Measures of Synteny Conservation Between Species Pairs

Elizabeth Ann Housworth*,[1] and John Postlethwait†

Mathematics Department, †Institute of Neuroscience, University of Oregon, Eugene, Oregon 97403

Manuscript received September 12, 2001
Accepted for publication June 3, 2002

ABSTRACT

Measures of conserved synteny are important for estimating the relative rates of chromosomal evolution in various lineages. We present a natural way to view the synteny conservation between two species from an Oxford grid—an $r \times c$ table summarizing the number of orthologous genes on each of the chromosomes 1 through r of the first species that are on each of the chromosomes 1 through c of the second species. This viewpoint suggests a natural statistic, which we denote by ρ and call syntenic correlation, designed to measure the amount of synteny conservation between two species. This measure allows syntenic conservation to be compared across many pairs of species. We improve the previous methods for estimating the true number of conserved syntenies given the observed number of conserved syntenies by taking into account the dependency of the numbers of orthologues observed in the chromosome pairings between the two species and by determining both point and interval estimators. We also discuss the application of our methods to genomes that contain chromosomes of highly variable lengths and to estimators of the true number of conserved segments between species pairs.

GENOME evolution in multichromosomal organisms involves the translocation of genes between chromosomes, the rearrangement of genes on chromosomes, splitting and fusion of chromosomes, and gene and genome duplication events. Comprehensive measures of rearrangement distances, even restricted to pairs of chromosomes, one from each species, are computationally difficult to obtain. These measures are feasible only with highly conserved orthologous gene arrangements (Sankoff *et al.* 1992; Graur and Li 2000), *e.g.*, for the Herpesviruses (Hannenhalli *et al.* 1995).

Recent articles modeling and measuring genome evolution have concentrated on estimating the true number of conserved syntenies or the true total number of conserved chromosomal segments between pairs of species (Sankoff and Nadeau 1996; Ehrlich *et al.* 1997; Sankoff *et al.* 1997; Waddington *et al.* 1999; Kumar *et al.* 2001). Synteny refers to genes on the same chromosome and the original definition of a conserved synteny between two species was the presence of two or more orthologues syntenic in each of the two species. However, many of the previous works identified a conserved synteny by the presence of one or more markers or orthologues, not two or more. We use the latter method and define a conserved synteny as the presence of one or more orthologues on a pair of chromosomes (one chromosome from each species). See Figure 1 for a syntenic plot of orthologues in humans and cats.

Measures of conserved synteny ignore gene rearrangements on the chromosomes while measures involving the number of conserved segments take into consideration separate intrachromosomal rearrangements of blocks of orthologues while ignoring rearrangements within those blocks. Estimates of the true number of conserved syntenies clearly underestimate the true number of conserved segments between the genomes of two species. Measures of the total number of conserved syntenies or the total number of conserved segments are computationally feasible to obtain and provide a gross measure of genomic distance between pairs of species.

Both the estimators for synteny and segment conservation in recent articles (Sankoff and Nadeau 1996; Ehrlich *et al.* 1997; Waddington *et al.* 1999; Kumar *et al.* 2001) have been developed under the assumption that the proportion of genes observed in one syntenic group or segment is independent of the proportion observed in another. This approximation was justified (Sankoff and Nadeau 1996) by the argument that the relative lengths of any two segments are only very weakly correlated. However, because these measures involve not a few groups or segments but all observed groups or segments, this dependency is increasingly important as a larger percentage of the genome is mapped. We show in this article that it is a relatively simple mathematical matter to take this dependency into account and that doing so provides a simple statistical estimator for the true number of conserved syntenies. In the discussion, we consider the application of our method to estimates of the true total number of conserved segments between pairs of species.

Neither the true total number of conserved syntenies nor the true total number of conserved segments is

[1]*Corresponding author:* Mathematics and Biology Departments, Indiana University, Bloomington, IN 47405.
E-mail: ehouswor@indiana.edu

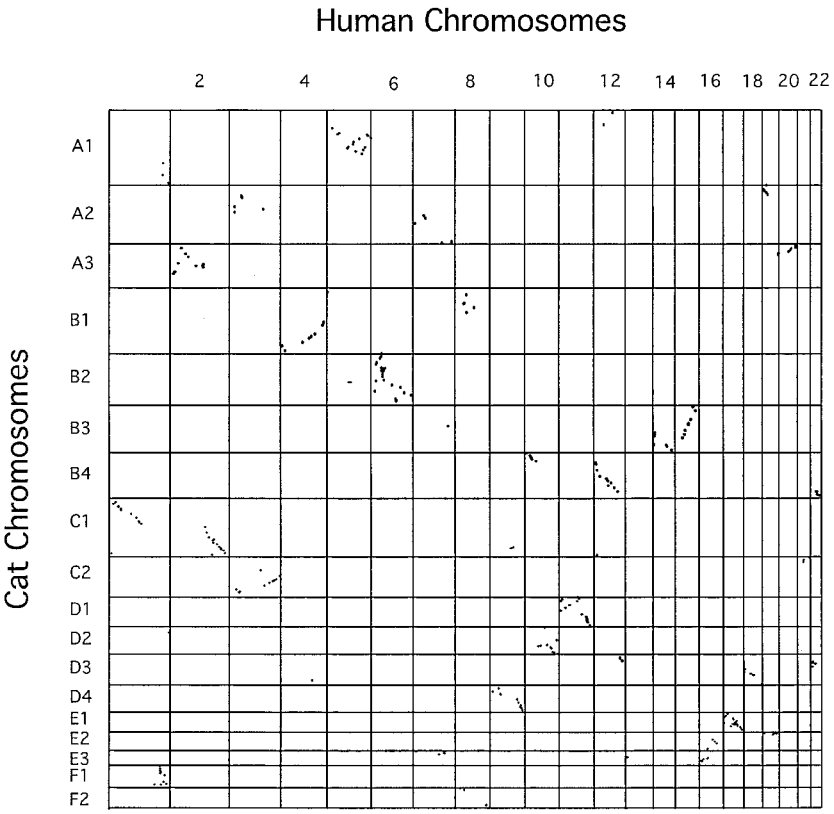

FIGURE 1.—Synteny plot for the orthologuous genes between humans and cats. Dots indicate the relative position of each orthologous gene pair on the chromosomes of humans and cats. The width and length of each box are proportional to the lengths of the chromosomes determining the box. The data are taken from MURPHY *et al.* (2000).

particularly useful in comparing genomic distances between pairs of species because raw counts do not provide adjustments or standardizations for such basic genomic differences between pairs of species such as genome sizes or numbers of chromosomes. Further, the orthologous genes that are yet to be found may be ones not subject to much genetic or genomic constraint. These orthologues may be scattered widely on the chromosomes of species and may inflate the number of conserved syntenies or the number of conserved segments while the bulk of the genome may be highly conserved. We introduce a measure of genomic conservation, which we call syntenic correlation, which corresponds to a measure of how far the orthologues are from being independently scattered in the genomes of the two species. This measure is standardized to be between zero, for completely randomized arrangements of orthologues between the genomes, and one, for two genomes with perfect synteny conservation. Further, this measure can be used to compare genomic distances (*i.e.*, Oxford grids) between many pairs of species.

METHODS

Multivariate distribution of gene counts: Measures of synteny conservation come essentially from looking at an Oxford grid, *i.e.*, an $r \times c$ table of the r chromosomes in species A and the c chromosomes in species B. The (i, j) entry is denoted $n_{i,j}$ and is the observed number of genes on species A chromosome i with an orthologue on species B chromosome j. A pictorial representation of this table is formed by placing a dot in the (i, j) box, representing the chromosome pair, in the orthologue's relative position on each chromosome (see Figure 1). A box with one or more entries is counted as a conserved synteny. The distribution of the n observed orthologues then follows a multinomial distribution with $r \times c$ classes (the boxes or pairs of chromosomes), each orthologue having chance $p_{i,j}$ of landing in the (i, j) class.

It is easiest for the analysis that follows to change notation to avoid the multidimensional subscript. Let $r \times c = m$. Label the possible chromosome pairs 1, 2, . . . , m and the corresponding probabilities $p_1, p_2, \ldots, p_m$. Label the observed number of orthologues n_1 for the first chromosome pair, n_2 for the second, . . . , n_m for the mth pair. The multinomial distribution for the number of orthologs found on each chromosome pair is then

$$f(n_1, n_2, \ldots, n_m | p_1, p_2, \ldots, p_m) = \binom{n}{n_1, n_2, \ldots, n_m} p_1^{n_1} p_2^{n_2} p_m^{n_m}$$

for $n_1 \geq 0$, $n_2 \geq 0$, . . . , $n_m \geq 0$, and $n_1 + n_2 + \cdots n_m = n$ and where 0^0 and 0! are interpreted as 1 in this context.

Let $l \leq m$ be the number of chromosome pairs on which orthologous genes will ever be found. The goal is to find an estimate for l, the true number of conserved syntenies that will be found after the genomes of both species have been completely mapped and analyzed.

Multivariate distribution of the lengths of the syntenic groups: Consider the ancestral genome with all of its chromosomes concatenated and with the ancestral genes blocked by their syntenic groups that will be conserved between the two daughter species. The concatenated ancestral genome is to be broken into l segments (conserved syntenies) with lengths of proportions p_1 through p_l. The last proportion, p_l, is determined from the other proportions as $p_l = 1 - (p_1 + p_2 + \cdots p_{l-1})$. If the number of breaks in any interval of the ancestral genome is modeled as Poisson, the realized lengths of the segments on the ancestral genome are modeled from an exponential distribution. The joint density function of the proportional lengths is given by $(l - 1)!$ over the region $p_1 > 0$, $p_2 > 0$, . . . , $p_{l-1} > 0$ and $p_1 + p_2 + \cdots + p_{l-1} < 1$. That is, the joint density of the proportional lengths is uniform over the ancestral genome scaled to unit length. This distribution is the member of the Dirichlet family of distributions (further described below) with all of its l parameters equal to one.

It may be preferable to model the lengths of the conserved syntenies or segments with several gamma distributions, all on the same scale, but with shapes that depend on the sizes of the chromosomes making up the pairs. In this case, the joint density function of the proportional lengths follows a Dirichlet distribution whose parameters are determined by the shape parameters of the gamma distributions (see Fristedt and Gray 1997, pp. 156–157). In Bayesian statistics language, this Dirichlet distribution on the proportional syntenic lengths is the conjugate prior to the multinomial probabilities that pairs of chromosomes from the two species contain orthologous genes; the parameters of the Dirichlet distribution may be chosen to take into account the relative lengths of the chromosomes in the two species. Choosing a nonuniform Dirichlet distribution amounts to choosing an informative rather than a noninformative prior distribution. The actual parameters chosen to model the chromosome lengths would impart the level of strength for the information given by the prior distribution.

Specifically, if the length of the block of genes from the ancestral genome that will constitute the orthologues of the jth chromosome pair is modeled by a gamma distribution with scale λ and shape parameter α_j, then the joint distribution of the proportional lengths follows a Dirichlet distribution with parameters $\{\alpha_j\}$. Let $\alpha = \Sigma\alpha_j$. The density function of this Dirichlet distribution is

$$f(p_1, p_2, \ldots, p_{l-1}|\alpha_1, \alpha_2, \ldots, \alpha_l) = \frac{\Gamma(\alpha)p_1^{(\alpha_1-1)}p_2^{(\alpha_2-1)} \cdots p_{l-1}^{(\alpha_{l-1}-1)}(1 - (p_1 + p_2 + \cdots p_{l-1}))^{(\alpha_l-1)}}{\Gamma(\alpha_1)\Gamma(\alpha_2) \cdots \Gamma(\alpha_l)}$$

over the region $p_1 > 0$, $p_2 > 0$, . . . , $p_{l-1} > 0$, and $p_1 + p_2 + \cdots + p_{l-1} < 1$.

Distribution of the total number of conserved syntenies: We assume that the proportional lengths of the syntenic segments are uniformly distributed. The uniform assumption is noninformative and corresponds to standard likelihood methods. The data consist of counts of orthologous gene pairs in the conserved syntenies found: $(n_1, n_2, \ldots, n_k)$. These counts are in a collection of k chromosome pairs. Another $l - k$ chromosome pairings to which orthologous genes have not yet been mapped actually contain orthologues yet to be discovered.

The likelihood function of the true total number of conserved syntenies, l, is found by integrating the multinomial distribution against the joint uniform distribution on these proportional lengths. We must include the number of ways to choose $l - k$ of the $m - k$ chromosome pairings to which orthologous genes have not yet been mapped to actually contain orthologues yet to be discovered and we must include the fact that these particular l conserved syntenies are only one choice out of all the equally likely collections of l conserved syntenies chosen from the m chromosome pairings. Then we have the modification of Theorem 1 of Sankoff and Nadeau (1996),

$$\begin{aligned}\mathrm{Lik}(l|(n_1, n_2, \ldots, n_k)) &= f(n_1, n_2, \ldots, n_k|l) \\ &= \int \frac{\binom{m-k}{l-k}}{\binom{m}{l}} \binom{n}{n_1, n_2, \ldots, n_k} \\ &\quad \times p_1^{n_1}p_2^{n_2} \cdots p_k^{n_k}(l-1)! \prod_{i=1}^{l-1} dp_i \\ &= \frac{\binom{m-k}{l-k}}{\binom{n+l-1}{l-1}\binom{m}{l}}\end{aligned}$$

for $l = k, k + 1, \ldots, m$, where the integral is over the region where $p_1 > 0$, $p_2 > 0$, . . . , $p_{l-1} > 0$, and $p_1 + p_2 + \cdots + p_{l-1} < 1$.

The maximum-likelihood estimator for the true number of conserved syntenies is then the value of l that maximizes the function above. This estimator depends on the total number of orthologous genes mapped (n) and the observed number of conserved syntenies (k). The maximum-likelihood estimator depends on the total number of pairs of chromosomes $(m = r \times c)$ between the two species only through the constraint that $l \leq m$ because

$$\frac{\binom{m-k}{l-k}}{\binom{n+l-1}{l-1}\binom{m}{l}} = \frac{\binom{l}{k}}{\binom{n+l-1}{l-1}\binom{m}{k}}.$$

The formula for the density of the counts of the number of orthologous genes in the k conserved syntenies found given that there are l conserved syntenies in total has the following probabilistic interpretation: The denominator is the number of ways of choosing l out of the total of m possible conserved syntenies to be filled times the number of ways to fill l conserved syntenies with n orthologous genes. The numerator is the number of ways of choosing $l - k$ of the unseen conserved syntenies from the $m - k$ possibilities. The probability space includes not only the actual counts $(n_1, n_2, \ldots, n_k)$ observed but also which of the m possible conserved syntenies (cells in the table) get those counts.

An interval estimate for the true number of conserved syntenies, l, can be obtained by recognizing that we have essentially calculated the posterior distribution of l given the noninformative prior distribution that each chromosome pairing has equal chance of ever containing or not containing orthologues and that the orthologues are uniformly distributed among the chromosome pairings that actually contain orthologues. Under this noninformative prior, the posterior distribution on l is simply proportional to the likelihood function of l. That is,

$$f(l|n_1, n_2, \ldots, n_k) \propto f(n_1, n_2, \ldots, n_k|l).$$

The proportionality constant required to give a probability distribution is given by

$$\frac{1}{C} = \sum_{l=k}^{m} f(n_1, n_2, \ldots, n_k|l) = \sum_{l=k}^{m} \frac{\binom{m-k}{l-k}}{\binom{n+l-1}{l-1}\binom{m}{l}}.$$

An interval estimate (at a 95% level) on the true number of conserved syntenies (of the form $[k, L]$ where k is the observed number and L is the upper bound on the number) is determined by finding the smallest value of L that satisfies

$$0.95 \leq \sum_{l=k}^{L} f(l|n_1, n_2, \ldots, n_k) = C\sum_{l=k}^{L} \frac{\binom{m-k}{l-k}}{\binom{n+l-1}{l-1}\binom{m}{l}}.$$

Syntenic correlation: We introduce a measure of syntenic correlation that can be used to compare genomic distances across many pairs of species. Similar measures have been developed by Bengsston *et al.* (1993) and discussed in Zakharov and Valeev (1988). For instance, Bengsston *et al.* take a pair-wise approach, counting the pairs of genes syntenic in both species and normalizing by the square root of the product of the number of syntenic pairs in each individual species. This measure, however, has a nonzero lower bound that depends on the probabilities that a pair of genes will be syntenic in each species. Our correlation measure falls between zero and one; it is one if the two genomes have identical syntenic groups and zero if the orthologous genes are randomly scattered between the two genomes.

Reverting to our original multivariate notation describing the $r \times c$ table summarizing the number of orthologues in each synteny, let $n_{i,j}$ be the observed number of genes on species A chromosome i with an orthologue on species B chromosome j. Let $e_{i,j}$ be the expected number of genes in the cell assuming that the genes are scattered independently in the two genomes. That is, $e_{i,j} = n_{\cdot,j}n_{i,\cdot}/n$, where $n_{i,\cdot}$ is the row total of the number of genes on species A chromosome i with an orthologue anywhere in species B's genome, $n_{\cdot,j}$ is the column total of the number of genes on species B chromosome j with an orthologue anywhere in A's genome, and n is the total number of orthologous genes mapped between the two species. Then a measure of syntenic correlation is given by

$$\rho = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{i,j} - e_{i,j})^2}{n \min\{r-1, c-1\} e_{i,j}}.$$

This measure of association has the following properties: It always makes sense as long as $0^2/0$ is interpreted as being 0; the value of ρ lies between 0 and 1; the value is 1 if and only if, for one of the two species, knowing which chromosome an orthologue belongs to in that species determines which chromosome the orthologue is on in the other species; the value is 0 if and only if the counts of orthologues are perfectly independently scattered on the chromosomes of the two species; and the value is not changed by reordering the chromosomes in the two species.

The use of this scaled chi-square statistic as a measure of association is not new. It was proposed by Cramér as a measure of the degree of dependence or association between the arguments of a contingency table (Cramér 1946). While we believe that ρ is a useful measure of syntenic correlation, other statisticians have argued against the use of modified versions of the chi-square statistic as a measure of the degree of association (Fisher 1938; Goodman and Kruskal 1954). We thus include another measure for comparison.

One of the alternative measures of association proposed by Goodman and Kruskal (1954; first proposed by Guttman 1941) would, in our application, measure the proportion of errors made in assigning a gene to a chromosome in one species that can be eliminated by knowing which chromosome the orthologue belongs to in the other species. Suppose an orthologue chosen at random must be assigned to a chromosome in a species. The most likely chromosome for this assignment is the one that contains the largest proportion of genes mapped and the chance of making an error is 1 minus this largest proportion. If additionally we know which chromosome in the other species contains the or-

thologue, then we consider only the distribution of orthologues that map to this chromosome; that is, we find the chromosome in the original species that contains the largest proportion of orthologues from this one chromosome in the other species. The probability of making an error when one knows which chromosome from the other species contains the orthologue is 1 minus the sum of these maximum proportions over all the chromosomes in the other species. The proposed measure of association λ is the difference between the probabilities of making an error with no information and with the chromosome of the other species known divided by the chance of making an error with no information from the other species. To obtain a symmetric measure, assume a gene is taken from each of the two species with probability 1/2 each.

Let $m_{i,\cdot}$ be the maximum number of orthologues mapped from species A chromosome i to any chromosome in species B. Similarly, let $m_{\cdot,j}$ be the maximum number of orthologues mapped from species B chromosome j to any chromosome in species A. Let m_A be the maximum number of genes mapped to any single chromosome in species A and m_B be the maximum number of genes mapped to any single chromosome in species B. Recall that n is the total number of genes mapped. The proposed measure of association is then

$$\lambda = \frac{\sum_{i=1}^{r} m_{i,\cdot} + \sum_{j=1}^{c} m_{\cdot,j} - (m_A + m_B)}{2n - (m_A + m_B)}.$$

This measure of association has the following properties: It makes sense as long as not all the orthologous genes mapped lie in only one chromosome pairing; the value of λ lies between 0 and 1; the value is 1 if and only if the counts of orthologues are concentrated in chromosome pairings (cells of the table), no two of which are in the same row or column; the value is 0 whenever knowing the chromosome on which an orthologue resides in the other species is of no help in determining the chromosome the gene resides on in the focal species; the value is not changed by reordering the chromosomes in the two species.

If the orthologues are scattered independently on the chromosomes of the two species, then this measure of chromosome prediction ability is 0. However, $\lambda = 0$ whenever the same chromosome in the focal species is most likely to contain a gene no matter which chromosome contains the orthologue in the other species. An example where $\lambda = 0$ without the genes being scattered independently may be constructed by ensuring that chromosome 1 of species A always contains more orthologues than any other chromosome from species A for each given chromosome in species B and that chromosome 2 of species B plays the same role for species B. Clearly the orthologues do not need to be independently scattered when constructing this example. Thus, this measure assesses the predictive value of the conditional distributions for gene assignments to chromosomes in the other species but it does not measure the randomness of the distribution of the orthologues among chromosome pairs.

RESULTS

To compare our method for estimating the true number of conserved syntenies to the method of Sankoff and Nadeau (1996), we consider the human-mouse data provided in Ehrlich *et al.* (1997): $k = 91$ observed conserved syntenies, $n = 1152$ orthologous genes mapped, $m = r \times c = 19 \times 22 = 418$ chromosome pairs. Using the Sankoff-Nadeau techniques, Ehrlich *et al.* (1997) reported an estimated 141 true total number of conserved syntenies between mouse and man. Our method gives a point estimate of 98 and a 95% interval estimate of [91, 105] conserved syntenies. Thus, modeling the dependency of the segment lengths on each other results in smaller estimates for the true number of conserved syntenies that will ultimately be found between mice and humans.

We report the observed, estimated, and 95% upper bound estimate of conserved syntenies between all species pairs of man, cow, rat, and mice in Table 1. We also include the cat-human data from Murphy *et al.* (2000). We report the measure of syntenic correlation and the measure of chromosome prediction between these pairs of species with 95% confidence intervals obtained through resampling procedures. Note that the syntenic correlation between humans and cats ($\rho = 0.66$) is not statistically significantly different from the syntenic correlation between mice and rats ($\rho = 0.69$) even though the time since divergence for humans and cats (~92 mya; Kumar and Hedges 1999) is much greater than for mice and rats (~40.7 mya; Kumar and Hedges 1999). These results are in keeping with the conclusions of Murphy *et al.* (2000) regarding the remarkable degree of conservation of genome organization between cats and humans.

DISCUSSION

Recent articles on estimating the total number of conserved syntenies or segments between pairs of species (Sankoff and Nadeau 1996; Ehrlich *et al.* 1997; Waddington *et al.* 1999; Kumar *et al.* 2001) use the approximation that the lengths of the syntenic groups or segments are independent of each other. This assumption is clearly only an approximation: In a finite genome, if one segment is unusually long, it forces the other segments to be shorter. While it is clearly true that, in practice, the lengths of any two syntenic groups or segments are only very weakly correlated, the joint dependency of the entire collection of these lengths contributes significantly to the estimators of the total number of conserved syntenies or segments.

TABLE 1

Statistical results

	Mouse (19)	Rat (20)	Cattle (29)	Human (22)	Cat (18)
Mouse	—	ρ: 0.69 ± 0.04 λ: 0.80 ± 0.03	ρ: 0.36 ± 0.07 λ: 0.48 ± 0.07	ρ: 0.31 ± 0.01 λ: 0.41 ± 0.02	
Rat	[58, 62, 68] (752)	—	ρ: 0.39 ± 0.10 λ: 0.51 ± 0.08	ρ: 0.32 ± 0.04 λ: 0.45 ± 0.05	
Cattle	[104, 138, 154] (416)	[94, 149, 174] (252)	—	ρ: 0.64 ± 0.05 λ: 0.70 ± 0.04	
Human	[157, 164, 170] (3521)	[99, 113, 122] (776)	[72, 84, 93] (482)	—	ρ: 0.66 ± 0.05 λ: 0.71 ± 0.06
Cat				[39, 44, 50] (324)	—

Entries above the diagonal are syntenic correlations with 95% confidence intervals obtained through resampling procedures. Entries below the diagonal are of the form [observed number of syntenies, maximum-likelihood estimate of the true number of syntenies, 95% upper bound on the true number of syntenies]. The number of orthologues used for the analysis is in parentheses underneath. Numbers in parentheses next to the species in the column headings are the numbers of autosomes. The data involving human-mouse-rat comparisons were taken from the Mouse Genome Database (2001) at Jackson Laboratory on July 28, 2001 (URL: http://www.informatics.jax.org/). The data involving comparisons with cattle were generated from BovBase (2001) from the Roslin Institute (http://www.ri.bbsrc.ac.uk/bovmap/arkbov/) and LocusLink from the National Institutes of Health (http://www.ncbi.nih.gov/genome/guide/human) in June, 2001. The human-cat data are from the article of Murphy *et al.* (2000).

Further, many recent approaches choose one member of the pair of species being compared to provide critical information for the model. Sankoff and Nadeau (1996) and Ehrlich *et al.* (1997) choose one of the two species to provide the number of chromosomal breakpoints in their model. They subtract this number from the total number of conserved syntenies to calculate the syntenic distance between the pair of species. Waddington *et al.* (1999) choose one of the two species to provide the chromosome lengths that go into their β-distribution model of segment lengths. This model uses one of the two species as a donor species and the other as a receiver species of conserved segments. Reversing the roles does not necessarily lead to the same estimate of the total number of conserved segments. Kumar *et al.* (2001) present their model as being useful when the relative order of markers or genes in a primary genome is known while only the synteny of the orthologous markers or genes is known in the secondary genome. The primary genome is concatenated and the conserved segments chosen from it are assumed to have lengths that are independently distributed and follow a gamma distribution with a shape and scale parameter to be estimated from the data.

In this article, we have demonstrated how to take the dependency of the number of genes in conserved syntenies into account when estimating the true total number of conserved syntenies and measuring syntenic correlation. Our methods are symmetrical and do not require the specification of a focal genome. We believe that extending our methods to estimating the total number of conserved segments is fundamentally more problematic. The following extension of our model to the problem of estimating the true number of conserved segments demonstrates why. The following is closely related to the Kumar *et al.* (2001) model with the shape parameter of their gamma distribution taken to be 1 so that the distribution is exponential.

Suppose we observed k conserved segments containing $n_1, n_2, \ldots, n_k$ orthologues, respectively, where $n = \Sigma n_i$ is the total number of orthologues mapped between the two species. Suppose that the actual l conserved segments from the ancestral genome have lengths that are independently distributed and follow an exponential distribution with parameter λ. Then the proportional lengths follow a uniform Dirichlet distribution (Fristedt and Gray 1997, pp. 156–157). To estimate the total number of conserved segments, l, consider the likelihood function

$$\begin{aligned}\mathrm{Lik}(l|n_1, n_2, \ldots, n_k) &= f(n_1, n_2, \ldots, n_k|l)\\ &= \int\binom{n}{n_1, n_2, \ldots, n_k}\\ &\quad \times p_1^{n_1}p_2^{n_2}\cdots p_k^{n_k}(l-1)!\prod_{i=1}^{l-1}dp_i\\ &= \frac{1}{\binom{n+l-1}{l-1}}\end{aligned}$$

$l = k, k + 1, \ldots$. This distribution is uniform on the probability space, which includes not only the actual counts $(n_1, n_2, \ldots, n_k)$ observed but also which k of the l total conserved segments get those counts.

This likelihood function has its maximum when $l = k$ [because $\mathrm{Lik}(l+1|n_1, n_2, \ldots, n_k) = l/(n+l)\ \mathrm{Lik}(l|n_1, n_2, \ldots, n_k)$]. In short, without an informative, proper, prior distribution on the true number of conserved segments or information about the actual observed segment lengths proportional to the length of the genome, our most likely single estimate of the total number of conserved segments that will ever be found is simply the number observed at present.

The difference between estimating the total number of conserved syntenies and the total number of conserved segments is that, in the case of conserved syntenies, we in effect assume a noninformative prior distribution to model which chromosome pairs will contribute a conserved synteny. Given that there are exactly l conserved syntenies, each combination of l chromosome pairs out of the m possible pairs is assumed to be equally likely. In the case of counting conserved segments, the noninformative prior is improper because there is no upper bound on the true number, l, and the result is that our best guess for the true number of conserved segments is the number of conserved segments observed (much as our best guess for the probability a randomly chosen new gene will land in each cell in the Oxford grid is simply the observed proportion of genes in the cell).

Additionally, the observed number of conserved syntenies is a sufficient statistic for estimating the total number of conserved syntenies but the same is not true for segments. In other words, the information encoded by the numbers of orthologues found in the observed conserved syntenies and by the number and positions of the observed conserved syntenies that is useful in estimating the total number of conserved syntenies is completely summarized by k, the observed number of conserved syntenies.

Mathematically, we compute the likelihood function for l, the total number of conserved syntenies, given the observed number, k, as

$$\mathrm{Lik}(l|k) = f(k|l) = \frac{\binom{m}{k}\binom{n-1}{k-1}\binom{m-k}{l-k}}{\binom{n+l-1}{l-1}\binom{m}{l}}.$$

This formula is obtained from the density function of the raw data by counting the number of ways to choose the k observed conserved syntenies from the m possible ones and the number of ways of distributing the n orthologues between those k conserved syntenies so that none are empty (Feller 1968, p. 38). Since these additional terms do not depend on l, we lose no information about l when we summarize the information given in the observed numbers of genes in the k observed conserved syntenies into just the number k. [Note that, after simplifying, the formula for $f(k|l)$ above for syntenies reduces to the formula for $f(k|l)$ below for segments.]

Because we have no upper bound for the number of conserved segments, we lose information about the total number of conserved segments when we summarize the data by reporting only the observed number. The likelihood function for the total number of conserved segments, l, given the observed number, k, is

$$\mathrm{Lik}(l|k) = f(k|l) = \frac{\binom{l}{k}\binom{n-1}{k-1}}{\binom{n+l-1}{l-1}},$$

which is obtained from the density function of the raw data by counting the number of ways to choose the k observed segments from the total number of segments, l, and distributing the n orthologues so that none of the k segments is empty. Since one of the additional terms does depend on l, we have different information about l when we summarize the data into just the count of the number of conserved segments. Note that this formula is given in Theorem 3 (Sankoff *et al.* 1997), although our conclusions about the sufficiency of k are at odds with their Theorem 2.

One way around these difficulties may be to use the following proper prior distribution: Assume an arbitrarily large, artificial number of possible conserved segments, m, and assume that, prior to obtaining data, each possible segment has equal chances of ever containing orthologues or not. This approach corresponds to the approach used in estimating conserved syntenies. For sufficiently large choices of m, the maximum-likelihood estimator for the true number of segments, l, will not depend on m and the posterior distribution of l will depend only weakly on m.

Neither the raw number of conserved segments nor the raw number of conserved syntenies provides an adequate measure of genomic distance. While the measures proposed by Bengsston *et al.* (1993) and discussed in Zakharov and Valeev (1988) have been criticized for failing to estimate the total number of conserved syntenies (both observed and unobserved) and for giving disproportionate weight to segments in which many genes have been mapped (Sankoff and Nadeau 1996; Ehrlich *et al.* 1997; Nadeau and Sankoff 1998), these measures do attempt to standardize genomic distances so that they can be compared across many pairs of species. Under the necessary and universal model assumption of random gene discovery, our proposed syntenic correlation provides a standardized measure of genomic distances that avoids all these difficulties. It can be used to compare the genomic distances of many pairs of

species, does not require the specification of a primary and secondary genome, does not give undue weight to segments in which many genes have been mapped (assuming random gene discovery), and relies on a modification of the well-understood chi-square statistic for testing independent gene scattering on the two genomes. Our correlation measures how far the orthologous genes are from being independently scattered on the two genomes.

The caveat to the above work, of ourselves and of others, is the typical caveat for all observational data: The orthologous genes that have been mapped must represent a random sample of all the orthologous genes that will be discovered. Indeed, the orthologues found so far may be the ones that are more easily found due to mutational constraints on their divergence and these constraints may also require higher levels of synteny correlation. The ones left to be discovered may be more divergent due to fewer restrictions on their evolution and this relaxation of mutational constraint may also allow them to be more scattered in the genome. Nonetheless, even if orthologues are eventually found on all chromosome pairs from the two species and even when the entire genomes of many pairs of species have been mapped, our syntenic correlation measure will provide a useful and nontrivial measure of syntenic conservation, allowing for the summary and comparison of Oxford grids for many pairs of species.

We thank Phuong Ngo-Hazelett for help with the construction and formatting of the Oxford grids analyzed in this article, Sasha Richardson for help constructing the synteny plot given in Figure 1, and Michael Lynch for suggestions improving the legibility of some of the formulas. We heartily thank David Sankoff and two anonymous reviewers for their constructive comments. One of the anonymous reviewers was particularly helpful, providing references for the use of scaled versions of the chi-square statistic as a measure of association and voicing concerns that enabled us to improve the article substantially. This work was supported by a National Science Foundation interdisciplinary grant in the mathematical sciences DMS 0075143 to E.A.H. and National Institutes of Health grant R01RR10715 to J.P.

LITERATURE CITED

Bengtsson, B. O., K. K. Levan and G. Levan, 1993 Measuring genome reorganization from synteny data. Cytogenet. Cell Genet. **64:** 198–200.

BovBase, 2001 The Roslin Institute, Edinburgh (http://www.ri.bbsrc.ac.uk/bovmap/arkbov/), July 28, 2001.

Cramér, H., 1946 *Mathematical Methods of Statistics*. Princeton University Press, Princeton, NJ.

Ehrlich, J., D. Sankoff and J. H. Nadeau, 1997 Synteny conservation and chromosome rearrangements during mammalian evolution. Genetics **147:** 289–296.

Feller, W., 1968 *An Introduction to Probability Theory and Its Applications*, Vol. I. John Wiley & Sons, New York.

Fisher, R. A., 1938 *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.

Fridstedt, B., and L. Gray, 1997 *A Modern Approach to Probability Theory*. Birkhäuser, Boston.

Goodman, L. A., and W. H. Kruskal, 1954 Measures of association for cross classifications. J. Am. Stat. Assoc. **49:** 732–764.

Graur, D., and W.-H. Li, 2000 *Fundamentals of Molecular Evolution*. Sinauer Associates, Sunderland, MA.

Guttman, L., 1941 An outline of the statistical theory of prediction. Supplementary study B-1, pp. 253–318 in *The Prediction of Personal Adjustment*, edited by P. Horst, P. Wallin and L. Guttman. Bulletin 48, Social Science Research Council, New York.

Hannenhalli, S., C. Chappey, E. V. Koonin and P. A. Pevzner, 1995 Genome sequence comparison and scenarios for gene rearrangements: a test case. Genomics **30:** 299–311.

Kumar, S., and S. B. Hedges, 1999 A molecular timescale for vertebrate evolution. Nature **392:** 917–920.

Kumar, S., S. R. Gadagkar, A. Filipski and X. Gu, 2001 Determination of the number of conserved chromosomal segments between species. Genetics **157:** 1387–1395.

Mouse Genome Database (MGB), 2001 Mouse Genome Informatics Web Site, The Jackson Laboratory, Bar Harbor ME (http://www.informatics.jax.org/), July 28, 2001.

Murphy, W. J., S. Sun, Z. Chen, N. Yuhki, D. Hirschmann *et al.*, 2000 A radiation hybrid map of the cat genome: implications for comparative mapping. Genome Res. **10:** 691–702.

Nadeau, J. H., and D. Sankoff, 1998 Counting on comparative maps. Trends Genet. **14:** 495–501.

Sankoff, D., and J. H. Nadeau, 1996 Conserved synteny as a measure of genome rearrangement. Discrete Appl. Math. **71:** 247–257.

Sankoff, D., G. Leduc, N. Antoine, B. Paquin, B. F. Lang *et al.*, 1992 Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome. Proc. Natl. Acad. Sci. USA **89:** 6575–6579.

Sankoff, D., M.-N. Parent, I. Marchland and V. Ferretti, 1997 On the Nadeau-Taylor theory of conserved chromosome segments, pp. 262–274 in *Combinatorial Pattern Matching. Eighth Annual Symposium*, edited by A. Apostolico and J. Hein. Lecture Notes in Computer Science 1264, Springer Verlag, Berlin.

Waddington, D., A. Springbett and D. W. Burt, 1999 A chromosome-based model for estimating the number of conserved segments between pairs of species from comparative genetic maps. Genetics **154:** 323–332.

Zakharov, I. A., and A. K. Valeev, 1988 Quantitative analysis of evolution of mammalian genomes by comparison of genetic maps. Proc. Acad. Sci. USSR **301:** 1213–1218. Genetika **28:** 77–81.

Communicating editor: G. A. Churchill