Published in final edited form as:
Theor Popul Biol. 2008 May ; 73(3): 369–373.

The IBD process along four chromosomes

E. A. Thompson
Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195, U.S.A.

Abstract

In this note, we present the continuous-time Markov rate matrix that models identity by descent (*ibd*) patterns among four chromosomes in a population. The equilibrium distribution of this Markov process along a chromosome is the set of 4-gene state probabilities given by the Ewens sampling formula. This model will facilitate inference of identity by descent among the four chromosomes of a pair of individuals, using data at dense SNP loci among which there may be linkage disequilibrium.

Keywords

Identity by descent; continuous genome; multiple gene identities; Ewens sampling formula; linkage disequilibrium

1 Introduction

All methods of gene mapping rely, at some level, on the detection of genome shared *ibd* by individuals who are similar with regard to some trait whose genetic determinants are to be found. Thus some prior model for *ibd* is required, based either on known or hypothesized pedigree information, or on a model for coancestry in a population.

Traditionally, from Thompson (1975) to Anderson and Weir (2007), pairwise relationships between individuals have been estimated on the basis of data at unlinked loci. Even methods of *ibd* detection for linkage inferences have used widely spaced markers, such as those of a typical microsatellite genome scan at an average spacing of 5 centiMorgans or 5×10^6bp. (Weeks and Lange, 1988; Thompson, 2000). The increasing availability of dense SNP data, at an average spacing of 0.01cM or 10,000 bp (300K over the genome), provides potential for much more accurate identification of small segments of genome shared *ibd* between a pair, or among a set, of haplotypes.

At the pedigree level, in the absence of genetic interference, the sequence of inheritance vectors (Lander and Green, 1987) at locations of markers along a chromosome can be modeled as a first-order Markov chain. Where more remote relationships are not accurately known, a simple Markov model for the *ibd* between a pair of chromosomes can be used to analyze marker data and obtain probabilities conditional on these data (Leutenegger et al., 2003). Whereas the pointwise probability of *ibd* decreases as 2^{-m}, where m is the number of meioses separating a pair of individuals, the length of shared segments decreases only as m^{-1}. Use of sets of remotely

Address for correspondence: Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195, U.S.A., Phone: 206-685-0108, Fax: 206-686-7419, Email: eathomp@u.washington.edu.

related individuals, such as distant cousins, decreases the confounding effects of genetic heterogeneity of the trait relative to a population study, and decreases the confounding effects of shared environment relative to studies of close relatives.

The model of Leutenegger et al. (2003) is a continuous-time Markov process for *ibd* between two chromosomes in a population, where "time" is genetic distance along the chromosomes. This model can be used to analyze, for example, data on potentially inbred individuals, but analysis of sharing between affected relatives requires at least consideration of the four chromosomes of a pair of individuals. Although the marginal (single-locus) equilibrium probabilities of patterns of identity by descent among multiple chromosomes are well known (Balding and Nichols, 1994; Weir, 2001), an appropriate model for the changes in *ibd* patterns along a chromosome has proved elusive for more than a chromosome pair (Leutenegger, 2003).

Here we present the infinitesimal-rate matrix (Karlin and Taylor, 1975) for changes in *ibd* pattern among four chromosomes that has the correct population equilibrium probabilities of all 4-gene patterns of *ibd*. Thus this matrix serves as a prior Markov model for *ibd* along a chromosome for a pair of remote relatives from a population. Combined with hidden Markov models for population haplotypes such as that of Scheet and Stephens (2006) or of Browning (2006), it can serve as a basis for inferring *ibd* segments shared by individuals on the basis of dense SNP marker data among which there may be linkage disequilibrium (LD).

2 Identity along four chromosomes

2.1 The case of two chromosomes

As background, we first review briefly the model of Leutenegger et al. (2003) for *ibd* between a pair of chromosomes. In effect, that model is a continuous-time Markov process, with rate-matrix between the two states of no-*ibd* (0) and *ibd* (1)

$$\begin{pmatrix} -af & af \\ a(1-f) & -a(1-f) \end{pmatrix}.$$

In this model, the equilibrium marginal probability of *ibd* is f and the second parameter a determines the rate of change between the two states in units of genetic distance. The *ibd* segments are exponential with expected length $(a(1-f))^{-1}$, and the relative rate of gain vs loss of *ibd* is $f/(1-f)$. In fitting this model using marker data, Leutenegger et al. (2003) allowed for typing error, so that heterozygosity at a marker did not absolutely preclude *ibd*. With this proviso and with sufficiently dense polymorphic markers, it was found that *ibd* segments could be well identified, and that posterior probabilities of *ibd* were not very sensitive to the precise values of f and a. While these parameters could be estimated from the data, any reasonable choice gave similar results for purposes of identifying *ibd* segments. Although in reality a complex relationship between the parents of the individual result in segments of *ibd* of a variety of expected lengths, the single parameter a provides a sufficiently flexible model that *ibd* segments can be accurately imputed. For example, if relatively close relationships are suspected, a choice of $f = 0.1$ and $a = 0.1$ per centiMorgan (cM) is appropriate Leutenegger (2003).

For remote relationships, or for detecting *ibd* segments among chromosomes in a population, the overall level of *ibd* is at least an order of magnitude lower, and *ibd* segments are smaller. However, if the marginal prior probability of *ibd*, f, is chosen too small, *ibd* can be hard to detect: $f = 0.01$ provides a compromise. Even where *ibd* is *a priori* improbable, *ibd* segments are few rather than small. Even at 100 meioses (50 generations) separation, a shared segment is of expected length 1 cM, corresponding closely to a value $a = 1$ per cM. With dense marker data, segments of *ibd* much larger than the expected length are easily detected despite long lengths being *a priori* improbable. Further, due to the *a priori* exponential distribution of segment lengths, segments of less than the expected length are *a priori* quite probable and can also be detected given dense data. Thus the values $f = 0.01$ and $a = 1$ per cM are expected to work well for analysis of remote relationships at the population level.

2.2 The fifteen states for four chromosomes

Second, we review patterns of gene identity among four ordered genes. There are 15 patterns of gene identity, labeled by four digits (Thompson, 1974). Genes with the same label are *ibd* and those with different digit labels are not. For example, the state 1223 indicates that the second and third genes are *ibd*. If the four genes are the maternal and paternal genes of two individuals, then for the purposes of single-locus genotypes on the pair, the 15 states reduce to 9 state classes. These state classes are characterized by whether the 2 genes of the first individual are *ibd* (1 if so, 0 otherwise), similarly for the second individual, and by the number of distinct genes shared between the individuals. For example, the state class 101 consists of the two states 1112 and 1121. We will use this labeling of state classes in order to keep clear the distinction between the 15- and the 9-state representations. Although, for many purposes, the nine states are sufficient, if LD along a chromosome is to be modeled, it will be necessary to distinguish the two chromosomes within an individual. The 9 state-classes and corresponding 15 states are shown in Table 1.

2.3 Gene identity in an equilibrium population

Third, we review the single-locus multigene identity probabilities at equilibrium under an infinite alleles model (Ewens, 1972). In a sample of size n, let a_i be the number of allelic types present in i copies. Then the number of observed alleles is $k = \Sigma\, a_i$, $n = \Sigma\, ia_i$ and

$$P_n(a_1,\ldots,a_n)=\frac{n!}{\theta(\theta+1)\ldots(\theta+n-1)}\prod_{j=1}^{n}\left(\frac{\theta}{j}\right)^{a_j}\frac{1}{a_j!} \tag{1}$$

where θ is the population mutation rate parameter. Note that we are not here assuming the infinite alleles model for our observed marker data. We use this model only for the latent *ibd* among sampled genes. Under an infinite-alleles model, genes deriving from each mutation event are identical by descent. Under the model (1) the coancestry coefficient β is given by

$$\beta=P_2(a_2=1)=\frac{2}{\theta(\theta+1)}\frac{\theta}{2}=\frac{1}{1+\theta}$$

Ewens' sampling formula (1) may be rewritten in terms of coancestry β as

$$P_n(a_1,\ldots,a_n)=\frac{n!\beta^{n-k}(1-\beta)^{k-1}}{(1+\beta)(1+2\beta)\ldots.(1+(n-2)\beta)}\prod_{j=1}^{n}(j^{a_j}a_j!)^{-1} \tag{2}$$

We follow Balding and Nichols (1994) in expressing higher order gene identities in terms of β by considering equilibrium under an infinite alleles model using a coalescent framework. However, rather than the parametrization of Balding and Nichols (1994) we use equation (2) directly to obtain the probabilities of the 15 states. For convenience, write $\eta = \beta(1 - \beta)/(1 + \beta)(1 + 2\beta)$.

Then

$$P_4(a_4=1)=6\beta^2\eta/(1-\beta);\qquad P_4(a_2=2)=3\beta\eta$$
$$P_4(a_3=a_1=1)=8\beta\eta;\qquad P_4(a_1=2,a_2=1)=6\eta(1-\beta)$$
$$\text{and}\quad P_4(a_1=4)=\eta(1-\beta)^2/\beta.$$

Since all configurations with the same values of a_j have the same probability, the probabilities in the last two columns of Table 1 immediately follow.

2.4 State-changes along a continuous genome

We can now present the continuous-chromosome rate-matrix, Q, that has these required equilibrium probabilities. From each of the 15 states, we consider the ways in which *ibd* can be gained (rate g) or lost (rate h). The resulting matrix is shown in Table 2. Consider, for example, the state 1213. The pair of *ibd* genes can lose identity at rate h, leading to state 1234. On the other hand *ibd* may be gained in several ways: the gene labeled 2 may become *ibd* to either of the two labeled 1, giving state 1112 (rate $2g$); the gene labeled 3 may become *ibd* to either of the two labeled 1, giving state 1211 (rate $2g$); the genes labeled 2 and 3 may become *ibd* giving state 1212 (rate g). Transition rates from other states with three distinct genes may be derived similarly. As another example consider the state 1111. Any one of the four genes may lose *ibd* with the remainder, leading to rate h changes to each of states 1112, 1121, 1211, and 1222.

Because we are considering change rates over infinitesimal chromosome lengths, only single events need to be considered. For example, there is 0 rate of switching from 1122 to 1234, since this would require two separate losses of *ibd*. Additionally, since in the infinite alleles model new alleles arise a singletons, here only singletons can "coalesce" to become *ibd* either with each other or with a larger group. Hence the 3 states with $a_2 = 2$ cannot instantaneously gain *ibd*. A transition must be made via one of the states with $a_1 = 2$, $a_2 = 1$. On the other hand, the other states with two distinct genes ($a_1 = a_3 = 1$) can gain *ibd*, moving at rate $3g$ to the state $a_4 = 1$, since the singleton gene can, without constraint, become *ibd* to any of the other three.

In Table 2 the detailed 15 states label the rows, while for clarity the nine state classes are used to label groups of corresponding columns. For reasons of space the diagonal terms are omitted. Since each row sum of any rate-matrix is zero (Karlin and Taylor, 1975), they are easily determined as $-4h$ for state 1111; $-2h$ for states 1122, 1212 and 1221, $-3h -3g$ for states 1112, 1121, 1211 and 1222; $-h -5g$ for states 1123, 1233, 1213, 1231, 1223 and 1232; and $-6g$ for state 1234.

The equilibrium probabilities of this rate-matrix must satisfy $\pi Q = 0$ (Karlin and Taylor, 1975), and are a function of $g/h = \alpha$ only. Further, the equilibrium probability for states with k distinct genes have relative probabilities of order α^{-k}. It is easily shown that this rate-matrix has equilibrium probabilities

$$\pi(1122)=\pi(1212)=\pi(1221)$$
$$\pi(1112)=\pi(1211)=\pi(1222)=\pi(1121)$$
$$\pi(1123)=\pi(1233)=\pi(1231)=\pi(1213)=\pi(1223)=\pi(1232)$$

and further that

$$\begin{aligned} \pi(1111) &= 3\alpha \quad \pi(1112) \\ \pi(1112) &= 2 \quad \pi(1122) \\ \pi(1122) &= \alpha \quad \pi(1123) \\ \text{and} \quad \pi(1123) &= \alpha \quad \pi(1234) \end{aligned}$$

in agreement with the equilibrium probabilities of Table 1 with $g/h = \alpha = \beta/(1 - \beta)$.

2.5 The nine-state *ibd* model

In the event that paternal and maternal chromosomes need not be distinguished, and in particular if analysis is in the absence of LD, we can reduce for the 15-states to the 9 genotypically distinguishable classes. The relevant matrix is shown in Table 3, and is obtained simply by collapsing the 15-state matrix into the 9 classes. Since within any state class there are no transitions among the states, the rate of leaving the class is the same as that of leaving each state in the class, and transitions occur to the classes corresponding to the states in Table 2. For example, the two states 1212 and 1221 are each left at rate 2h, and so also is the combined state class 002 = {1212, 1221}. As discussed above the only possible transitions from this class is to the class 001, which therefore occurs at rate 2*h*. Likewise all four states in the class 001 = {1213, 1231, 1223, 1232} are left at rate 5*g* + *h* and so also is the class, and the transitions to other classes at rates 2*g*, 2*g*, *g* and *h* follow: compare rows 11 to 14 of Table 2 with the penultimate row of Table 3.

Note that only certain transitions are possible in a single step, but since this is a continuous-time process all transitions are possible between any two distinct points of the genome. The equilibrium solution of the process has

$$\begin{aligned} \pi(101) &= \pi(011) = 2\pi(002) = 4\pi(110), \\ \pi(001) &= 4\pi(100) = 4\pi(010), \\ \pi(111) &= \tfrac{3}{4}\alpha \quad (\pi(101)+\pi(011)) = 6\alpha\,\pi(110), \\ \text{and} \quad \pi(000) &= (6\alpha)^{-1}(\pi(100)+\pi(010)+\pi(001)) = \alpha^{-1}\pi(100). \end{aligned}$$

Indeed, the solution is that of Table 1, again with $g/h = \alpha = \beta/(1 - \beta)$. Note α is the analogue of the ratio $f/(1 - f)$ for the two-state model.

3 Discussion

Markov models, and more generally hidden Markov models (HMM), are pervasive in complex stochastic systems, not least because they permit likelihood and probability computations (Baum et al., 1970; Baum, 1972). The continuous chromosome *ibd* model of section 2.4 can be used as an HMM underlying genetic marker data on pairs of relatives, and allow the inference of small segments of genome shared *ibd* from data on dense SNP markers at a spacing of a few thousand base pairs (bp). However, at this scale, failure to take linkage disequilibrium (LD) into account in modeling haplotype frequencies would lead to severe biases. Haplotypes that are common in the population would be modeled as rare, through multiplication of allele frequencies over loci, leading to false inference of *ibd*. Other haplotypes, not existing in the population, would be given positive frequencies, and might be falsely imputed in the case of unphased genotypic data.

For likelihood computations on small pedigrees (Abecasis and Wigginton, 2005), the approach of clustered SNPs has been used to accommodate LD among marker loci. That is, SNPs in high LD are combined into “super-markers” within which no recombination is permitted. Between clusters there is assumed to be no LD. For simple relationships, such as sib pairs, an alternative approach is to incorporate LD in parental haplotypes, using a Markov model or HMM for the allelic types along a chromosome. For example, the cluster-based hidden Markov model for population haplotypes of Scheet and Stephens (2006) may be combined with the hidden Markov model of inheritance vectors (Lander and Green, 1987) to provide a tractable model for lod-score or *ibd*-based methods of linkage inference in the presence of LD in parental haplotypes (Fu and Thompson, 2007).

While for small pedigrees the dichotomy of no-LD or no-recombination will often suffice, it is less clear that this is so for remote relatives separated by many meioses. A more flexible model allowing both recombination and LD is desirable. Just as for the Markov inheritance vectors, the Markov model of Section 2.4 for latent *ibd* can be combined with hidden Markov LD-models for population haplotypes (Scheet and Stephens, 2006; Browning, 2006) to provide a framework for inference of *ibd* segments in a pair of remote relatives, using data at dense SNP markers exhibiting LD.

Acknowledgment

This research was supported in part by NIH grant GM46255.

References

Abecasis G, Wigginton J. Handling marker-marker linkage disequilibrium: Pedigree analysis with clustered markers. American Journal of Human Genetics 2005;77:754–767. [PubMed: 16252236]

Anderson AD, Weir BS. A maximum-likelihood method for the estimation of pairwise relatedness in structured populations. Genetics 2007;176:421–440. [PubMed: 17339212]

Balding DJ, Nichols RA. DNA profile match probability calculations: How to allow for population stratification, relatedness, database selection, and single bands. Forensic Science Int 1994;64:125–140.

Baum, LE. In: Shisha, O., editor. An inequality and associated maximization technique in statistical estimation for probabilistic functions on Markov processes; Inequalities-III; Proceedings of the Third Symposium on Inequalities; University of California Los Angeles, 1969. New York: Academic Press; 1972. p. 1-8.

Baum LE, Petrie T, Soules G, Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions on Markov chains. Annals of Mathematical Statistics 1970;41:164–171.

Browning SR. Multilocus association mapping using variable-length Markov chains. American Journal of Human Genetics 2006;78:903–913. [PubMed: 16685642]

Ewens WJ. The sampling theory of selectively neutral alleles. Theoretical Population Biology 1972;3:87–112. [PubMed: 4667078]

Fu, AQ.; Thompson, EA. Technical report # 519. University of Washington: Department of Statistics; 2007. Inference of identity-by-descent in sib pairs: Analysis with and without linkage disequilibrium.

Karlin, S.; Taylor, HM. A First Course in Stochastic Processes. 2 nd edition. New York, NY: Academic Press; 1975.

Lander ES, Green P. Construction of multilocus genetic linkage maps in humans. Proceedings of the National Academy of Sciences (USA) 1987;84(8):2363–2367.

Leutenegger A, Prum B, Genin E, Verny C, Clerget-Darpoux F, Thompson EA. Estimation of the inbreeding coefficient through use of genomic data. American Journal of Human Genetics 2003;73:516–523. [PubMed: 12900793]

Leutenegger, AL. Estimation of random genome sharing: Consequences for linkage detection. Ph.D. Thesis. University of Washington and Université Paris-Sud 11; 2003.

Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. American Journal of Human Genetics 2006;78:629–644. [PubMed: 16532393]

Thompson EA. Gene identities and multiple relationships. Biometrics 1974;30:667–680. [PubMed: 4429760]

Thompson EA. The estimation of pairwise relationship. Annals of Human Genetics 1975;39:173–188. [PubMed: 1052764]

Thompson EA. MCMC estimation of multi-locus genome sharing and multipoint gene location scores. International Statistical Review 2000;68:53–73.

Weeks DE, Lange K. The affected pedigree member method of linkage analysis. American Journal of Human Genetics 1988;42:315–326. [PubMed: 3422543]

Weir, BS. Forensics. Wiley, New York: Handbook of Statistical Genetics; 2001. p. 721-739.

Table 1

The state classes among 4 genes

State class	*ibd* states	class characterization	equilibrium state probability	equilibrium class probability
1	1111	1 1 1	$6\beta^2\eta/(1-\beta)$	—
2	1122	1 1 0	$\beta\eta$	—
3	1112 and 1121	1 0 1	$2\beta\eta$	$4\beta\eta$
4	1123	1 0 0	$\eta(1-\beta)$	—
5	1211 and 1222	0 1 1	$2\beta\eta$	$4\beta\eta$
6	1233	0 1 0	$\eta(1-\beta)$	—
7	1212 and 1221	0 0 2	$\beta\eta$	$2\beta\eta$
8	1213, 1231, 1223 and 1232	0 0 1	$\eta(1-\beta)$	$4\eta(1-\beta)$
9	1234	0 0 0	$\eta(1-\beta)^2/\beta$	—

Table 2

Transitions among the fifteen *ibd* states between four chromosomes.

ibd	One or both individuals inbred								non-inbred states						
state	**111**	**110**	**101**		**100**	**011**		**101**	**002**		**001**				**000**
1111	—	0	h	h	0	h	h	0	0	0	0	0	0	0	0
1122	0	—	0	0	h	0	0	h	0	0	0	0	0	0	0
1112	3g	0	—	0	h	0	0	0	0	0	h	0	h	0	0
1121	3g	0	0	—	h	0	0	0	0	0	0	h	0	h	0
1123	0	g	2g	2g	—	0	0	0	0	0	0	0	0	0	h
1211	3g	0	0	0	0	—	0	h	0	0	h	h	0	0	0
1222	3g	0	0	0	0	0	—	h	0	0	0	0	h	h	0
1233	0	g	0	0	0	2g	2g	—	0	0	0	0	0	0	h
1212	0	0	0	0	0	0	0	0	—	0	h	0	0	h	0
1221	0	0	0	0	0	0	0	0	0	—	0	h	h	0	0
1213	0	0	2g	0	0	2g	0	0	g	0	—	0	0	0	h
1231	0	0	0	2g	0	2g	0	0	0	g	0	—	0	0	h
1223	0	0	2g	0	0	0	2g	0	0	g	0	0	—	0	h
1232	0	0	0	2g	0	0	2g	0	g	0	0	0	0	—	h
1234	0	0	0	0	g	0	0	g	0	0	g	g	g	g	—

For reasons of space, the diagonal terms are omitted. They are such that each row sums to 0.

Table 3

Transitions among the nine *ibd* states between four chromosomes

***ibd* state**	**inbred-states**						**non-inbred**		
111	−4h	0	2h	0	2h	0	0	0	0
110	0	−2h	0	h	0	h	0	0	0
101	3g	0	−3h-3g	h	0	0	0	2h	0
100	0	g	4g	−h-5g	0	0	0	0	h
011	3g	0	0	0	−3h-3g	h	0	2h	0
010	0	g	0	0	4g	−h-5g	0	0	h
002	0	0	0	0	0	0	−2h	2h	0
001	0	0	2g	0	2g	0	g	−h-5g	h
000	0	0	0	g	0	g	0	4g	−6g