	−4.25	.70	−2.44	.0078	.0052	.0130
22	−6.30	.71	−3.66	.0001	.0001	.0002
10:						
5 Alleles:						
1	−1.04	.59	−.45	.3037	.1442	.4480
5	−2.43	.59	−1.37	.0633	.0367	.1000
10	−3.44	.60	−2.00	.0128	.0079	.0207
22	−5.10	.60	−3.01	.0004	.0002	.0006
10 Alleles:						
1	−1.19	.65	−.52	.2906	.1488	.4395
5	−2.65	.65	−1.48	.0565	.0350	.0915
10	−3.75	.65	−2.16	.0104	.0067	.0171
22	−5.57	.65	−3.25	.0003	.0002	.0005
25:						
5 Alleles:						
1	−.82	.55	−.40	.3388	.1827	.5215
5	−1.82	.55	−1.02	.0976	.0640	.1615
10	−2.58	.55	−1.47	.0282	.0193	.0476
22	−3.82	.55	−2.20	.0020	.0014	.0034
10 Alleles:						
1	−.99	.59	−.50	.3012	.1589	.4601
5	−2.21	.59	−1.27	.0735	.0465	.1200
10	−3.13	.59	−1.82	.0173	.0113	.0286
22	−4.64	.59	−2.72	.0007	.0005	.0012

[a] Infinitely dense and fully informative.

for obtaining the regression equation. Taking T to be the total length of the genotyped genome divided by 150 allows use of the regression equation in practice.)

Unrelated Individuals and MZ Twins

The simulations were repeated for true unrelated pairs. Results are shown in table 3, for only the most and least informative marker maps. Using all simulations, not just those shown in table 3, I determined the OCVs and fitted a regression equation:

$$\log_{10}(-\mathrm{OCV}) = .421 + .506 \log_{10}\mathrm{T} + 1.162 \log_{10}\mathrm{AMIC} + .472 (\log_{10}\mathrm{AMIC})^2 \ . \quad (5)$$

This regression equation explained 99.9% of the variance in $\log_{10}(-\mathrm{OCV})$. This OCV differentiates between unrelated individuals and half-sibs (not sibs), but only when $\hat{Y}_j$ is computed under the assumption that a sibling relationship exists. In other words, no recalculation of allele sharing or of $\hat{Y}_j$ is required.

For MZ twins (or, in practice, duplicate DNA samples), $\hat{Y}_j$ shows little variability and 100% power, provided that at least five chromosomes are included and the classification criterion is $\geqslant\sim 3$. In this case, OCVs need not be computed; it suffices to set the classification value at 3.27, corresponding to a probability of $\leqslant .0005$ that a true sib will be declared an MZ twin (regardless of the number of chromosomes included), so that virtually no MZ twins will be declared true sibs (provided that no fewer than five chromosomes are used).

P/O Pairs

P/O pairs are always expected to share exactly one allele IBD, so that $\hat{Y}_j$ cannot be used to discriminate between sib pairs and P/O pairs. Because such mistakes can occur in real data if blood samples are mislabeled, I provide a second Markov-process statistic to detect P/O pairs. For chromosomal location s, use

$$X_s^* = \sum_{j=1}^{N} (\hat{f}_{j2s} + \hat{f}_{j0s} - \hat{f}_{j1s}) \ . \quad (6)$$

For a fully informative location s, the Gaussian-process statistic

$$Z_s^* = \sum_{j=1}^{N} \frac{X_s^*}{N^{\frac{1}{2}}}$$

has a standard normal distribution in a large sample of sib pairs, with covariance function $\exp(-\beta|t|)$, where now $\beta = .08$. Now consider a single sib pair. Integrating over the genome in the same manner as has been used above, we obtain a new statistic, Y_j^*, which compares the proportion of the genome sharing one allele IBD versus the proportion sharing zero or two alleles IBD. For a true sib pair, Y_j^* has an expected mean of 0 and a variance of 1.

I simulated sib pairs and P/O pairs to test the performance of $\hat{Y}_j^*$. For true sib pairs, the empirical type I error rates were close to the nominal values of .05, .01, and .001, indicating that the normal approximation is excellent, provided that more than one chromosome is included (not shown). For true P/O pairs, table 4 shows results for the most (after the ideal) and least informative marker maps. For the ideal map, $\hat{Y}_j^*$ has a variance of 0 when applied to P/O pairs, and so it is not useful for determination of the OCV. The OCV needed for a particular data set can be determined by

$$\log_{10}(-\mathrm{OCV}) = .475 + .518 \log_{10}\mathrm{T} + 2.220 \log_{10}\mathrm{AMIC} \ , \quad (7)$$

the best-fitting regression model for differentiation between sib pairs and P/O pairs. It explained 99.8% of the variance of $\log_{10}(-\mathrm{OCV})$; addition of quadratic terms did not increase the percentage of explained variation.

Table 3

Characteristics of $\hat{Y}_j$ for Unrelated Individuals

Intermarker Distance and No. of Chromosomes	$\hat{Y}_j$			Minimum Probability of Misclassification as		
	Mean	SD	OCV	Unrelated	Half-Sibs	Total
10; 10 alleles:						
1	−2.24	.25	−1.82	.1669	.0458	.2127
5	−5.01	.25	−4.29	.0058	.0020	.0078
10	−7.08	.25	−6.11	.0001	.0001	.0002
22	−10.50	.25	−9.10	.0000	.0000	.0000
25; 5 alleles:						
1	−1.42	.46	−1.07	.3228	.2253	.5481
5	−3.17	.46	−2.52	.1010	.0794	.1804
10	−4.48	.46	−3.59	.0330	.0266	.0597
22	−6.64	.46	−5.34	.0029	.0024	.0052

Table 4

Characteristics of Y_j^* for P/O Pairs

Intermarker Distance and No. of Chromosomes	Y_j^*			Minimum Probability of Misclassification as		
	Mean	SD	OCV	Sibs	P/O	Total
10; 10 alleles:						
1	−1.91	.201	−1.43	.0085	.0483	.0568
5	−4.28	.201	−3.42	.0000	.0000	.0000
10	−6.05	.201	−4.87	.0000	.0000	.0000
22	−8.98	.201	−7.27	.0000	.0000	.0000
25; 5 alleles:						
1	−.68	.251	−.33	.0836	.2518	.3355
5	−1.52	.252	−.95	.0125	.0297	.0422
10	−2.15	.251	−1.39	.0014	.0031	.0044
22	−3.18	.254	−2.09	.0000	.0000	.0000

Strategy for Classification of Sib Pairs and Nonsib Pairs

Below, I suggest an algorithm for use of $\hat{Y}_j$ and $\hat{Y}_j^*$ to classify pairs in a sib-pair linkage study. As part of the analysis, I recommend visual examination of histograms of $\hat{Y}_j$ and $\hat{Y}_j^*$. One of the advantages of the Markov approach is that histograms of the Markov statistics provide the linkage analyst with an immediate sense of the nature and extent of the nonsib problem in a particular data set. These histograms also can be used informally to ensure that the formal algorithm gives sensible results, to adjust classification values if desired, or to decide on the classification method. For example, if the half-sib and sib distributions overlap substantially and the proportion of half-sib pairs appears to be nontrivial, then a Bayesian approach may be preferred. Conversely, examination of the histograms may reveal that little, if any, formal testing is necessary; if the sib and half-sib distributions appear to be completely separate, then agonizing over choice of method or classification value is unnecessary.

In practice, OCVs can be obtained automatically in the course of the multipoint calculation, by saving both the information content for the total data set and the pair-specific allele-sharing probabilities at each point on the marker map. AMIC is computed by averaging the information content over the map. In real data sets, in which most nonsibs are half-sibs or unrelated individuals, AMIC will be overestimated, since it is computed under the assumption that all putative sib pairs are true sib pairs. In sets of putative DZ twins, AMIC will usually be underestimated, if most nonsibs are MZ twins. In reality, AMIC depends on both the true relationship and the assumed relationship. However, the investigator will not know, prior to relationship estimation, which pairs are true sibs. For the classification strategy to be effective, it is necessary that the misclassification rates not be too sensitive to misspecification of the OCV. A summary of the procedure is as follows:

1. Obtain the AMIC from the existing data by calculating the Kruglyak and Lander (1995) marker information content at 1-cM intervals throughout the genome and then averaging these values. Obtain T by dividing the total length, in centimorgans, of the (genotyped) genome by 150.

2. Use T and AMIC and the regression equations (4), (5), and (7) to obtain the OCVs. Put the OCV for MZ twins equal to 3.27.

3. For each pair, compute the multipoint allele-sharing probabilities at equally spaced points throughout the genome, using a standard multipoint algorithm. Use equations (1)–(3) to obtain $\hat{Y}_j$, the estimate of Y_j. Also compute $\hat{Y}_j^*$, using equations (1) (after putting $\beta = .08$), (2), and (6).

4. Classify each pair, using the classification values obtained in step 2. When $\hat{Y}_j$ is used, each pair will fall into one of four mutually exclusive categories: unrelated, half-sib, sib, and MZ twin.

5. For each pair classified as a sib in step 4, reclassify as sib or P/O pair, using $\hat{Y}_j^*$.

I tested this strategy by using a simulated data set with 10,000 sib pairs, a random 10% of which were actually half-sib pairs. All putative parents were untyped. To make the data more realistic, I used an 11-marker map for each chromosome, with varying intermarker distances (10–20 cM, mean 15 cM) and varying numbers of equally frequent alleles (3–10, mean 5.5). The estimated value of AMIC when all the pairs were used was .587; when only sib pairs were use, it was .521. When the estimated classification values were used, the total misclassification rates were .1265, .0329, and .0018 for 5, 10, and 22 chromosomes, respectively, which compare favorably to the optimal total misclassification rates of .1213, .0250, and .0011, respectively.

I then simulated a data set with 10,000 relative pairs, randomly choosing 10% to be half-sib pairs, 10% to be unrelateds, 10% to be MZ twins, 10% to be P/O pairs,

and the remaining 60% to be true sib pairs. Again, all putative parents were untyped, and the same marker map was used. This data set contains an unusually high number of nonsib pairs. Genotyping error was included by changing a random 1% of the alleles before multipoint computation. Inclusion of genotyping error introduces a downward bias in the relationship statistics for all relationships, since true mean allele sharing is underestimated. For this data set, AMIC was estimated to be .535, which is close to the sib-pair value of .521. The classification results are given in table 5, and histograms of the $\hat{Y}_j$ and $\hat{Y}_j^*$ values are given in figure 2. When either 10 chromosomes or the entire genome was used, few pairs were misclassified, and the corresponding histograms show that the types of pairs are easily distinguished.

When five chromosomes were used, misclassification rates were larger, as anticipated. There is substantial overlap of the distributions, and the classification method is more sensitive to choice of cut point. In addition, because the classification value that distinguishes half-sibs and sibs is high, the sib distribution is truncated in the lower tail, which will lead to an increase in the false-positive rate of linkage results if only the pairs classified as sib pairs are included in the subsequent linkage analysis. If the goal of relationship estimation is to eliminate nonsib pairs from the data set, then methods that incorporate prior probabilities will be more appropriate if substantially less than half of the genome is genotyped.

I performed relationship testing on a set of 49 sib pairs affected with intracranial aneurysm (Ronkainen et al. 1997). Markers were genotyped at ~10-cM intervals (Weber screening set 8) throughout the genome, and statistics $\hat{Y}_j$ and $\hat{Y}_j^*$ were calculated. Of the $\hat{Y}_j$ values, 1 was −4.35, 1 was −2.83, 6 were between −2.5 and −1.0, 35 between −1.0 and 1.0, 6 were between 1.0 and 2.5, and 0 were >2.5. The OCV for sibs and half-sibs was calculated to be ~−3.46, so that one putative sib pair ($\hat{Y}_j = -4.35$) was classified as a half-sib pair. The putative sib pair with $\hat{Y}_j = -2.83$ was classified as a true sib pair. This pair was a member of a sib trio with pairwise $\hat{Y}_j$ equal to −2.83 (pair 1-2), 1.48 (pair 2-3), and −.82 (pair 1-3). Viewed together, these pairwise results indicate that the three offspring are indeed sibs. $\hat{Y}_j$ values were also calculated for four putative half-sib pairs, yielding values of −4.57, −4.30, −5.12, and −4.83, confirming these half-sib relationships. No pairs were classified as P/O pairs, unrelated pairs, or MZ twins. The same classifications were obtained by the Boehnke and Cox (1997) method.

Discussion

I have proposed methods for estimation of relationships in sib-pair studies that are based on Markov-process models and that may be viewed as multipoint IBD extensions of methods that average identity-by-state (IBS) or IBD results of individual markers (e.g., see

Table 5

Classification Proportions for 10,000 Putative Sib Pairs

No. of Chromosomes and Type of Classification	Proportion in Which True Relationship Is				
	Sib	Half-Sib	Unrelated	MZ Twin	Parent
5:[a]					
Sib	.9178	.0462	.0000	.0000	.0251
Half-sib	.0787	.8541	.0186	.0000	.0000
Unrelated	.0000	.0997	.9814	.0000	.0000
MZ twin	.0002	.0000	.0000	1.0000	.0000
Parent	.0033	.0000	.0000	.0000	.9749
10:[b]					
Sib	.9757	.0052	.0000	.0000	.0048
Half-sib	.0241	.9906	.0049	.0000	.0000
Unrelated	.0000	.0042	.9951	.0000	.0000
MZ twin	.0002	.0000	.0000	1.0000	.0000
Parent	.0000	.0000	.0000	.0000	.9952
22:[c]					
Sib	.9988	.0000	.0000	.0000	.0000
Half-sib	.0012	1.0000	.0000	.0000	.0000
Unrelated	.0000	.0000	1.0000	.0000	.0000
MZ twin	.0000	.0000	.0000	1.0000	.0000
Parent	.0000	.0000	.0000	.0000	1.0000

[a] Estimated OCV values are as follows: unrelated (−3.12) half-sib (−1.25) sib (3.27) MZ twin; parent (−1.71) sib.

[b] Estimated OCV values are as follows: unrelated (−4.43) half-sib (−1.80) sib (3.27) MZ twin; parent (−2.45) sib.

[c] Estimated OCV values are as follows: unrelated (−6.60) half-sib (−2.72) sib (3.27) MZ twin; parent (−3.69) sib.

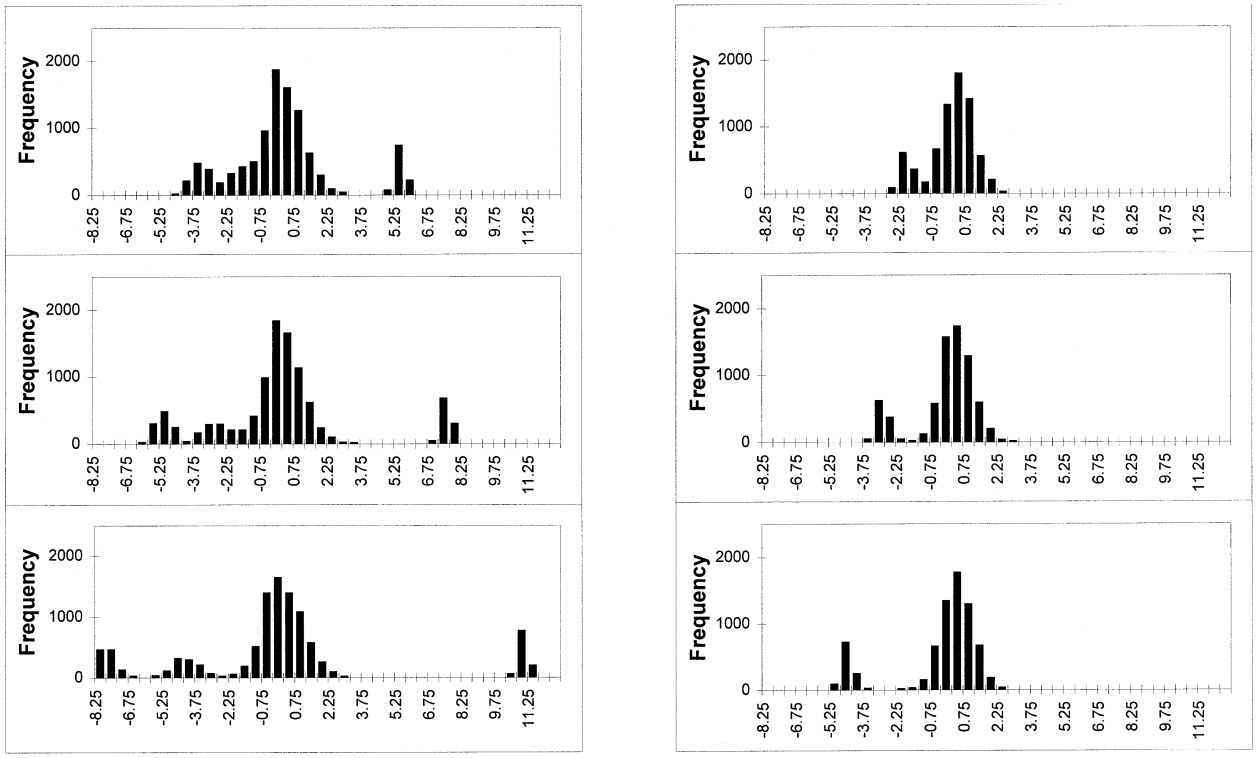

Figure 2 Histograms of Markov-process statistics: $\hat{Y}_j$ (*left*) and $\hat{Y}_j^*$ (*right*). From top to bottom, histograms contain results for 5, 10, and 22 chromosomes, respectively. For histograms on the right, only sib pairs and P/O pairs are included.

Chakraborty and Jin 1993*b;* Ehm and Wagner 1996). The methods use multipoint IBD calculations and are extremely powerful when genomewide autosomal marker data are available. The new methods require only one multipoint calculation per pedigree and thus save time compared with likelihood-ratio comparisons, such as those used by Boehnke and Cox (1997). In addition, the new methods can detect MZ twins (or duplicate samples) even if the possibility of genotyping error is not allowed in the multipoint calculations. A relative pair with different observed genotypes at even one marker will always be formally excluded as MZ twins when the likelihood-ratio method is used, unless more-complicated and time-consuming algorithms that allow for genotyping error are used (Boehnke and Cox 1997). Although the emphasis of the present study is on classification, the proposed statistics can be used to test the null hypothesis—that the pair is a sib pair—against the general alternative—that the pair is not a sib pair—if simply eliminating all nonsib pairs from the data set is the only goal. Another advantage is that a histogram of the relationship statistics gives the user a visual display that can be used as an exploratory tool or to confirm the classification results.

Their primary disadvantage is that the optimal classification criteria are functions of the informativity of the marker map and the amount of typed genome and thus depend on the particular marker set and population under study. However, I have shown, using simulations, that useful classification criteria can be chosen in parallel with computation of the statistics themselves, so that little or no extra effort is required. When more that half the genome is genotyped, misclassification is rare; when less than half of the genome is genotyped, then Bayesian methods may be preferred (e.g., see Goring and Ott 1997), particularly if detection of nonsib pairs is the goal. Alternatively, if the prior probability of nonsib pairs is small, then the $\hat{Y}_j$ distribution can simply be truncated at the extreme tail. For example, if 1% of pairs are expected to be half-sib pairs, then the lower 1% of the $\hat{Y}_j$ might be discarded. Severe asymmetric truncation of the sib $\hat{Y}_j$ distribution should be avoided, to avoid inflation of the number of false-positive linkage results.

As with all current methods of testing or classifying genetic relationships, the new methods are slightly biased in an affected-sib-pair study, since affected sib pairs are chosen for their presumed increased allele sharing at disease loci. In the context of a genome scan, this bias is small (Goring and Ott 1997). Calculations using the results of Feingold et al. (1993) show that the bias is negligible for relationship testing unless many trait loci, each with large effect, are present. To illustrate, assume

that five disease loci, each in the middle of a different chromosome and each giving a mean allele-sharing proportion of .6 at the location of a disease locus, contribute to a disease and that the variance of the $\hat{Y}_j$ is .8. Then $E(\hat{Y}_j) = .018$. If each of these five loci gives a mean allele-sharing proportion of 1, requiring that each locus be recessive and that each affected individual carry the disease homozygote at all five loci, then $E(\hat{Y}_j) = .09$. Such extreme cases are unlikely to occur in practice. For the vast majority of complex diseases, relationship testing under the assumption of random sampling of pairs and based on genomewide marker data can be safely performed.

The new methods of evaluation of genetic relationship will be implemented into the S.A.G.E. (version 4.0) computer software package. I recommend use of the entire genome, whenever possible, to classify relationships. At a minimum, using half the genome should guarantee a misclassification rate of $\leqslant 5\%$. Relationship estimation using small numbers of markers/chromosomes have high rates of misclassification and can discard large numbers of true sib pairs, particularly if prior probabilities are not taken into account. In theory, classification values that incorporate prior probabilities could be developed by standard approaches (e.g., see Johnson and Wichern 1998); for example, to classify sibs (s) and half-sibs (h) on the basis of their respective prior probabilities p_s and p_h, one could compare $p_s f_s(\hat{Y}_j)$ with $p_h f_h(\hat{Y}_j)$, where f_s and f_h are the (normal) density functions of $\hat{Y}_j$, where pairs are assumed to be sibs and half-sibs, respectively. The procedure could be modified further, to incorporate costs of misclassification. The means and variances of these densities depend on AMIC and T and could be obtained by regression formulas similar to those used to get OCV. Regardless of the method used to classify pairs, the ability to accurately distinguish sibs from half-sibs worsens faster, as the number of markers decreases, than the ability to accurately distinguish sibs from MZ twins, P/O pairs, or unrelated pairs. If the goal of relationship testing is to eliminate nonsib pairs, then a genomewide overall increase in allele sharing, as well as in false-positive evidence for linkage, can result if a large proportion of true sibs who happen by chance to share fewer alleles than expected genomewide are excluded from the data set.

Acknowledgments

The work was supported in part by U.S. Public Health Service grants HG01577 (from the National Center for Human Genome Research) and RR03655 (from the National Center for Research Resources). I thank Michael Boehnke for helpful discussion and thank Gerard Tromp for use of the intracranial aneurysm data. Some of the results in this study were obtained by the computer program MAPMAKER/SIBS (Kruglyak and Lander 1995).

References

Boehnke M, Cox NJ (1997) Accurate inference of relationships in sib-pair linkage studies. Am J Hum Genet 61:423–429

Chakraborty R, Jin L (1993*a*) Determination of relatedness between individuals using DNA fingerprinting. Hum Biol 65:875–895

Chakraborty R, Jin L (1993*b*) A unified approach to study hypervariable polymorphisms: statistical considerations of determining relatedness and population distances. In: Pena SDJ, Chakraborty R, Epplen JT, Jeffreys A (eds) DNA fingerprinting: state of the science. Birkhäuser-Verlag, Basel, pp 153–175

Ehm MG, Wagner M (1996) Test statistic to detect errors in sib-pair relationships. Am J Hum Genet Suppl 59:A217

Feingold E, Brown PO, Siegmund D (1993) Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. Am J Hum Genet 53:234–251

Goring HHH, Ott J (1997) Relationship estimation in affected sib pair analysis of late-onset diseases. Eur J Hum Genet 5: 69–77

Idury RM, Elston RC (1997) A faster and more general hidden Markov model algorithm for multipoint likelihood calculations. Hum Hered 47:197–202

Johnson RA, Wichern DW (1998) Applied multivariate statistical analysis. Prentice-Hall, Englewood Cliffs, NJ

Kosambi DD (1944) The estimation of map distances from recombination values. Ann Eugenics 12:172–175

Kruglyak L, Lander ES (1995) Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am J Hum Genet 57:439–454

Olson JM (1998) Discussion of Teng and Siegmund article. Biometrics 54:1266–1270

Parzen E (1962) Stochastic processes. Holden-Day, San Francisco

Ronkainen A, Hernesniemi J, Puranen M, Niemitukia L, Vanninen R, Ryynanen M, Kuivaniemi H, et al (1997) Familial intracranial aneurysms. Lancet 349:380–384

Stivers DN, Zhong Y, Hanis CL, Chakraborty R (1996) RELTYPE: a computer program for determining biological relatedness between individuals based on allele sharing at microsatellite loci. Am J Hum Genet Suppl 59:A190

Teng J, Siegmund D (1998) Multipoint linkage analysis using affected relative pairs and partially informative markers. Biometrics 54:1247–1265