Evolution, 57(7), 2003, pp. 1465–1477

THE SHAPES OF NEUTRAL GENE GENEALOGIES IN TWO SPECIES: PROBABILITIES OF MONOPHYLY, PARAPHYLY, AND POLYPHYLY IN A COALESCENT MODEL

NOAH A. ROSENBERG
Molecular and Computational Biology, University of Southern California, 1042 West 36th Place, DRB 289, Los Angeles, California 90089
E-mail: noahr@usc.edu

Abstract.—The genealogies of samples of orthologous regions from multiple species can be classified by their shapes. Using a neutral coalescent model of two species, I give exact probabilities of each of four possible genealogical shapes: reciprocal monophyly, two types of paraphyly, and polyphyly. After the divergence that forms two species, each of which has population size N, polyphyly is the most likely genealogical shape for the lineages of the two species. At $\sim 1.300N$ generations after divergence, paraphyly becomes most likely, and reciprocal monophyly becomes most likely at $\sim 1.665N$ generations. For a given species, the time at which 99% of its loci acquire monophyletic genealogies is $\sim 5.298N$ generations, assuming all loci in its sister species are monophyletic. The probability that all lineages of two species are reciprocally monophyletic given that a sample from the two species has a reciprocally monophyletic genealogy increases rapidly with sample size, as does the probability that the most recent common ancestor (MRCA) for a sample is also the MRCA for all lineages from the two species. The results have potential applications for the testing of evolutionary hypotheses.

Key words.—Coalescence, gene tree, labeled histories, population divergence, speciation, Yule model.

Received September 7, 2002. Accepted February 4, 2003.

The genealogy for all copies of a particular genomic region in a particular species can be classified into one of two categories. Either the lineages are *monophyletic*, that is, they comprise *all* the extant descendants of their most recent common ancestor (MRCA), or they are not monophyletic. The latter scenario, equivalently, requires that lineages of one or more additional species also descend from this MRCA.

Consider all copies of orthologous regions in an ordered pair of species, (A, B). The genealogy of these lineages can be placed into one of four categories: $C1$, the lineages of each species are separately monophyletic; $C2$, the lineages of species A are monophyletic, and the lineages of species B are not monophyletic; $C3$, the lineages of species B are monophyletic, and the lineages of species A are not monophyletic; and $C4$, Neither the lineages of species A nor the lineages of species B are monophyletic.

These scenarios (Fig. 1) can also be labeled *monophyly of A and B* or *reciprocal monophyly, paraphyly of B with respect to A, paraphyly of A with respect to B,* and *polyphyly of A and B,* respectively. The lineages of species A, for example, are monophyletic in $C1$ and $C2$, paraphyletic with respect to B in $C3$, and polyphyletic with respect to B in $C4$. In this article these terms are used to describe both a set of lineages and the unique genealogy for the set of lineages; for example, a species is described as monophyletic if the genealogy of all of its lineages is monophyletic. When referring to a locus, the terms implicitly describe the genealogy of all copies of the locus in a species. Two *genealogies* are identical if and only if they have both the same branching order and the same times of branching.

Typically, random pairs of species have reciprocally monophyletic genealogies at nearly all of their orthologous loci. However, the lineages of two closely related species are often not reciprocally monophyletic (Takahata and Nei 1985; Neigel and Avise 1986; Palumbi et al. 2001; Hudson and Coyne 2002). For true biological species that have not exchanged migrants, lack of reciprocal monophyly requires many lineages ancestral to extant lineages to have been present in ancestral species. These ancestral lineages must have coalesced in an order that included more than one interspecific coalescence (Fig. 1ii–iv). Because recently diverged pairs of species might be represented by multiple ancestral lineages at their times of divergence, their genealogies can have many interspecific coalescences and, thus, might not show reciprocal monophyly.

Probabilities of genealogical shapes for two species have been studied in various population structure models for samples of size 2 (Tajima 1983; Takahata and Nei 1985; Takahata and Slatkin 1990; Wakeley 2000). Some more general cases are treated in simulations (Neigel and Avise 1986; Hudson and Coyne 2002) and in a few theoretical results (Brown 1994; Rosenberg 2002). Here I allow samples of arbitrary size from unstructured species and obtain exact probabilities of each of the four genealogical shapes for the lineages of the two species. Because only the shapes of genealogies are studied, underlying relationships masked by stochastic differences in the numbers of mutations that accumulate along different lineages are not considered. This issue has been treated in related circumstances (Hey 1991; Clark 1997; Wakeley and Hey 1997).

I first review models used in the calculations. The main results, namely the probabilities of the four genealogical shapes, are presented in equations (14)–(17). The properties of these probabilities are then explored in detail.

THE COALESCENT MODEL: ONE SPECIES

I assume a neutral coalescent model to describe the genealogy of the lineages of each of two species (Nordborg 2001). For each species, n lineages, each of which is identified by a distinctive label, are sampled in the present, and the model provides a probability distribution on the set of possible genealogies for these lineages. Going backward in time, the coalescent model has two components: a model that spec-

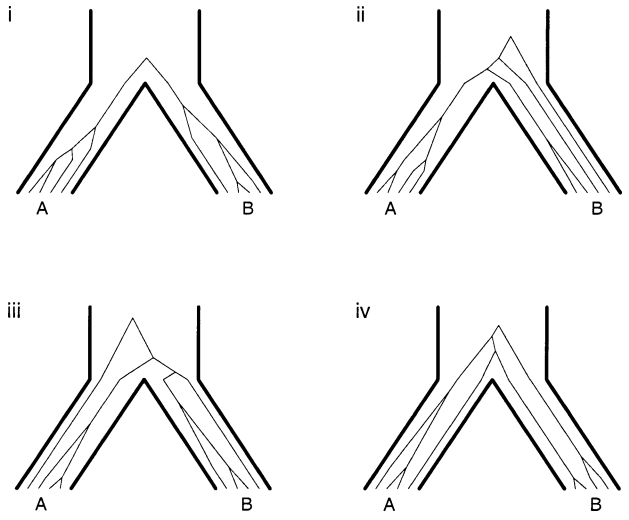


FIG. 1. The four types of genealogies possible for the lineages of species *A* and *B*. (i) Monophyly of *A* and *B*. (ii) Paraphyly of *B* with respect to *A*. (iii). Paraphyly of *A* with respect to *B*. (iv) Polyphyly of *A* and *B*.

ifies the distributions of waiting times until coalescence and a model that specifies coalescence probabilities for pairs of lineages. The latter is the Yule model (see next section). The waiting time until coalescence of n to $n - 1$ lineages is exponentially distributed with mean $n(n - 1)/2$ coalescences per unit of time. A useful result under the coalescent model is given by Tavaré (1984, eq. 6.1): if $g_{nj}(T)$ is the probability that the n lineages derive from j lineages that existed T *coalescent time units* in the past, then

$$g_{nj}(T) = \sum_{k=j}^{n} e^{-k(k-1)T/2} \frac{(2k-1)(-1)^{k-j} j_{(k-1)} n_{[k]}}{j!(k-j)! n_{(k)}}, \qquad (1)$$

where $a_{(k)} = a(a+1)\cdots(a+k-1)$ and $a_{[k]} = a(a-1)\cdots(a-k+1)$ for $k \geq 1$, with $a_{(0)} = a_{[0]} = 1$. Except when $1 \leq j \leq n$, $g_{nj}(T) = 0$. Note that $g_{nj}(0) = \delta_{nj}$, where δ_{nj} is Kronecker's delta. The limiting case of $n = \infty$ yields (Tavaré 1984, eq. 6.3):

$$g_{\infty j}(T) = \sum_{k=j}^{\infty} e^{-k(k-1)T/2} \frac{(2k-1)(-1)^{k-j} j_{(k-1)}}{j!(k-j)!}. \qquad (2)$$

For a population of constant size N lineages, in which the variance of the number of offspring produced by an individual equals one, T coalescent units equals TN generations. For models that have different values of this parameter or that incorporate any of various types of more complex reproductive behavior, equations (1) and (2) still hold, but a different scaling of coalescent units into units of absolute time is used (Nordborg 2001). Thus, results that follow can be applied for diverse evolutionary models within species, only with changes of scale to the time parameters. I allow two species to experience different amounts of coalescent time during the same period of absolute time; in the constant-size model, this assumption corresponds to different population sizes for the two species.

Although approximations can assist in the computation (Griffiths 1984), equations (1) and (2) can be difficult to evaluate numerically for small positive values of T, especially for large n and j. As T increases, however, the number of ancestral lineages is likely quite small, so that $g_{nj}(T)$ is negligible for all but very small values of j. In this article, $g_{nj}(T)$ is only evaluated for $T \geq 0.1$, where evaluation is straightforward (and for $T = 0$, where it is trivial). The following assumption, which has no impact on the first several decimal places of results shown, is also made: $g_{nj}(T) = 0$ if $T \geq 0.1$, $n \geq 90$, and $j \geq 50$, or if $T \geq 1$, $n \geq 20$, and $j \geq 10$.

THE YULE MODEL

In the coalescent model, at any time in the past, all pairs of lineages have equal probabilities of being the next pair to coalesce. The choice of which lineages coalesce is independent of when the coalescence occurs. This rule specifies a uniform probability distribution on the set of possible sequences of coalescences, and has been termed the *Markov model* or *Yule model* (Yule 1924; Aldous 2001); the sequence of coalescences, from n lineages to the MRCA of the lineages, is a ''random-joining sequence'' (Maddison and Slatkin 1991). The Yule model has been used to describe the branching tree not only for lineages within species, but also for species themselves (Harding 1971; Slowinski and Guyer 1989; Maddison and Slatkin 1991; Brown 1994; Aldous 2001; Steel and McKenzie 2001).

Sequences of Coalescences

The set of possible sequences of coalescences for n lineages is equivalent to the set of *labeled histories* for n taxa. Stated precisely, two genealogies have the same labeled history if and only if they have the same coalescences in the same temporal order. Under the Yule model, each of the possible sequences of coalescences has the same probability of being the labeled history for a random genealogy. The number of possible labeled histories for n lineages is obtained from the fact that there are $\binom{n}{2}$ choices for the two most recent lineages to coalesce, $\binom{n-1}{2}$ choices for the next coalescence, and so forth. The total number of labeled histories, H_n, equals (Edwards 1970):

$$H_n = \binom{n}{2}\binom{n-1}{2}\cdots\binom{3}{2}\binom{2}{2} = \frac{n!(n-1)!}{2^{n-1}}. \qquad (3)$$

More generally, the number of sequences of coalescences that reduce n lineages to k lineages is

$$I_{n,k} = \binom{n}{2}\binom{n-1}{2}\cdots\binom{k+2}{2}\binom{k+1}{2} = \frac{n!(n-1)!}{2^{n-k}k!(k-1)!}. \qquad (4)$$

As a special case, $I_{n,1} = H_n$.

The Interweaving of Two Sequences of Coalescences

Suppose that a particular sequence S_1 of s_1 coalescences occurs for a set of $s_1 + 1$ lineages, and that a particular sequence S_2 of s_2 coalescences occurs for a different set of

$s_2 + 1$ lineages. The number of sequences of $s_1 + s_2$ coalescences that include both S_1 and S_2 as subsequences is of interest. There are $\binom{s_1 + s_2}{s_1}$ ways to choose which positions in a sequence of length $s_1 + s_2$ are occupied by the subsequence S_1. Once these positions are chosen, the subsequence S_1 fills them in a unique way, with its first coalescence in the first position chosen, its second coalescence in the second position, and so forth. The subsequence S_2 fills the remaining positions in a unique way. Thus, the number of ways that s_1 and s_2 coalescences can be ''interwoven,'' preserving the order of coalescence in each subset is:

$$W_2(s_1, s_2) = \binom{s_1 + s_2}{s_1}. \tag{5}$$

Sequences of Coalescences for Subsamples

Consider n lineages that have a randomly chosen labeled history, and a random subsample consisting of j of these lineages. If $Z_1(T)$ and $Z_2(T)$ respectively denote the numbers of lineages ancestral to the sample and the subsample at time T coalescent units before the present, then (eq. 2.3 of Saunders et al. 1984)

$$\begin{aligned} S(l_2, l_1, n, j) &= \Pr[Z_2(T) = l_2 \,|\, Z_1(T) = l_1; n, j] \\ &= \frac{\binom{l_1}{l_2}\binom{n - l_1}{j - l_2}\binom{n + l_2}{n}}{\binom{n}{j}\binom{j + l_1}{j}} \frac{n l_2 (j + l_1)}{(n + l_2) l_1 j}. \end{aligned} \tag{6}$$

The limit as $n \to \infty$, is (Griffiths and Tavaré, 2003):

$$S(l_2, l_1, \infty, j) = \frac{\binom{l_1}{l_2}\binom{j}{l_2}}{\binom{j + l_1}{j}} \frac{l_2 (j + l_1)}{l_1 j}. \tag{7}$$

The probability that the MRCA of the subsample is identical to the MRCA of the sample is (eq. 3.1 of Saunders et al. 1984; see also Sanderson 1996)

$$\begin{aligned} &\Pr[\min\{T : Z_2(T) = 1\} = \min\{T : Z_1(T) = 1\}] \\ &\quad = \frac{j - 1}{j + 1} \frac{n + 1}{n - 1}. \end{aligned} \tag{8}$$

Theory: Two Species

Here I consider probabilities of genealogical shapes in a divergence model. At time t years in the past, a species evolving according to the coalescent model diverges into two species, A and B, each of which also evolves by the coalescent model (Fig. 2). For species A and B, T_A and T_B coalescent time units have elapsed since the divergence, respectively. Total population sizes in the present are N_A and N_B; the ancestral population size need not be specified. Samples from species A and B have sizes r_A and r_B. Both species are assumed to have the same generation time.

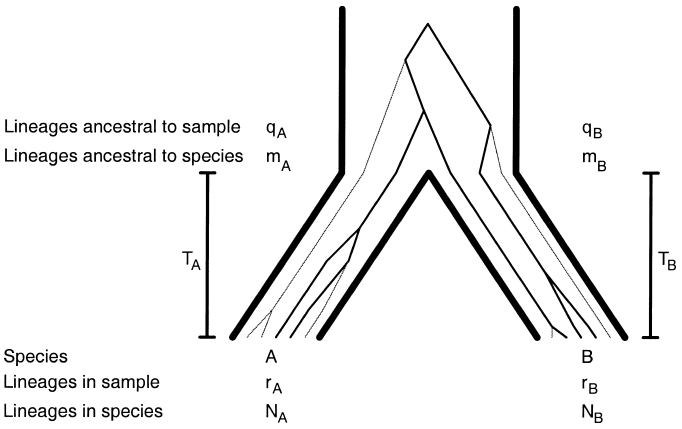


FIG. 2. Notation for the divergence model. Lineages ancestral to a sample are drawn more darkly than non-ancestral lineages. In this example, $N_A = 5$, $r_A = 2$, $m_A = 2$, $q_A = 1$; $N_B = 5$, $r_B = 3$, $m_B = 3$, $q_B = 2$. Times T_A and T_B are measured in coalescent units and refer to the same period of absolute time.

Probabilities of Monophyly, Paraphyly, and Polyphyly

Let X denote the classification of the $r_A + r_B$ sampled lineages (either $C1$, $C2$, $C3$, or $C4$) and let $Z_A(T)$ and $Z_B(T)$ denote the numbers of lineages ancestral to species A and B, respectively, at time T coalescent units in the past. Suppose the r_A and r_B lineages of species A and B have q_A and q_B ancestral lineages at the time of divergence. This event has probability $g_{r_A,q_A}(T_A) g_{r_B,q_B}(T_B)$.

To have $X = C1$, the MRCA for the $q_A + q_B$ ancestral lineages must separate the genealogy of the lineages into two species-specific subgenealogies. Using equation (3), H_{q_A} and H_{q_B} sequences of coalescences can monophyletically join the lineages of species A and B, respectively. Using equation (5), each pair of sequences—one for species A and one for species B—can be interwoven in $W_2(q_A - 1, q_B - 1)$ ways. There are $H_{q_A+q_B}$ possible sequences of coalescences. We obtain (Brown 1994):

$$\begin{aligned} &\Pr[X = C1 \,|\, Z_A(T_A) = q_A, Z_B(T_B) = q_B] \\ &\quad = \frac{H_{q_A} H_{q_B} W_2(q_A - 1, q_B - 1)}{H_{q_A+q_B}} \\ &\quad = \frac{2}{\binom{q_A + q_B}{q_A}(q_A + q_B - 1)}. \end{aligned} \tag{9}$$

This equation solves the recursion for the quantity $C(q_A, q_B)$ of Hudson and Coyne (2002) and for the quantity $1 - F_2^{A,B}(q_A, q_B, 0)$ of Rosenberg (2002).

To compute conditional probabilities that the $q_A + q_B$ lineages fall into the other categories, we first need the conditional probability that the lineages of species A are monophyletic, that is, the probability that $X = C1$ or $X = C2$. Let E_k be the event that the q_A lineages of species A are monophyletic and that the most recent interspecific coalescence occurs when k lineages ancestral to the lineages of species B are present ($1 \le k \le q_B$). The number of sequences that coalesce q_A lineages to one lineage is H_{q_A}. The number of

sequences that coalesce q_B lineages to k lineages is $I_{q_B,k}$. The number of ways to interweave a sequence of coalescences from species A and one from species B is $W_2(q_A - 1, q_B - k)$. There are k choices for the lineage from species B that participates in the coalescence with the MRCA of species A, and there are H_{k-1} ways to coalesce the last $k - 1$ lineages. Thus, we have

$$\Pr(E_k) = \frac{H_{q_A} I_{q_B,k} W_2(q_A - 1, q_B - k)\, k H_{k-1}}{H_{q_A+q_B}} = \frac{2k\binom{q_B}{k}}{q_B\binom{q_A+q_B}{q_A}\binom{q_A+q_B-1}{k}}. \quad (10)$$

A combinatorial identity (Appendix 1) facilitates computation of the probability of E, the event that the q_A lineages of species A are monophyletic:

$$\Pr(E) = \sum_{k=1}^{q_B} \Pr(E_k) = \frac{2}{q_B\binom{q_A+q_B}{q_A}} \sum_{k=1}^{q_B} \frac{k\binom{q_B}{k}}{\binom{q_A+q_B-1}{k}} = \frac{2}{\binom{q_A+q_B}{q_A}} \frac{q_A+q_B}{q_A(q_A+1)}. \quad (11)$$

This result simplifies a calculation of Brown (1994, eq. 12) and gives the closed-form solution for the quantity $C^*(q_A, q_B)$ of Hudson and Coyne (2002). The probability of F, the event that the q_B lineages of species B are monophyletic, is computed analogously. We now obtain

$$\begin{aligned}\Pr[X = C2 \,|\, Z_A(T_A) = q_A, Z_B(T_B) = q_B] &= \Pr(E) - \Pr[X = C1 \,|\, Z_A(T_A) = q_A, Z_B(T_B) = q_B] \\ &= \frac{2}{\binom{q_A+q_B}{q_A}} \frac{(2q_A+q_B)(q_B-1)}{q_A(q_A+1)(q_A+q_B-1)}. \end{aligned} \quad (12)$$

A similar expression gives the conditional probability of $X = C3$, and the conditional probability of $X = C4$ is one minus the sum of the other three probabilities. This probability can also be obtained as follows:

$$\begin{aligned}\Pr[X = C4 \,|\, Z_A(T_A) = q_A, Z_B(T_B) = q_B] &= 1 - \Pr(E) - \Pr(F) + \Pr(E \cap F) \\ &= 1 - \frac{2}{\binom{q_A+q_B}{q_A}}\left[\frac{q_A+q_B}{q_A(q_A+1)} + \frac{q_A+q_B}{q_B(q_B+1)} - \frac{1}{q_A+q_B-1}\right]. \end{aligned} \quad (13)$$

Finally, summing over possible values of q_A and q_B, the unconditional probabilities are

$$\Pr(X = C1) = \sum_{q_A=1}^{r_A}\sum_{q_B=1}^{r_B} g_{r_A,q_A}(T_A) g_{r_B,q_B}(T_B) \times \frac{2}{\binom{q_A+q_B}{q_A}(q_A+q_B-1)} \quad (14)$$

$$\Pr(X = C2) = \sum_{q_A=1}^{r_A}\sum_{q_B=1}^{r_B} g_{r_A,q_A}(T_A) g_{r_B,q_B}(T_B) \times \frac{2}{\binom{q_A+q_B}{q_A}} \frac{(2q_A+q_B)(q_B-1)}{q_A(q_A+1)(q_A+q_B-1)} \quad (15)$$

$$\Pr(X = C3) = \sum_{q_A=1}^{r_A}\sum_{q_B=1}^{r_B} g_{r_A,q_A}(T_A) g_{r_B,q_B}(T_B) \times \frac{2}{\binom{q_A+q_B}{q_A}} \frac{(q_A+2q_B)(q_A-1)}{q_B(q_B+1)(q_A+q_B-1)} \quad (16)$$

$$\Pr(X = C4) = \sum_{q_A=1}^{r_A}\sum_{q_B=1}^{r_B} g_{r_A,q_A}(T_A) g_{r_B,q_B}(T_B) \times \left\{1 - \frac{2}{\binom{q_A+q_B}{q_A}}\left[\frac{q_A+q_B}{q_A(q_A+1)} + \frac{q_A+q_B}{q_B(q_B+1)} - \frac{1}{q_A+q_B-1}\right]\right\}. \quad (17)$$

Appendix 2 discusses an alternate derivation of (14). Note that if $T_A = T_B = 0$, then $q_A = r_A$, $q_B = r_B$, and the unconditional probabilities reduce to equations (9), (12), and (13). The probabilities of monophyly, paraphyly, and polyphyly given by (14)–(17) exhibit complex dependencies on the four parameters. It is convenient to first consider the effects of the times T_A and T_B, and then the effects of the sample sizes r_A and r_B.

Effects of Time

Suppose that T_A and T_B increase so that T_B/T_A is always equal to a constant K. This constant allows coalescent time to elapse at different rates in the two species. In the constant population size model, $T_B/T_A = K$ corresponds to $N_B/N_A = 1/K$.

As observed by Neigel and Avise (1986), at their time of divergence, a sample from a pair of species is likely to show polyphyly. As time progresses, paraphyly becomes more likely, and eventually, reciprocal monophyly is most likely. In the limit as time approaches ∞, reciprocal monophyly has probability one. This intuitive result holds for all values of the sample sizes, as long as they are more than one, and regardless of the value of K. (If one of the sample sizes equals one, polyphyly and one type of paraphyly are impossible, so that the transition proceeds from a paraphyly stage to a mono-

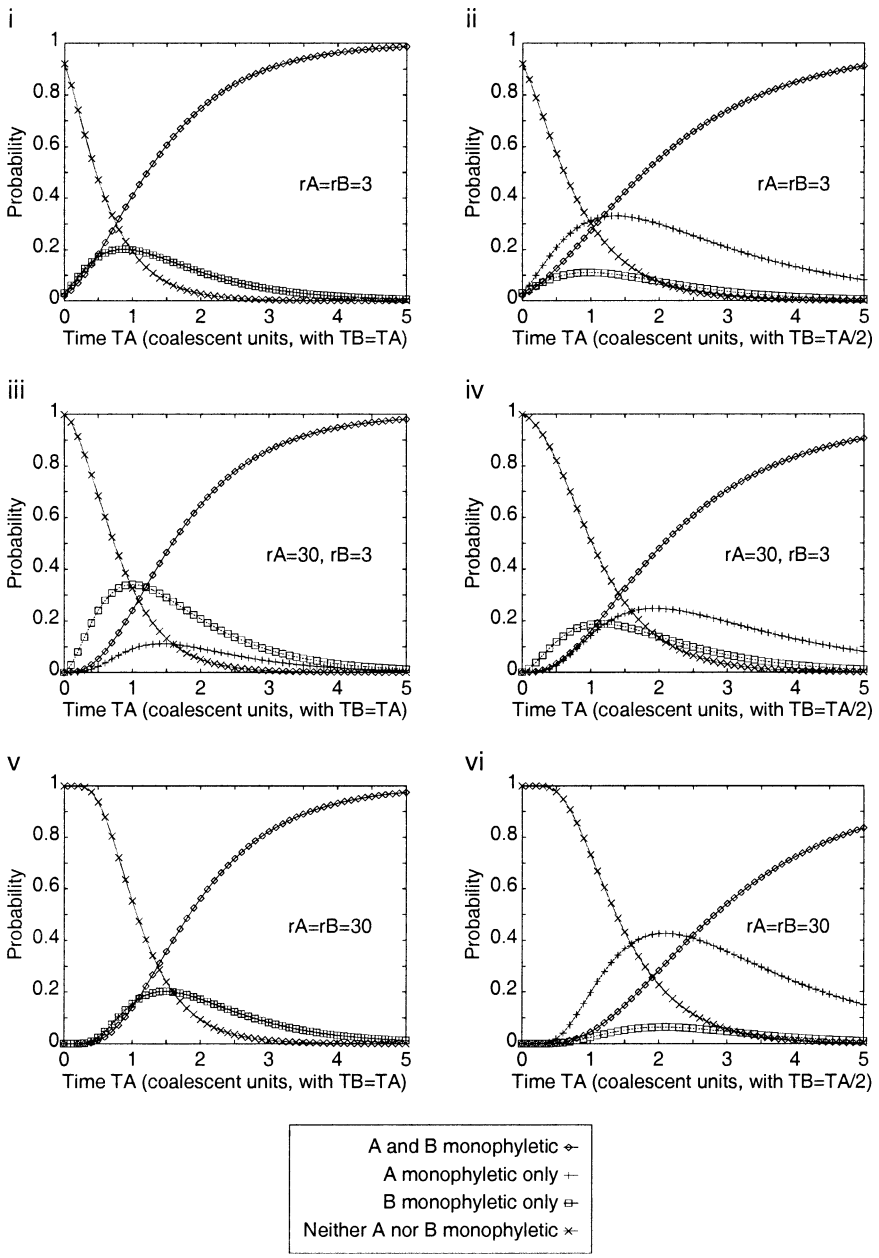


FIG. 3. Probabilities of the four types of genealogies as functions of time, obtained from equations (14)–(17). (i) $r_A = r_B = 3$, $T_B = T_A$. (ii) $r_A = r_B = 3$, $T_B = T_A/2$. (iii) $r_A = 30$, $r_B = 3$, $T_B = T_A$. (iv) $r_A = 30$, $r_B = 3$, $T_B = T_A/2$. (v) $r_A = r_B = 30$, $T_B = T_A$. (vi) $r_A = r_B = 30$, $T_B = T_A/2$.

phyly stage; if both samples have size one, I use the convention that reciprocal monophyly is guaranteed from time 0.)

Thus, in Figure 3, which shows (14)–(17) as functions of time, various parameter values produce the same qualitative limiting behavior. This result is explained by considering the numbers of lineages ancestral to the sample at the time of divergence. For small times T_A and T_B, many lineages ancestral to the sample are present at divergence. In the ancestral species these likely coalesce in such a way as to produce polyphyly. As time since divergence becomes large, sufficient time exists for all lineages to coalesce within species. At the time of divergence, likely only one lineage per species remains, and monophyly is achieved for the lineages of each species. In passing from the stage in which neither species is monophyletic to the stage in which monophyly occurs in both species, an intermediate stage is encountered in which it is likely that the lineages of only one of the species are monophyletic.

Sample sizes affect this phenomenon in that for larger sample sizes, more time must elapse before the transition between stages occurs. A comparison of Figures 3i and 3v shows that increases in sample size do not have a great impact on these transition times.

Asymmetry in sample sizes and rates of coalescence, although it does not affect the general behavior, causes the two

types of paraphyly to have different probabilities. If time elapses more rapidly in one species, the lineages of that species reach monophyly more rapidly; the species that has experienced fewer coalescent units in the same amount of absolute time is the one that is more likely to exhibit paraphyly. Similarly, if coalescence occurs at the same rate in both species, the species that has a smaller sample is likely to achieve monophyly faster, and the lineages of the other species are more likely to be paraphyletic. This difference between the two probabilities of paraphyly can be substantial (Figs. 3ii, iii, iv, vi); it depends more on relative speed of coalescence than on relative sample sizes (compare the difference between Figs. 3iii and 3iv to that between Figs. 3ii and vi).

Effects of Sample Size

As in other genealogical phenomena, the effect on (14)–(17) of increasing sample size is less profound than the effect of increasing time. Increases in sample size shift the time at which paraphyly has highest probability (cf. Figs. 3i, v), but only slightly. The probabilities of the four types of genealogies approach their large-sample limits rapidly (Fig. 4), regardless of which type of genealogy is ultimately the most likely.

Rapid convergence in sample size results from the fact that as lineages are added to a sample, they add ancestral lineages at a much slower rate. If we consider ancestors to a sample at T coalescent units in the past, adding lineages to the sample is unlikely to increase the number of ancestors if T is large. We can determine sample sizes r such that the distribution of the number of ancestral lineages at time T given a sample of size r is close to the large-sample limiting distribution. As the distribution of the number of ancestral lineages converges with increases in sample size, so too do the probabilities of monophyly, paraphyly, and polyphyly.

A previous argument (Rosenberg 2002, table 3) can be used to identify sample sizes at which probabilities of monophyly, paraphyly, and polyphyly approach their limiting values. The sample size r is chosen large enough in each species so that the mean number of ancestral lineages given a sample of size r differs from the corresponding mean for an infinite sample size by less than a specified tolerance. As can be seen by comparing Figures 4i and 4iii, if times are larger, sample sizes required for nearing the limit are smaller.

Order of the Probabilities

The parameter space can be partitioned based on the relative order of the probabilities of the four types of genealogies. Suppose $r_A = r_B = r$ and $T_A = T_B = T$. The probabilities of the two types of paraphyly are equal, so they are grouped into one category. The parameter space can contain as many as six regions, corresponding to the six orderings of the probabilities of monophyly, paraphyly, and polyphyly. The fact that evolution proceeds from polyphyly to paraphyly to monophyly suggests that the orderings Pr(*monophyly*) > Pr(*polyphyly*) > Pr(*paraphyly*) and Pr(*polyphyly*) > Pr(*monophyly*) > Pr(*paraphyly*), should never be observed. Indeed they are not seen.

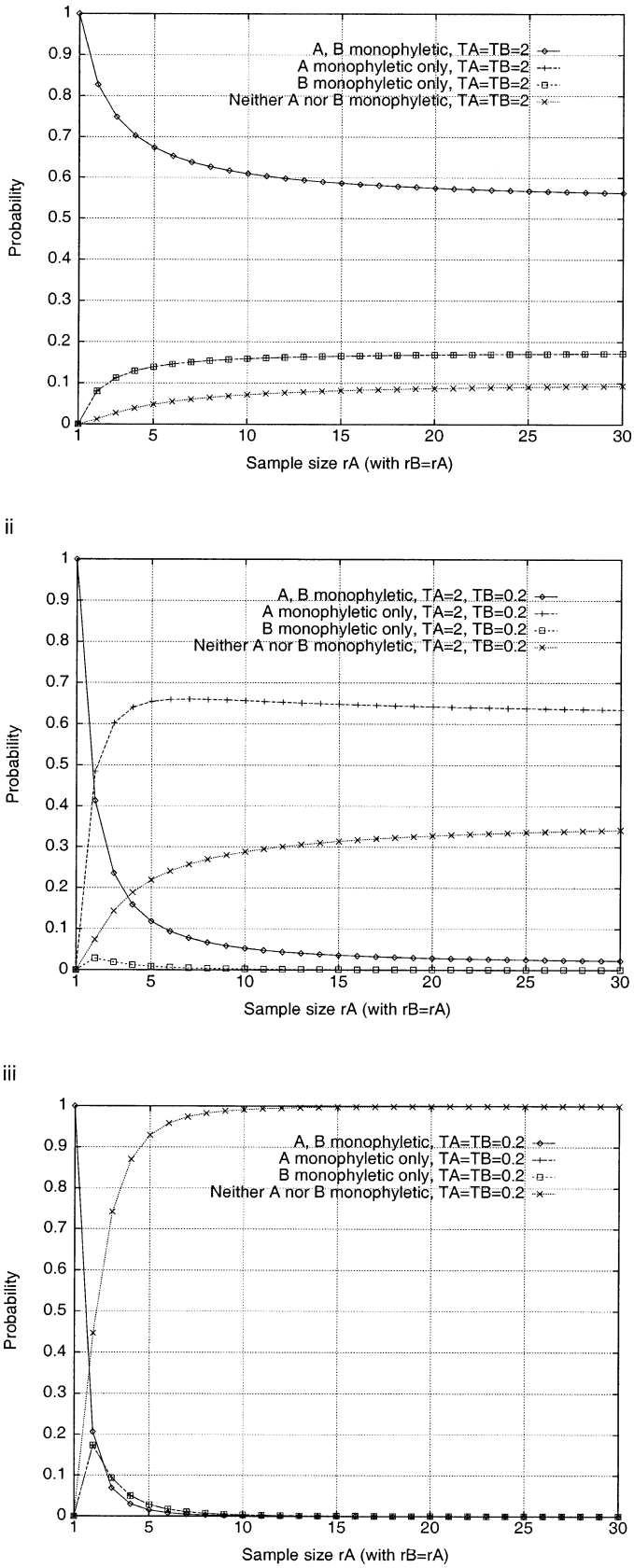


FIG. 4. Probabilities of the four types of genealogies as functions of sample size, obtained from equations (14)–(17). For all graphs, both species had the same sample size, that is, $r_A = r_B$. (i) $T_A = T_B = 2$. (ii) $T_A = 2$, $T_B = 0.2$. (iii) $T_A = T_B = 0.2$.

Transition times between different orderings of the probabilities are values of T for which two of $\Pr(X = C1)$, $\Pr(X = C2) + \Pr(X = C3)$, and $\Pr(X = C4)$, as given by (14)–(17), are equal (Fig. 5). The three equations that must be solved are polynomials in e^{-T}. For samples of size r these polynomials have degree $r(r - 1)$. With $r = 2$, exact solutions to the quadratic equations give the transition times: $\Pr(X = C2) + \Pr(X = C3) = \Pr(X = C4)$ is solved by $\ln(4/3) \approx 0.2877$; $\ln[(2 + \sqrt{6})/3] \approx 0.3942$ solves $\Pr(X = C1) = \Pr(X = C4)$, and $\ln[(4 + \sqrt{2})/3] \approx 0.5904$ solves $\Pr(X = C1) = \Pr(X = C2) + \Pr(X = C3)$. For larger r, transition times are larger, and the equations can be solved numerically. Even for small sample sizes, the transition times between orderings are close to their (approximate) limiting values of 1.29989, 1.44026, and 1.66536 (computed using ∞ for r).

Figure 5ii shows transition times for the case in which the sample from species B has size 1. This case arises because it is useful to consider whether lineages of a species are monophyletic or paraphyletic with respect to a monophyletic sister species. Because polyphyly and one type of paraphyly are not possible if $r_B = 1$, only the transition that occurs when $\Pr(X = C1) = \Pr(X = C3) = 1/2$ must be considered. For $r_A = 2$, the transition time is $\ln(4/3) \approx 0.2877$. For larger samples the transition time can be obtained numerically, and the large-sample limit is about 1.32663.

Genealogy of All Lineages in the Two Species

The genealogy for the sample of lineages provides information about the genealogy for all lineages in the two species. A polyphyletic sample indicates that the genealogy for the species is polyphyletic; a monophyletic sample suggests but does not guarantee monophyly for the species. Using X for the classification of the sample genealogy as before, and Y for that of the genealogy of all lineages from the two species, $\Pr(Y \mid X)$ is of interest. For six of 16 combinations (X, Y), the sample configuration excludes the species configuration.

The r_A and r_B sampled lineages have q_A and q_B ancestral lineages at the time of divergence, the species population sizes are N_A and N_B, and the numbers of lineages ancestral to species at the time of divergence are m_A and m_B (Fig. 2). Using equation (6), this configuration of ancestral lineages, denoted G, has probability $g_{N_A,m_A}(T_A)\ S(q_A, m_A, N_A, r_A)\ g_{N_B,m_B}(T_B)\ S(q_B, m_B, N_B, r_B)$. Here I show the probability that the lineages of both species have monophyletic genealogies given that their samples are monophyletic. Bayes's theorem yields

$$\Pr(Y = C1 \mid X = C1, G) = \frac{\Pr(Y = C1 \mid G)\ \Pr(X = C1 \mid Y = C1, G)}{\Pr(X = C1 \mid G)}. \quad (18)$$

Because the sampled lineages form a subset of the whole set of lineages, $\Pr(X = C1 \mid Y = C1, G) = 1$. $\Pr(X = C1 \mid G)$ and $\Pr(Y = C1 \mid G)$ are computed using equation (9). Summing over lineage configurations, the desired probability is

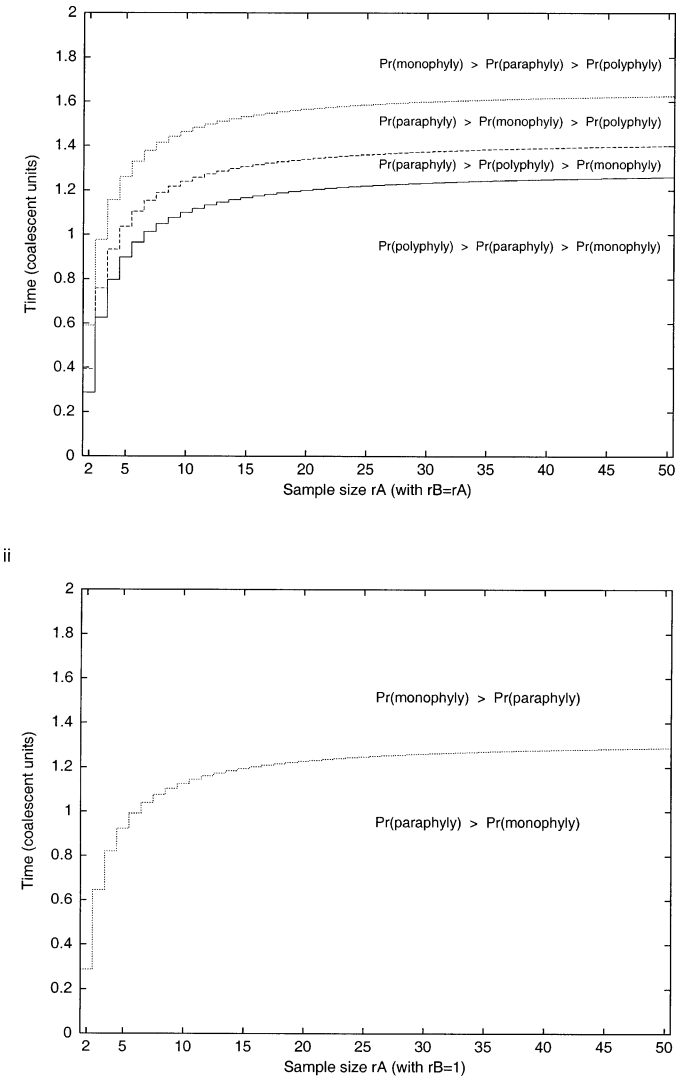
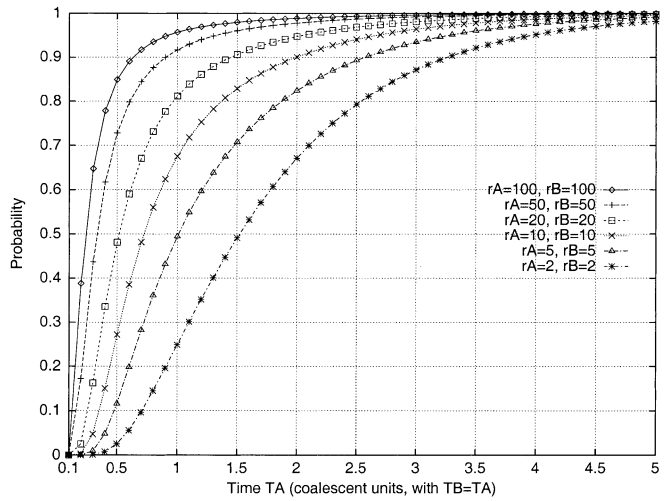


FIG. 5. Partition of the parameter space based on the ordering of the probabilities of monophyly, paraphyly, and polyphyly computed from equations (14)–(17). Time was assumed to have elapsed at the same rate for both species. (i) $r_A = r_B$. The sum of the probabilities of the two types of paraphyly was used for the probability of paraphyly. (ii) $r_B = 1$. In this case only one type of paraphyly is possible and polyphyly is not possible.

$$\Pr(Y = C1 \mid X = C1) = \sum_{m_A=1}^{N_A} \sum_{m_B=1}^{N_B} \sum_{q_A=1}^{m_A} \sum_{q_B=1}^{m_B} g_{N_A,m_A}(T_A) \times S(q_A, m_A, N_A, r_A)\ g_{N_B,m_B}(T_B) \times S(q_B, m_B, N_B, r_B) \times \frac{\binom{q_A + q_B}{q_A}(q_A + q_B - 1)}{\binom{m_A + m_B}{m_A}(m_A + m_B - 1)}. \quad (19)$$

If coalescent time elapses at the same rate in both species, for large time since divergence, monophyly of the sample indicates that monophyly of all lineages from the two species

i



ii

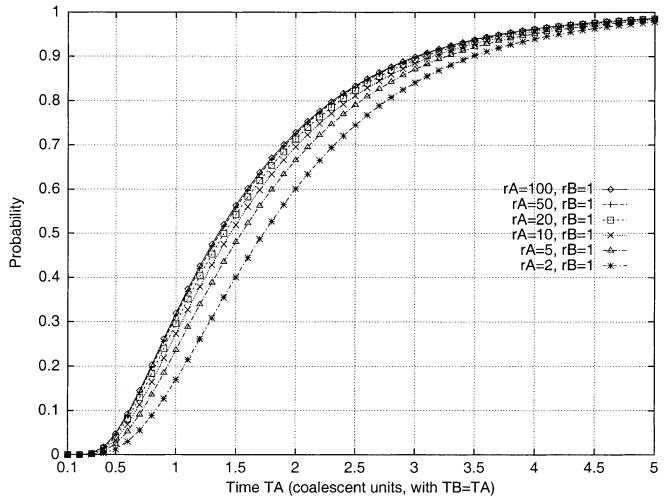


FIG. 6. Probability of reciprocal monophyly of two species given reciprocal monophyly of their samples, computed using equation (19). Results are similar as long as population sizes are large compared to sample sizes; thus, infinite population sizes were used for both species. (i) $r_B = r_A$. (ii) $r_B = 1$.

is likely (Fig. 6). For shorter times, monophyly of a larger sample gives considerably more evidence that two species are reciprocally monophyletic than does monophyly of a smaller sample (Fig. 6i). If only one lineage is sampled from species *B*, however, the probability of monophyly for species *A* given monophyly of the sample increases very slowly with sample size (Fig. 6ii; see also Nordborg 1998).

Most Recent Common Ancestor of All Lineages in the Two Species

It is of interest to know whether the MRCA of a sample is identical to the MRCA of all lineages from the two species. Let $Z_{sample}(T)$ and $Z_{species}(T)$ denote the numbers of lineages ancestral to the sample and to all lineages of the two species at time T coalescent units in the past. By summing over all lineage configurations the conditional probability that sample and species MRCAs are identical (eq. 8) times the probability of the lineage configuration, we obtain

$$\begin{aligned}\Pr[\min\{T: Z_{sample}(T) = 1\} = \min\{T: Z_{species}(T) = 1\}] \\ = \sum_{m_A=1}^{N_A}\sum_{m_B=1}^{N_B}\sum_{q_A=1}^{m_A}\sum_{q_B=1}^{m_B} g_{N_A,m_A}(T_A)\, S(q_A, m_A, N_A, r_A) \\ \times g_{N_B,m_B}(T_B) S(q_B, m_B, N_B, r_B) \\ \times \frac{(q_A + q_B - 1)}{(q_A + q_B + 1)}\frac{(m_A + m_B + 1)}{(m_A + m_B - 1)}. \qquad (20)\end{aligned}$$

For equal sample sizes and equal rates of coalescence, the probability that the sample MRCA is the species MRCA approaches one as time increases (Fig. 7i). At time 0, with $r_A = r_B = r$ and infinite population size, the result reduces to equation (8), or $(2r - 1)/(2r + 1)$. For large values of time, the sample and the species each likely have one lineage per species at the time of divergence, so the genealogy of the sample necessarily includes this lineage. The probability increases monotonically, so that for larger samples, it is more likely that the sample MRCA is the same as the species MRCA.

The case in which one of the species has only one sampled lineage is somewhat different (Fig. 7ii). Again, at time 0, equation (8) provides the desired probability. Also, for sufficiently large times, the sample and species each have one ancestral lineage, which necessarily coincides. However, a range of intermediate times exists during which the probability is smaller than the initial value of $r_A/(r_A + 2)$. In this range, the ancestry of the sample is usually reduced to a small number of lineages so that the $(q_A + q_B - 1)/(q_A + q_B + 1)$ term, or $q_A/(q_A + 2)$, is relatively small. The number of lineages ancestral to the species remains large enough that the $(m_A + m_B + 1)/(m_A + m_B - 1)$ term, or $(m_A + 2)/m_A$, is fairly close to one. Thus, in this range, the product of the two terms is usually smaller than the initial value.

Whole-Genome Monophyly

For two closely related species, equations (14)–(17) can predict the fractions of their genomes that are monophyletic, paraphyletic, and polyphyletic. For example, as the probability of monophyly for one region, (14) gives the expected fraction of a genome that is monophyletic.

Only about 1.665 coalescent units ($1.665N$ generations in the constant population size model) are needed before reciprocal monophyly is the most likely type of genealogy for two species. The time until most of their genomes are monophyletic is considerably longer (Table 1). The time until a certain fraction of the genome of species *A* is expected to be monophyletic depends only weakly on its sister species. Assuming that the sister species to *A* is monophyletic only reduces the time to monophyly of *A* by 0.6–0.8 coalescent units.

Using simulations with finite sample sizes, Hudson and Coyne (2002) obtained results for the time until reciprocal monophyly that are similar to the values in Table 1 (second row of their table 1). Corresponding values of the time until monophyly of species *A* in Table 1 and in Hudson and Coyne (2002, table 1) differ slightly, however: I assume species *A* has a sister species that contains one lineage that eventually

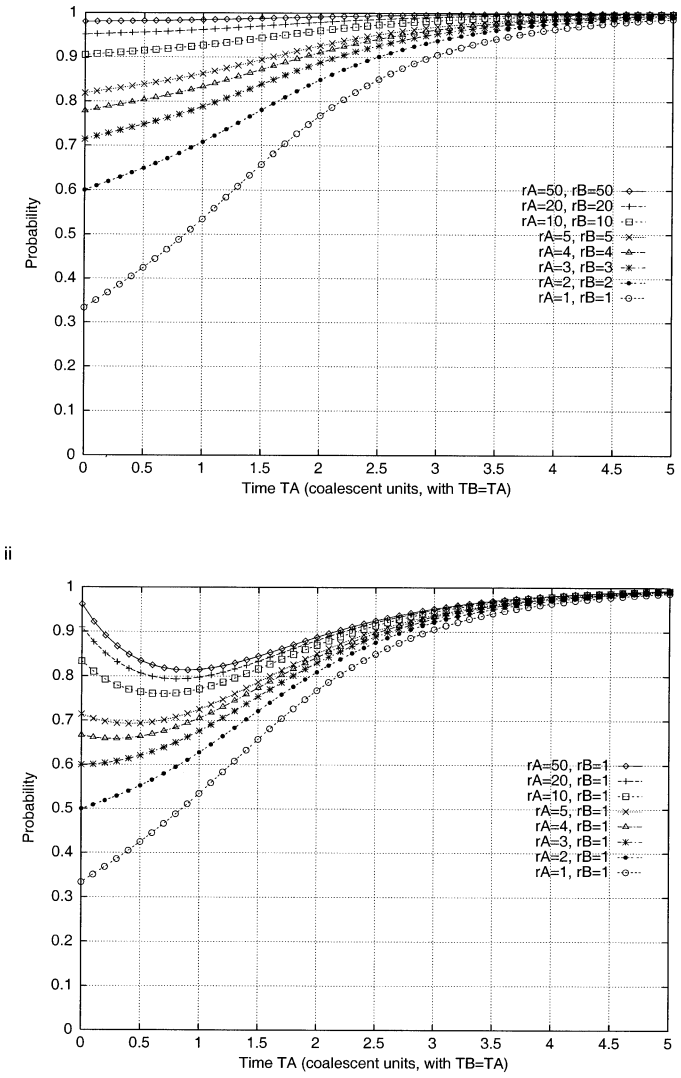


FIG. 7. Probability that the most recent common ancestor (MRCA) for a sample from two species is the same as the MRCA for all lineages in the two species, computed using equation (20). Results are similar as long as population sizes are large compared to sample sizes; thus, infinite population sizes were used for both species. (i) $r_A = r_B$. (ii) $r_B = 1$.

coalesces with the lineages of species *A*, whereas Hudson and Coyne (2002) study the time until monophyly of species *A* without allowing interspecific coalescences (see also Arbogast et al. 2002). Note that Hudson and Coyne (2002) considered $2N$ lineages per species but scaled time in units of N generations; thus, times in their table must be divided by two to be directly comparable with those here.

Given a number α, we can also determine the time until a whole genome is monophyletic with probability $1 - \alpha$. This calculation assumes neutrality of the whole genome; thus, nonmonophyletic regions that persist after this length of time are candidates for targets of natural selection (see Discussion). Let $M_1(T)$ be calculated from (14), using $r_A = \infty$, $r_B = 1$, $T_A = T$, and T_B arbitrary. This gives the probability of monophyly of a region in species *A*, assuming that species *B* is monophyletic. The calculations are analogous for $M_2(T)$,

TABLE 1. Waiting times until monophyly. $M_1^{-1}(1-\alpha)$ is the waiting time after divergence until 100 $(1 - \alpha)$% of the genome of species *A* is expected to be monophyletic, assuming species *B* is monophyletic. $M_2^{-1}(1 - \alpha)$ is the corresponding time, assuming species *A* and *B* have the same parameters. Times, measured in coalescent units, are obtained by solving equation (14) for T_A, assuming $Pr(X = C1) = 1 - \alpha$. For M_1^{-1}, $r_B = 1$ and the value of T_B does not matter. For M_2^{-1}, $r_B = r_A$ and $T_B = T_A$. For both columns, $r_A = \infty$.

α	$M_1^{-1}(1-\alpha)$	$M_2^{-1}(1-\alpha)$
0.50	1.327	1.903
0.10	2.994	3.662
0.05	3.794	4.369
0.01	5.298	5.989
0.001	7.601	8.294
10^{-4}	9.903	10.597
10^{-5}	12.206	12.899
10^{-6}	14.509	15.202
10^{-7}	16.811	17.504

the probability of reciprocal monophyly of a region, using $r_A = r_B = \infty$, $T_A = T_B = T$. Two sites in a genome that are separated by sufficient distance l have independent genealogies. Thus, a genome of total length C base pairs can be loosely approximated by u disjoint subunits, each of length l ($C = ul$), such that each pair of subunits is independent and such that no recombination occurs within any subunit. The probability of monophyly for any subunit is $M_1(T)$. To determine T_α so that the whole genome is monophyletic with probability $1 - \alpha$ after time T_α, we must solve

$$1 - \alpha = [M_1(T_\alpha)]^u. \tag{21}$$

If M_1^{-1} is the inverse function of M_1 (Table 1), the solution is

$$T_\alpha = M_1^{-1}[(1 - \alpha)^{1/u}] \approx M_1^{-1}(e^{-\alpha/u}). \tag{22}$$

As an example, consider two genomes of length 3000 megabases (Mb) in which sites separated by 0.2 Mb have independent genealogies. These genomes have $u = 15,000$, and the time until both genomes are fully reciprocally monophyletic with probability 0.95 is M_2^{-1} (0.9999966), or 13.972 coalescent units. The bottom row of table 1 of Hudson and Coyne (2002) contains a similar computation.

DISCUSSION

Equations (14)–(17) give the closed-form probabilities of the four types of genealogies under the coalescent two-species divergence model, superseding recursively defined solutions of Hudson and Coyne (2002) and Rosenberg (2002). Note, however, that (14) disagrees with the probability of monophyly suggested by Palumbi et al. (2001, eq. 1) for $N_B = \infty$ and $q_B = 1$. Unlike (14) and the approaches of Hudson and Coyne (2002) and Rosenberg (2002), Palumbi et al. (2001) assumed $q_A = 1$, or that monophyly of species *A* requires all lineages of species *A* to coalesce more recently than divergence. Because monophyly of *A* is also possible for $q_A > 1$ if the ancestral lineages of species *A* coalesce intraspecifically, the formula of Palumbi et al. (2001) can be regarded as an underestimate of the probability of monophyly of *A*. The discrepancy between (14) and equation (1) of Palumbi et al. (2001) is largest at small divergence times, for which $q_A > 1$ is a likely possibility.

Figures 3 and 4 confirm the stronger dependence of (14)–(17) on divergence times compared to sample sizes. For two recently diverged species, however, Figures 6 and 7 show that sample size noticeably affects the relationship between the genealogy of the sample and that of the species. For very small samples, one should be cautious not to equate reciprocal monophyly of a sample with reciprocal monophyly of the two species.

The concepts of monophyly, paraphyly, and polyphyly of lineages have frequently been employed in molecular ecology and phylogeography (Neigel and Avise 1986; Palumbi et al. 2001). Probabilities that sampled lineages exhibit monophyly, paraphyly, or polyphyly can be used in studies both of the history of a species as a whole (Wakeley and Hey 1997; Hudson and Coyne 2002) and of roles played in evolution by individual genes (Wang et al. 1999; Ting et al. 2000). In various contexts, the formulas here can assist in designing studies, making predictions, and interpreting data.

Paraphyly and Divergence Times

Because paraphyletic genealogies are most frequent for only a short period of time (Fig. 5i), observed dominance of paraphyly in multilocus datasets suggests that species lie in this intermediate period since divergence. If both types of paraphyly have similar frequencies and if paraphyletic genealogies occur more often than monophyletic and polyphyletic genealogies, the species are likely in the narrow band of time from $\sim$1.3 to 1.7 coalescent units since divergence ($\sim 1.3N$ to $1.7N$ generations, in the constant population size model).

Moreover, an observation that the frequencies of both types of paraphyly are about the same indicates that coalescence occurs at similar rates in the two species, whereas differences between the two frequencies provide evidence for a difference in coalescence rates. The species that experiences coalescence more slowly is more likely to be paraphyletic (Figs. 3ii, 3vi, 4ii). If constant population size is assumed, the species that has more paraphyly also has a larger population size.

Symmetric and Asymmetric Samples

Reasonably small samples are sufficient to characterize shapes of most two-species genealogies. Sample sizes that can be used to achieve desired precision in the number of ancestral lineages, and corresponding precision in genealogical shape probabilities, can be obtained from Rosenberg (2002, table 3).

A large sample from one species together with only a single lineage from the other provides much less information than moderate but similar sample sizes from both species. In Figure 6, if one species has sample size 1, increasing the sample size from the other species helps only minimally toward correct inference that the lineages of that species are monophyletic. By contrast, if both species have similar sample sizes, few lineages are needed to almost guarantee that sample monophyly implies species monophyly. Inferences about MRCAs behave similarly: caution is warranted in inference about the time to the MRCA of two closely related groups, when one group has a small sample size and the other does not. This comment especially applies to recent divergences, in which the MRCA of an asymmetric sample has considerable probability of differing from the species MRCA (Fig. 7ii).

These problems of asymmetric samples parallel results on the concordance of gene trees and species trees (Rosenberg 2002). If three species are studied by sampling many lineages for the two species that are predicted to be sister species, but only one lineage from the predicted outgroup species, the analysis will tend to place the predicted outgroup as the outgroup, even if this is not correct. By contrast, symmetric sampling less frequently leads to this erroneous inference.

Mitochondrial and Nuclear DNA

Recent studies investigate relationships between shapes of genealogies of mitochondrial DNA and those of nuclear DNA (Moore 1995; Palumbi et al. 2001). Under neutrality, with equal offspring distributions for diploid males and females, coalescence time elapses four times as fast for mitochondrial DNA as for nuclear DNA. As can be observed from Figures 3, 5, and 6, the behavior of gene genealogies can be quite different at time $4T$ (mitochondrial DNA) compared to time T (nuclear DNA).

For example, if three coalescent units have elapsed for mitochondria, in Figure 3iii, mitochondrial DNA has monophyly probability over 0.8. Nuclear loci in the same organism, however, having only experienced 0.75 coalescent units, have monophyly probability under 0.1. Using Figure 5i and large samples, mitochondrial and nuclear loci reside in the lower partition of the parameter space until $T \approx 0.325$, when paraphyly becomes more likely than polyphyly for mitochondrial DNA, but not for nuclear loci. The two types remain in separate partitions until $T \approx 1.665$, when both types are likely to be monophyletic.

Unusual Genealogical Shapes and Selection

Deviations from predictions of the neutral model here can help identify loci undergoing various forms of selection. Selected loci may be more or less likely than neutral loci to have reciprocally monophyletic genealogies. Balancing selection or selection for diversity increases the probability that many ancestral lineages per species persist into the distant past (Ioerger et al. 1990; Takahata and Nei 1990), decreasing the probability of monophyly. Loci that were under selection during species divergence or that have been involved in maintenance of differences between species might be more likely to be monophyletic in at least one of the species, because the MRCA of extant lineages in that species might be a novel mutant that lived around the time of divergence (Hey 1994; Wang et al. 1999; Ting et al. 2000).

Tests of genealogical shape have been devised in various contexts. Species trees are compared with the Yule model or other phylogenetic models, and agreement of inferred shapes of species trees with model predictions is taken as evidence for mechanisms of evolution that underlie the models (Slowinski and Guyer 1989; Mooers and Heard 1997; Aldous 2001; Harcourt-Brown et al. 2001). If the tree for a set of species has an improbable shape, special properties of those species are used to explain the anomaly. Within a species, statistics based on shapes of gene genealogies predicted by the neutral

coalescent are used to test if evolution of particular genes has occurred consistently with neutrality (Kreitman 2000). For genes that do not fit the model, the nature of the deviation assists in characterizing the alternative mode of evolution acting on the gene.

Corresponding tests might be derived for multispecies gene genealogies (Hey 1994; Palumbi et al. 2001). If divergence was ancient enough that the monophyly probability is near one, genomic regions that are observed to not be monophyletic might be under balancing selection. If divergence was recent enough that the probability of monophyly is near zero, then regions observed to be monophyletic might have been important in species divergences. As polymorphism data accumulate, a genomic approach may help to determine how many genes experience these types of selection.

Conclusions

In this article the probabilities of monophyly, paraphyly, and polyphyly have been computed under a coalescent model. The approach also enabled calculation of the probability that the lineages of two species are reciprocally monophyletic given that a sample has this property. Times at which the dominant genealogical shape transits from polyphyly to paraphyly or from paraphyly to monophyly have been obtained.

Note that genealogies have been discussed without concern for the fact that in practice they are estimated from DNA polymorphism data. The importance to the analysis of the shapes of genealogies makes it desirable to determine these shapes probabilistically. Recombination also has not been considered; the numerical results apply strictly only to regions in which no recombination has occurred. Genomes of organisms with low recombination rates may contain many such regions.

Similar formulas to those here might be obtained for more than two species. Many more types of genealogical shapes are then possible, including shapes in which the gene genealogy and species tree disagree. The larger number of possible shapes of gene genealogies can make analysis of gene tree shape cumbersome. It might be best to divide trees of many species into overlapping subsets of a few species each and then analyze genealogical shape in each subset. The results here are also based on simple models of within-species evolution; advances might treat more complex models of the divergence process (Teshima and Tajima 2002) and of population structure after divergence.

Acknowledgments

I thank H. Innan and S. Tavaré for discussions and M. Nordborg for comments on the manuscript. This research was supported by an National Science Foundation Postdoctoral Fellowship in Biological Informatics.

Literature Cited

Aldous, D. J. 2001. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. Stat. Sci. 16:23–34.

Arbogast, B. S., S. V. Edwards, J. Wakeley, P. Beerli, and J. B. Slowinski. 2002. Estimating divergence times from molecular data on phylogenetic and population genetic timescales. Annu. Rev. Ecol. Syst. 33:707–740.

Brown, J. K. M. 1994. Probabilities of evolutionary trees. Syst. Biol. 43:78–91.

Clark, A. 1997. Neutral behavior of shared polymorphism. Proc. Natl. Acad. Sci. USA 94:7730–7734.

Edwards, A. W. F. 1970. Estimation of the branch points of a branching diffusion process. J. R. Stat. Soc. Ser. B 32:155–174.

Gould, H. W. 1972. Combinatorial identities. Rev. ed. Gould Publications, Morgantown, WV.

Griffiths, R. C. 1984. Asymptotic line-of-descent distributions. J. Math. Biol. 21:67–75.

Griffiths, R. C., and S. Tavaré. 2003. The genealogy of a neutral mutation. *In* P. Green, N. Hjort, and S. Richardson, eds. Highly structured stochastic systems. Oxford Univ. Press, Oxford, U.K. *In press*.

Harcourt-Brown, K. G., P. N. Pearson, and M. Wilkinson. 2001. The imbalance of paleontological trees. Paleobiology 27: 188–204.

Harding, E. F. 1971. The probabilities of rooted tree-shapes generated by random bifurcation. Adv. Appl. Prob. 3:44–77.

Hey, J. 1991. The structure of genealogies and the distribution of fixed differences between DNA sequence samples from natural populations. Genetics 128:831–840.

———. 1994. Bridging phylogenetics and population genetics with gene tree models. Pp. 435–449 *in* B. Schierwater, B. Streit, G. P. Wagner, and R. DeSalle, eds. Molecular ecology and evolution: approaches and applications. Birkhäuser Verlag, Basel.

Hudson, R. R., and J. A. Coyne. 2002. Mathematical consequences of the genealogical species concept. Evolution 56:1557–1565.

Ioerger, T. R., A. G. Clark, and T.-H. Kao. 1990. Polymorphism at the self-incompatibility locus in Solanaceae predates speciation. Proc. Natl. Acad. Sci. USA 87:9732–9735.

Kreitman, M. 2000. Methods to detect selection in populations with applications to the human. Annu. Rev. Genomics Hum. Genet. 1:539–559.

Maddison, W. P., and M. Slatkin. 1991. Null models for the number of evolutionary steps in a character on a phylogenetic tree. Evolution 45:1184–1197.

Mooers, A. O., and S. B. Heard. 1997. Evolutionary process from phylogenetic tree shape. Q. Rev. Biol. 72:31–54.

Moore, W. S. 1995. Inferring phylogenies from mtDNA variation: mitochondrial-gene trees versus nuclear-gene trees. Evolution 49:718–726.

Neigel, J. E., and J. C. Avise. 1986. Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation. Pp. 515–534 *in* S. Karlin and E. Nevo, eds. Evolutionary processes and Theory. Academic Press, New York.

Nordborg, M. 1998. On the probability of Neanderthal ancestry. Am. J. Hum. Genet. 63:1237–1240.

———. 2001. Coalescent theory. Pp. 179–212 *in* D. J. Balding, C. Cannings, and M. Bishop, eds. Handbook of statistical genetics. Wiley, Chichester, U.K.

Palumbi, S. R., F. Cipriano., and M. P. Hare. 2001. Predicting nuclear gene coalescence from mitochondrial data: the three-times rule. Evolution 55:859–868.

Rosenberg, N. A. 2002. The probability of topological concordance of gene trees and species trees. Theor. Popul. Biol. 61:225–247.

Sanderson, M. J. 1996. How many taxa must be sampled to identify the root node of a large clade? Syst. Biol. 45:168–173.

Saunders, I. W., S. Tavaré, and G. A. Watterson. 1984. On the genealogy of nested subsamples from a haploid population. Adv. Appl. Prob. 16:471–491.

Slowinski, J. B., and C. Guyer. 1989. Testing the stochasticity of patterns of organismal diversity: an improved null model. Am. Nat. 134:907–921.

Steel, M., and A. McKenzie. 2001. Properties of phylogenetic trees generated by Yule-type speciation models. Math. Biosci. 170: 91–112.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460.

Takahata, N. 1989. Gene genealogy in three related populations: consistency probability between gene and population trees. Genetics 122:957–966.

Takahata, N., and M. Nei. 1985. Gene genealogy and variance of interpopulational nucleotide differences. Genetics 110:325–344.
———. 1990. Allelic genealogy under overdominant and frequency dependent selection and polymorphism of major histocompatibility complex loci. Genetics 124:967–978.
Takahata, N., and M. Slatkin. 1990. Genealogy of neutral genes in two partially isolated populations. Theor. Popul. Biol. 38: 331–350.
Tavaré, S. 1984. Line-of-descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. 26:119–164.
Teshima, K. M., and F. Tajima. 2002. The effect of migration during the divergence. Theor. Popul. Biol. 62:81–95.
Ting, C.-T., S.-C. Tsaur, and C.-I. Wu. 2000. The phylogeny of closely related species as revealed by the genealogy of a speciation gene, *Odysseus*. Proc. Natl. Acad. Sci. USA 97: 5313–5316.
Wakeley, J. 2000. The effects of subdivision on the genetic divergence of populations and species. Evolution 54:1092–1101.
Wakeley, J., and J. Hey. 1997. Estimating ancestral population parameters. Genetics 145:847–855.
Wang, R.-L., A. Stec., J. Hey., L. Lukens., and J. Doebley. 1999. The limits of selection during maize domestication. Nature 398: 236–239.
Yule, G. U. 1924. A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis, F. R. S. Philos. Trans. R. Soc. Lond. B 213:21–87.

Corresponding Editor: R. Harrison

Appendix 1

To simplify equation (11), identity 1 is needed. To prove identity 1, we need identities 2 and 3. Identity 2 is proven by induction on n, using $\binom{n}{k} = n!/[k!(n-k)!]$. Identity 3 follows from identity 2 by induction on n:

Identity 1: For positive integers m and n,

$$\sum_{k=0}^{n} \frac{k\binom{n}{k}}{\binom{m+n-1}{k}} = \frac{n(m+n)}{m(m+1)}. \tag{A1}$$

Identity 2: For positive integers m and n,

$$\sum_{k=0}^{n} \binom{m-1+k}{k} = \binom{m+n}{n}. \tag{A2}$$

Identity 3: For positive integers m and n,

$$\sum_{k=0}^{n} k\binom{m-1+k}{k} = \frac{n(m+n)}{m+1}\binom{m+n-1}{n}. \tag{A3}$$

Proof of Identity 1: We substitute $l = n - k$:

$$\begin{aligned}
&\sum_{k=0}^{n} \frac{k\binom{n}{k}}{\binom{m+n-1}{k}} \\
&= \sum_{k=0}^{n} k\frac{n!\,(m+n-1-k)!}{(n-k)!(m+n-1)!} \\
&= \sum_{k=0}^{n} k\frac{(m-1)!\,n!}{(m+n-1)!}\,\frac{(m+n-1-k)!}{(m-1)!\,(n-k)!} \\
&= \frac{1}{\binom{m+n-1}{n}} \sum_{k=0}^{n} k\binom{(m-1)+(n-k)}{n-k} \\
&= \frac{1}{\binom{m+n-1}{n}}\left[n\sum_{l=0}^{n}\binom{m-1+l}{l} - \sum_{l=0}^{n} l\binom{m-1+l}{l}\right].
\end{aligned} \tag{A4}$$

The proof is completed by applying identities 2 and 3. Identity 1 generalizes identity 4.10 of Gould (1972).

Appendix 2

A previous derivation of the probability of reciprocal monophyly (eq. 14) was given in terms of a recursive function (Rosenberg 2002, eq. 10). Here I give a closed-form expression for that function, which enables closed-form solution of other results in Rosenberg (2002) and Takahata (1989).

$F_k^{A,B}(a, b, c)$ (Rosenberg 2002, eq. 21) is the probability that in coalescing from a configuration that includes a lineages of species A, b lineages of species B, and c lineages of species C down to k total lineages, an interspecific coalescence occurs and the most recent interspecific coalescence links species A and B. The closed-form solution for $F_k^{A,B}$ is obtained using a counting approach. Suppose that a, b, and c are all positive, and suppose that there are x, y, and z ancestors of the a, b, and c lineages, respectively, when one of the x lineages and one of the y lineages engage in the most recent interspecific coalescence. The remaining $x + y + z - 1$ lineages can coalesce to k lineages in any order.

Using equation (4), there are $I_{a,x}$ sequences that can coalesce a lineages to x lineages. Similarly there are $I_{b,y}$ ways to coalesce b lineages to y lineages and $I_{c,z}$ ways to coalesce c lineages to z lineages. The total number of sequences that coalesce a to x, b to y, and c to z is $I_{a,x}I_{b,y}I_{c,z}W_3(a-x, b-y, c-z)$, where W_3 is the number of ways of interweaving three sequences of coalescences. Analogously to equation (5), W_3 is the trinomial coefficient

$$\begin{aligned}
W_3(a-x, b-y, c-z) &= \binom{a-x+b-y+c-z}{a-x, b-y, c-z} \\
&= \frac{(a-x+b-y+c-z)!}{(a-x)!\,(b-y)!\,(c-z)!}.
\end{aligned} \tag{A5}$$

There are xy ways to choose the two lineages that are the most recent to coalesce interspecifically. Finally there are $I_{x+y+z-1,k}$ ways to coalesce the $x + y + z - 1$ remaining lineages to k lineages. Thus, there are $I_{a,x}I_{b,y}I_{c,z}W_3(a-x, b-y, c-z)xyI_{x+y+z-1,k}$ sequences of coalescences that have the desired properties.

The most recent interspecific coalescence occurs when there are at least $k + 1$ lineages, so $x + y + z \geq k + 1$. Also, $1 \leq x \leq a$, $1 \leq y \leq b$, and $1 \leq z \leq c$. Using the allowable values of (x, y, z), the number of sequences of coalescences that satisfy the requirements is

$$\sum_{x=\max(1,k+1-b-c)}^{a}\ \sum_{y=\max(1,k+1-x-c)}^{b}\ \sum_{z=\max(1,k+1-x-y)}^{c} I_{a,x}I_{b,y}I_{c,z}\,W_3(a-x, b-y, c-z)xyI_{x+y+z-1,k}. \tag{A6}$$

The number of possible sequences of coalescences is $I_{a+b+c,k}$. The result is simplified using equations (4) and (A5):

$$F_k^{A,B}(a, b, c) = \sum_{x=\max(1,k+1-b-c)}^{a} \sum_{y=\max(1,k+1-x-c)}^{b} \sum_{z=\max(1,k+1-x-y)}^{c} \frac{I_{a,x}I_{b,y}I_{c,z}}{I_{a+b+c,k}} W_3(a - x, b - y, c - z)xyI_{x+y+z-I,k}$$

$$= \sum_{x=\max(1,k+1-b-c,1)}^{a} \sum_{y=\max(1,k+1-x-c)}^{b} \sum_{z=\max(1,k+1-x-y)}^{c} \frac{\binom{a}{x}\binom{b}{y}\binom{c}{z}}{\binom{a+b+c}{x+y+z}} \frac{\binom{x+y+z}{x,y,z}}{\binom{a+b+c}{a,b,c}} \frac{a+b+c}{abc} \frac{2(xy)^2 z}{(x+y+z)^2(x+y+z-1)}. \quad \text{(A7)}$$

If $c = 0$, the calculation is analogous and simpler.

$$F_k^{A,B}(a, b, 0) = \sum_{x=\max(1,k+1-b)}^{a} \sum_{y=\max(1,k+1-x)}^{b} \frac{I_{a,x}I_{b,y}}{I_{a+b,k}} W_2(a - x, b - y)xyI_{x+y-1,k}$$

$$= \sum_{x=\max(1,k+1-b)}^{a} \sum_{y=\max(1,k+1-x)}^{b} \frac{\binom{a}{x}\binom{b}{y}}{\binom{a+b}{x+y}} \frac{\binom{x+y}{x}}{\binom{a+b}{a}} \frac{a+b}{ab} \frac{2(xy)^2}{(x+y)^2(x+y-1)}. \quad \text{(A8)}$$

It is straightforward to show that (A7) gives the same numerical values as equation 21 of Rosenberg (2002).