Determination of the Number of Conserved Chromosomal Segments Between Species

Sudhir Kumar,* Sudhindra R. Gadagkar,* Alan Filipski* and Xun Gu†

**Department of Biology, Arizona State University, Tempe, Arizona 85287-1501 and †Zoology/Genetics Biology, Iowa State University, Ames, Iowa 50011*

Manuscript received July 28, 2000
Accepted for publication December 4, 2000

ABSTRACT

Genomic divergence between species can be quantified in terms of the number of chromosomal rearrangements that have occurred in the respective genomes following their divergence from a common ancestor. These rearrangements disrupt the structural similarity between genomes, with each rearrangement producing additional, albeit shorter, conserved segments. Here we propose a simple statistical approach on the basis of the distribution of the number of markers in contiguous sets of autosomal markers (CSAMs) to estimate the number of conserved segments. CSAM identification requires information on the relative locations of orthologous markers in one genome and only the chromosome number on which each marker resides in the other genome. We propose a simple mathematical model that can account for the effect of the nonuniformity of the breakpoints and markers on the observed distribution of the number of markers in different conserved segments. Computer simulations show that the number of CSAMs increases linearly with the number of chromosomal rearrangements under a variety of conditions. Using the CSAM approach, the estimate of the number of conserved segments between human and mouse genomes is 529 ± 84, with a mean conserved segment length of 2.8 cM. This length is <40% of that currently accepted for human and mouse genomes. This means that the mouse and human genomes have diverged at a rate of ~1.15 rearrangements per million years. By contrast, mouse and rat are diverging at a rate of only ~0.74 rearrangements per million years.

AFTER a speciation event, descendant genomes may diverge in overall structure as a result of intra- and interchromosomal rearrangements. Each rearrangement reduces the structural homology between the two genomes, while increasing the number of homologous chromosomal fragments. These homologous, but repositioned, fragments are referred to as conserved segments between the genomes compared. The genomic divergence due to chromosomal rearrangements between species can be estimated in terms of the number of conserved segments between their genomes (*e.g.*, Sankoff and Nadeau 1996; Ehrlich *et al.* 1997).

Conserved segments are identified by examining the relative order of contiguous landmarks in the chromosomes of the species being compared. Protein coding genes are frequently used as landmarks because they are numerous and their orthology relationships can be determined with great certainty even among distantly related species because of high levels of protein sequence conservation. For most organisms, however, genomic map information for only a fraction of these landmarks is currently available. As a consequence, many conserved segments are not "visible." A second reason for a conserved segment being unobserved is that there may simply be no identifying markers in those regions in one or both genomes.

Historically, genomic map information has consisted mostly of the knowledge of the chromosome number for a given gene in a genome. With recent genome-sequencing efforts and better mapping techniques, information is becoming available on the relative order as well as the actual physical location of genes. However, this progress has been made for only a select group of species. For these reasons, approaches that require only the chromosomal number for genes (conserved synteny approach) continue to be used (*e.g.*, Bengtsson *et al.* 1993; Zakharov *et al.* 1995; Sankoff and Nadeau 1996). A chromosome from one species is said to have a "conserved synteny" with a chromosome from another species if they have one or more markers in common. Thus, these measures require only knowledge of the chromosome number for the markers in both genomes.

However, statistical approaches utilizing conserved synteny data are known to have serious shortcomings. In particular, this method will provide only a lower bound of the number of observable conserved segments, and the number of conserved syntenies between two chromosomes cannot exceed $c_a \times c_b$, where c_a and c_b are the number of chromosomes in species a and b, respectively (Sankoff and Nadeau 1996).

For these reasons, more extensive genome map infor-

Corresponding author: Sudhir Kumar, Life Sciences 371, Department of Biology, Arizona State University, Tempe, AZ 85287-1501. E-mail: s.kumar@asu.edu

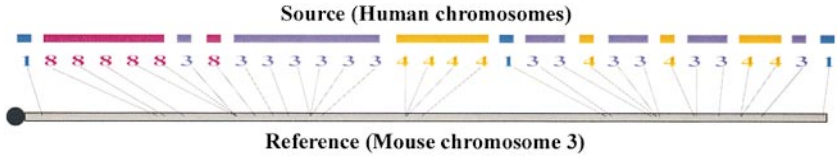

FIGURE 1.—A portion of mouse chromosome 3 showing the chromosome numbers of corresponding markers in the human genome (from July 6, 2000 release of the Mouse Genome Database). The number of CSAMs between the depicted portion of the mouse chromosome 3 and human chromosomes 1, 3, 4, and 8 are three, six, four, and two, respectively. Therefore, there are only four conserved syntenies but at least 15 conserved segments in this region.

mation needs to be used whenever available. The ideal approach for estimating the total number of conserved segments is to use conserved linkage data, which requires the knowledge of the relative order of markers in both genomes. This requirement, however, makes this approach impractical. For instance, the map location for most known human genes is still in the form of the cytogenetic band they reside in, and even for the more precisely mapped mouse genome, the relative gene order is known for only a few thousand genes (see BLAKE *et al.* 2000). In short, it is not yet possible to find many pairs of mammalian species for which a substantial number of conserved linkages can be identified. Therefore, we use an approach intermediate to the conserved synteny and conserved linkage approaches, which utilizes currently available information more effectively than the conserved synteny approach. In our intermediate approach, we use contiguous sets of autosomal markers (CSAMs). A CSAM is an uninterrupted set of markers in one genome (primary genome) that are syntenic in the other genome (secondary genome; Figure 1). Therefore CSAMs can be identified using relative marker order information in one genome and the chromosome number of those markers in the other genome. This intermediate approach differs from others (*e.g.*, NADEAU and TAYLOR 1984; WADDINGTON *et al.* 2000) in that no information is required about the physical distances between markers, which are not yet known with precision for most genomes.

In this article we discuss the relationship of the number of CSAMs with the number of rearrangements and present a mathematical formulation for estimating the number of unobserved CSAMs. We also show the usefulness of our approach by computer simulation and empirical data analyses with data from human, mouse, and rat genomes.

A MODEL FOR CSAM SIZE DISTRIBUTION

Let us consider the entire set of autosomal chromosomes linked together in a linear head-to-tail fashion to form a superchromosome, whose total length is denoted by C. Let there be n conserved segments in this genome, and m genes (markers) residing on these n conserved segments. If we consider each conserved segment to be a separate bin, irrespective of its true physical length (which is not known beforehand), the probability of observing k genes in a given conserved segment can be described by the Poisson distribution,

$$P(k|x) = \frac{x^k}{k!}e^{-x}, \tag{1}$$

where x, the expected number of genes in that segment, depends upon factors such as the physical length of the conserved segment and local marker density. The distribution of x over different conserved segments can be described by a gamma distribution with shape and size parameters, α and β, respectively,

$$\varphi(x) = \frac{\beta^x}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}. \tag{2}$$

Here we are proposing to use a gamma distribution to model an overdispersed Poisson distribution that is needed to describe counts of genes in a conserved segment. In this case, a more realistic distribution of genes and nonuniformity of breakpoints is modeled by the shape parameter, α. (β is a scaling factor.) It is worth noting that this gamma distribution is intended to model the observed data, which is affected by the extent of marker sampling, different types of chromosomal rearrangements (and their unknown relative proportions), and unknown differences in marker densities throughout the genome. For this reason, α is not a fundamental biological quantity, but is a descriptor of the observed data.

From Equations 1 and 2, it can be shown that the number of genes in a given conserved segment is expected to follow the negative binomial distribution

$$P(k) = \frac{\Gamma(\alpha + k)}{k!\Gamma(\alpha)}\left(\frac{1}{1+\beta}\right)^k\left(\frac{\beta}{1+\beta}\right)^\alpha, \quad \text{where } k = 0, 1, 2, \ldots. \tag{3}$$

This is a compound Poisson-gamma distribution, with the unobserved gamma random variable integrated out to give a negative binomial distribution of counts.

In Equation 3, there are n_0 zero-gene segments that cannot be observed. Therefore, we use a truncated negative binomial distribution to estimate n,

$$Q(k) = \frac{P(k)}{1 - P(0)}, \quad \text{where } k = 1, 2, 3 \ldots \tag{4}$$

so that $\Sigma_{i=1}^{\infty} Q(k) = 1$. In Equation 4,

$$P(0) = \left(\frac{\alpha n}{m + \alpha n}\right)^{\alpha}, \quad (5)$$

which is obtained by solving (3) with $k = 0$, noting that $\alpha/\beta = m/n$.

n can be estimated by equating the observed and expected values of the first and second moments. For the truncated binomial distribution given in (4), the first two moments about zero are given by

$$E[k] = \frac{m}{n[1 - P(0)]} \quad \text{and} \quad E[k^2] = \left(\frac{m}{\alpha n} + \frac{m}{n} + 1\right)\frac{m}{n[1 - P(0)]}, \quad (6)$$

respectively.

If $\bar{k}$ and $\bar{k}^2$ are observed first and second moments, respectively, and we define

$$b = \frac{\bar{k}}{\bar{k}^2} - 1, \quad (7)$$

then it can be shown that the estimate of n is the positive solution of

$$\ln\left(1 - \frac{m}{n\bar{k}}\right) = \frac{m}{bn - m}\left[\ln\frac{n}{n(1 + b) - m}\right]. \quad (8)$$

The estimate of n can be obtained by iteration. Sometimes the iterative procedure fails to converge, *e.g.*, when a large number of unobserved segments needs to be estimated using a relatively small number of markers. In this case, we suggest smoothing of the observed segment size distribution (*e.g.*, by taking a moving average with a window of size three) and truncating the tail (which often contains segments with functionally linked genes).

Once n is estimated, then the parameter modeling the nonrandomness of the gene distribution (α) and the scale parameter, β, are given by

$$\alpha = \frac{m}{bn - m} \quad \text{and} \quad \beta = \alpha\frac{n}{m}. \quad (9)$$

Now, given n and the number of observed CSAMs, the number of unobserved CSAMs (n_0) can be estimated. The standard error of the estimates can be obtained by the bootstrap procedure (Efron and Tibshirani 1993) in which the CSAMs are sampled with replacement. We suggest the resampling of CSAMs rather than markers, because the CSAM is the unit of measure. The same statistical estimation method can be applied to conserved syntenies, CSAMs, and conserved linkages.

Once the number of conserved segments has been estimated, an approximate (and conservative) estimate of the number of rearrangements can be obtained by $R = \frac{1}{2}(n - \max(c_a, c_b))$, where c_a and c_b are the number of chromosomes in species a and b, respectively, and n is the estimate of total number of conserved segments (observed + unobserved). The factor of $\frac{1}{2}$ is based on our computer simulations involving different proportions of different types of intra- and interchromosomal rearrangements (Table 1).

TABLE 1

Number of CSAMs per rearrangement (C/R) for different combinations of inter- and intrachromosomal rearrangements

% rearrangements				
Interchromosomal		Intrachromosomal		
Reciprocal	Nonreciprocal	"Simple"	In-place inversions	C/R
100	0	0	0	1.90
90	0	5	5	1.90
80	10	5	5	1.96
70	10	10	10	1.95
70	0	15	15	1.86
50	10	20	20	1.88

Note that the Sankoff and Nadeau (SN) model (Sankoff and Nadeau 1996) is not a direct special case of our model, although results with $\alpha = 1$ in our model produce estimates of n that are close to those obtained using their model (Sankoff and Nadeau 1996, p. 251). There have been some other recent sophisticated approaches to estimate the number of conserved segments (*e.g.*, Burt *et al.* 1999; Waddington *et al.* 2000). Waddington *et al.* assume that smaller chromosomes contain fewer conserved segments, and larger ones contain more, and develop a model that accounts for chromosome size differences. However, it is not clear that there is a necessary correlation between chromosome size and the number of conserved segments as the chromosome sizes can change considerably during the evolutionary history of species (*e.g.*, by simple fission) at different times. Also the empirical data do not seem to support this assumption (see Figure 5). Our model avoids making these types of assumptions.

COMPUTER SIMULATION

To assess the usefulness of our new approach and compare it to other approaches, we conducted computer simulations in the following manner. The process begins by creating a genome consisting of c chromosomes of specified lengths, with the location of centromeres assigned randomly. Positions of the given number (m) of genomic markers are then determined at random either under a uniform distribution (U) or clumped distribution (N) of marker density. Under the "clumped" scheme, a probability p and a distance d are specified for the clumping (we used $p = 0.5$ and $d = 0.04$ cM). With probability p, each marker is selected to be within distance d of the previously chosen marker on that chromosome arm, and with probability $(1 - p)$ it is chosen from a uniform density over the entire chromosomal arm.

TABLE 2

Initial length and breakpoint weight used in the computer simulations

Chromosome	Relative length	Relative breakpoint weight
1	2.0	1.2
2	1.8	1.0
3	1.4	0.6
4	1.4	1.4
5	1.2	1.9
6	1.2	2.1
7	1.2	0.5
8	1.2	2.8
9	1.0	0.7
10	1.0	1.6
11	1.0	1.4
12	1.0	2.5
13	0.8	2.2
14	0.6	3.1
15	0.6	1.2
16	0.6	0.7
17	0.6	1.3
18	0.6	1.5
19	0.4	1.8
20	0.4	3.5

Data are from Figure 3 of Ayme *et al.* (1976). Relative lengths are proportional to the depicted "frequency" expected according to the chromosomal length. Relative breakpoint weight is the ratio of the "frequency of all breaks observed" to the "frequency expected according to the chromosome length" as given in Ayme *et al.* (1976).

A specified number of rearrangements is then applied to this set of chromosomes to produce the second genome.

Simulating the process of rearrangement requires the selection of breakpoints for excision and insertion of chromosomal segments. In both cases, the breakpoints are chosen restricted to the chosen arm to ensure that each resultant chromosome has exactly one centromere. Breakpoints were also selected using a uniform (U) or a nonuniform (N) distribution. For the uniform chromosomal breakage scheme, the density function for selection of the breakpoint is uniform over the entire arm of the chromosome. In the nonuniform case, the assumption of uniformity is relaxed by specifying an initial length and a breakpoint weight, which is roughly proportional to the current chromosomal length and the observed breakpoint rate of the first 20 autosomal chromosomes in humans (Table 2). As each rearrangement is applied, the chromosomal set is updated to reflect the new position of the segment(s), while retaining identification of the original source chromosome. This process is repeated for a specified number of rearrangements. At the end of this process, each chromosome consists of an ordered list of segments, along with their lengths, origins, and other information. At this point markers are sprinkled onto the original chromosomes, as described above, and their evolutionary trajectories are computed. This information was used in this study for comparing true and estimated values of desired quantities.

Chromosomal rearrangement can take place in various ways. In this article, we consider both inter- and intrachromosomal rearrangement. Interchromosomal rearrangements considered were *simple translocations*, [an end (terminal) piece of one chromosome breaks off and attaches itself to the end of another chromosome], *reciprocal translocations* (two chromosomes exchange portions from their respective ends or terminals), and *intercalary transpositions* (the movement is from a nonterminal piece of one chromosome to a nonterminal position on another chromosome). Among the intrachromosomal rearrangements we considered *simple transpositions* (a fragment moves from one part of a chromosome to another part of the same chromosome) and *in-place inversions*. The simulation results we present here are from either 100% reciprocal rearrangements (U) or a particular mix of rearrangements (N), consisting of 10% intercalary transpositions, 50% reciprocal translocations, 20% simple intrachromosomal transpositions, and 20% in-place inversions.

Computer simulations denoted by UUU refer to the cases where the marker distribution was uniform, the breakpoint distribution was uniform, and the rearrangements were entirely reciprocal interchromosomal translocations. We use NNN to denote the cases in which the marker distribution was clumped (with probability = 0.5 that a given marker was to be located within 0.04 cM of the previous marker), the breakpoint distribution was based on the human chromosome size and breakpoint distribution as given in Table 2, and the rearrangement mix was in the proportions given in the previous paragraph. For brevity, extensive results from other intermediate combinations are not presented and are discussed when necessary.

RESULTS

Temporal distribution of conserved CSAM sizes: Figure 2 shows the expected patterns of generation of CSAMs (A and B) and conserved syntenies (C and D) with increasing number of chromosomal rearrangements, for UUU and NNN cases (solid circles). It is clear from Figure 2 that CSAMs accumulate linearly with increasing number of rearrangements and that additional CSAMs continue to accumulate at the same rate even when the number of rearrangements is very large. (The number of CSAMs per rearrangement is approximately two even for a very large number of rearrangements.) Furthermore, these relationships generally hold for both UUU and NNN cases. The slight discrepancy in the initial stages for the number of CSAMs per rearrangement (Figure 2B) is due to the

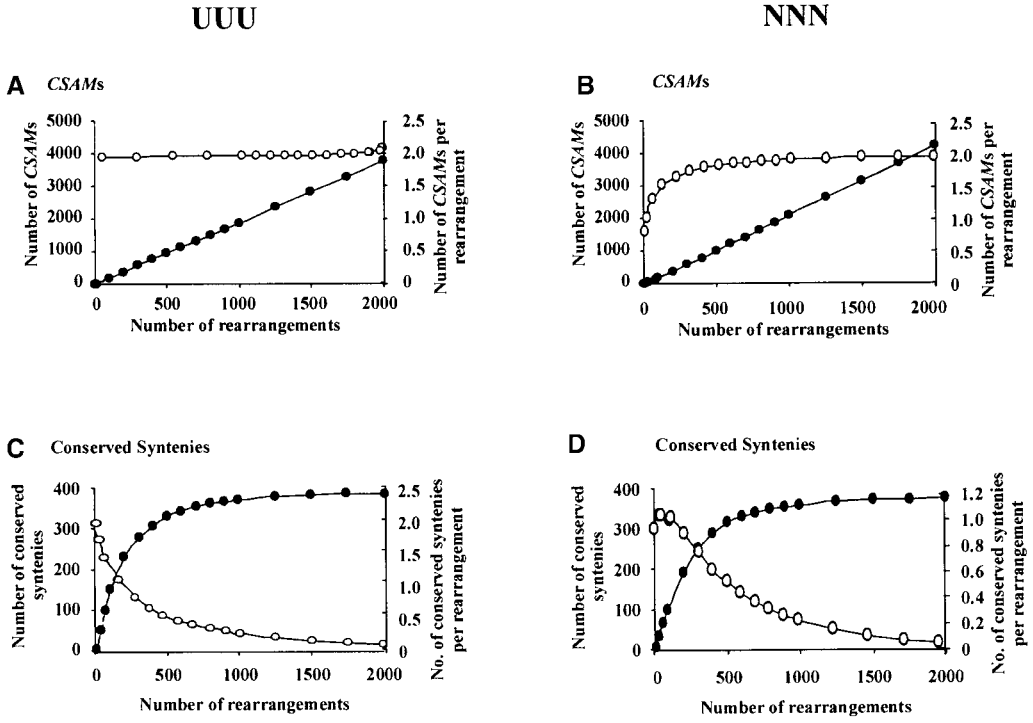

FIGURE 2.—Relationship of the expected numbers of CSAMs and conserved syntenies with the number of rearrangements (solid circles). A and C depict these relationships under the uniform case (UUU), whereas B and D show them under the nonuniform conditions (NNN). Open circles show the number of CSAMs and conserved syntenies per rearrangement. The results are from 100 simulation replicates.

inclusion of inversions in the rearrangement types (Table 1). In the initial stages, inversions do not result in new conserved segments. In the later stages, however, inversions may create more than one new segment, as, for example, if the fragment that is inverted is straddling two different conserved segments before it gets inverted. The consequences of these scenarios are reflected in Figure 2B. Our computer simulations also showed that CSAMs, although intermediate between conserved syntenies and conserved linkages, underestimate the number of conserved linkages by only 5–15%.

As expected, the number of conserved syntenies does not increase linearly with the number of rearrangements. Even though the shape of the curve in Figure 2C (solid circles) suggests weak linearity in the early portion, the number of syntenies per rearrangement (open circles) declines quickly even in this portion. As expected, the number of syntenies shows an upper bound of 400 because the genomes compared contain 20 chromosomes each. This nonlinear relationship of the true number of conserved syntenies with the number of rearrangements means that even the perfect estimation of all unobserved conserved syntenies will produce biased (lower) estimates of the number of conserved segments. In our simulations we also computed the Q value of BENGTSSON *et al.* (1993) and found that this statistic behaves in a manner similar to conserved syntenies, as it uses the number of observed conserved syntenies (results not shown).

Estimation of the number of conserved segments: Using landmarks such as genes we can count the number of observed conserved segments, *e.g.*, by counting the number of CSAMs containing at least one marker. However, as mentioned earlier, the number of unobserved segments needs to be estimated. Accurate estimation of this quantity is important to compute the genomic distance between the two species.

The accuracy with which the number of unobserved conserved segments can be estimated depends upon the accuracy with which the histogram of the observed conserved segments can be modeled. Figure 3 shows the temporal changes in the expected histogram of conserved segments (CSAMs) obtained from computer simulation (NNN). Before rearrangement, both species have 20 chromosomes with 50 genes each, and thus 20 conserved segments between the two species, each of size 50. With each rearrangement, the number of conserved segments increases and, consequently, the average segment size decreases. Obviously, when the number of rearrangements is few there are many large-sized segments, due to historical reasons. With time the number of rearrangements increases, which increases the number of small segments and reduces the number of large segments (Figure 3, A–C). In fact, Figure 3A shows that when there are only 50 rearrangements, there are 3 segments that are still of size 50; that is, they have not yet been broken up. These segments quickly get broken up, however, as the number of rearrangements increases (Figure 3, B and C).

For comparison, each panel of Figure 3 also shows the fit of the gamma model with the best-fit α-value and that of the gamma model with $\alpha = 1$, which approximates the SN model closely. The gamma model fits the observed conserved segments better, especially for

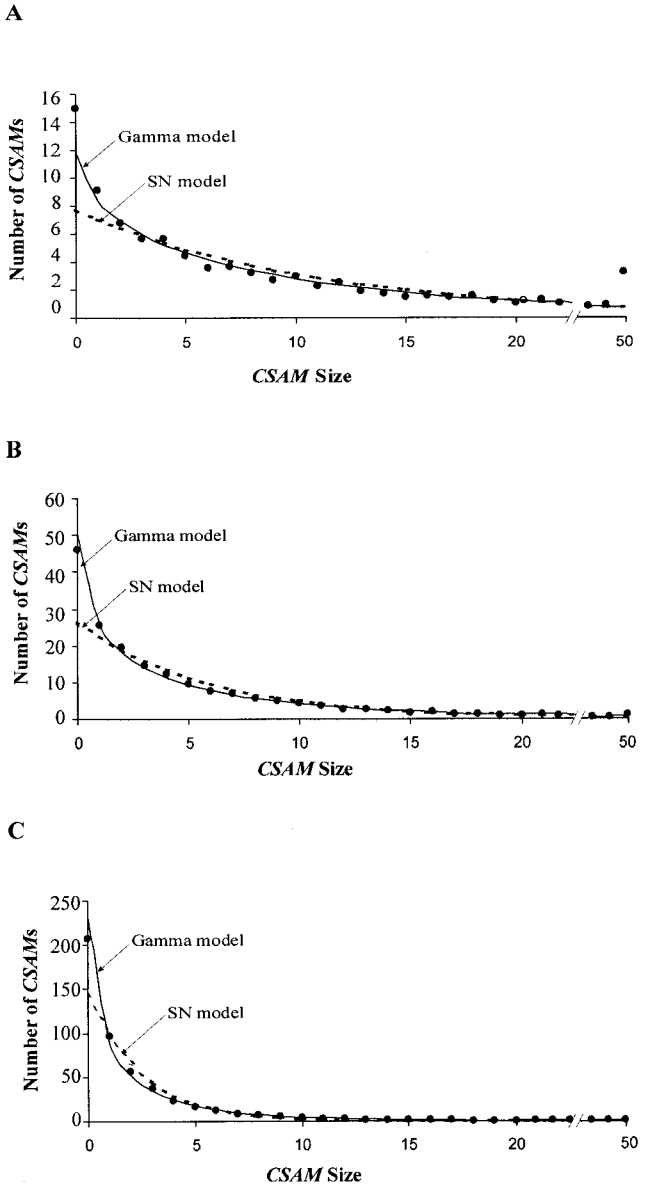

FIGURE 3.—True distribution of CSAMs after (A) 50, (B) 100, and (C) 250 rearrangements from 100 computer simulation replicates involving 1000 genes and nonuniform conditions (NNN). The fits of the gamma (solid line) and Sankoff-Nadeau (SN; broken line) models are shown. The values of the gamma parameter (α) for A, B, and C were 0.768, 0.558, and 0.455, respectively.

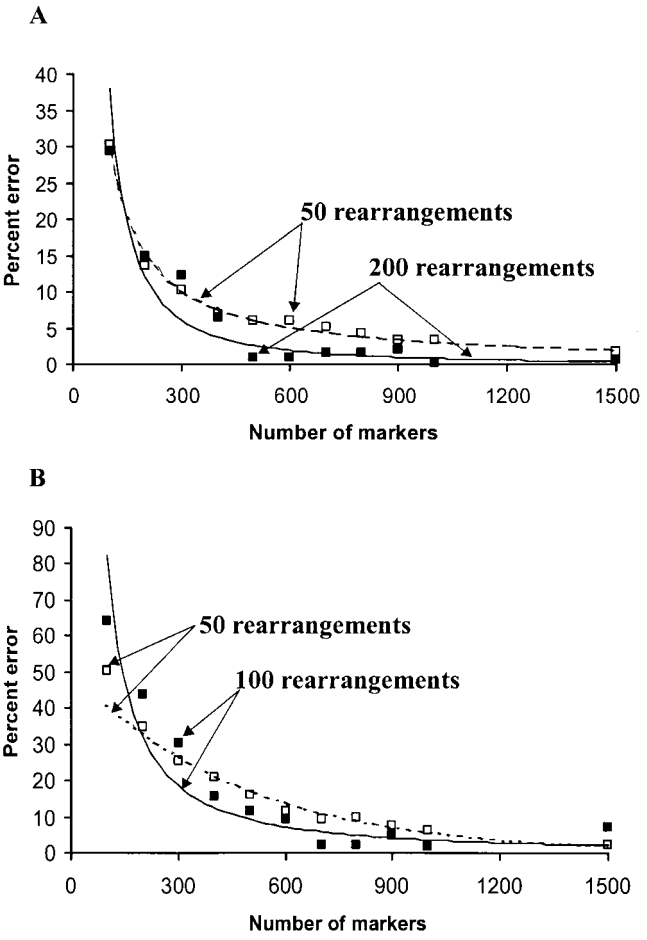

FIGURE 4.—Percentage error in the estimate of the number of conserved segments for different numbers of rearrangements, for conditions UUU (A) and NNN (B).

CSAMs containing small numbers of markers. For the complete uniform (UUU) case, fixed ($\alpha = 1$) as well as best-fit gamma models provide equally good fits to the observed distributions, as expected (results not shown).

As mentioned earlier, Equation 4 can be used to estimate the number of unobserved CSAMs. This involves estimation of α, which is generally a difficult problem when the 0 category is not available and the true value of α is small due either to the availability of only a small number of markers or because of a large number of rearrangements. This is because the probability distribution becomes L-shaped. All else being equal, increased nonuniformity of marker or breakpoint distribution also reduces α. If the value of α is fixed *a priori* (as in the case of our approximation of the SN model), estimation of n_0, the number of unobserved CSAMs, is straightforward. Concurrent estimation of α (along with n_0), as in the CSAM approach, yields good estimates of n_0 as long as the assumptions are UUU or NUN. When the distribution of chromosomal breakpoints is nonuniform, however, estimation of α is less reliable and can be significantly biased (n_0 is often overestimated). These results illustrate the difficulty in estimating the number of conserved segments, even when a single-parameter model is used. For this reason, parameter-rich models that attempt to estimate this quantity may experience difficulty without a large number of genes.

Number of markers needed to estimate genomic distance: Genomes of mammals generally consist of a very large number of genes (*e.g.*, 30,000–130,000 genes in humans; SCHULER *et al.* 1996; SCOTT 1999; EWING and GREEN 2000; LIANG *et al.* 2000; ROEST CROLLIUS *et al.* 2000). Clearly, it would be useful to know the minimum number of markers needed to reliably estimate the total number of conserved segments. Figure 4 shows the rela-

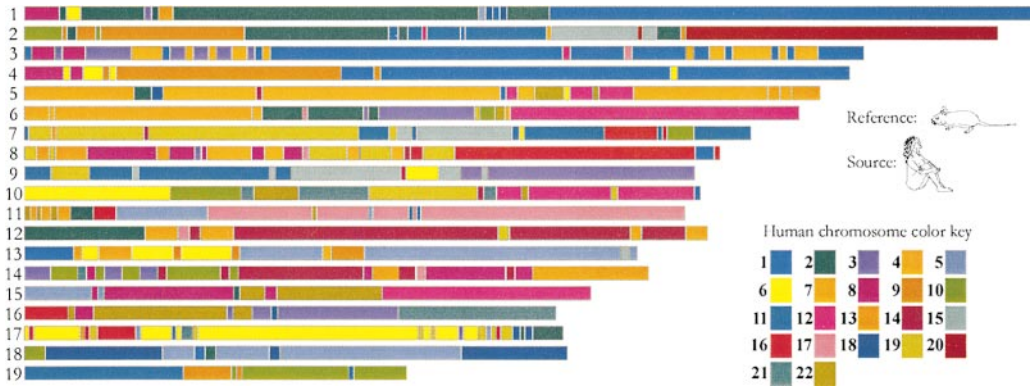

FIGURE 5.—A schematic depicting the comparative organization of the human and mouse genomes in a CSAM analysis of 2239 genes. Human chromosomes are denoted with a specific color shown in the color key, and the mouse chromosomes are painted by the human chromosome-specific colors for homologous markers. The chromosome lengths are proportional to their sizes in terms of the number of base pairs. Each observed conserved segment predicted by CSAM analysis is drawn proportional to the CSAM size (in terms of the number of genes), rather than chromosomal position of the genes, to better reflect the chromosomal homology as a function of the number of genes in each conserved segment.

tionship between the error (percentage difference between the true and the estimated value) and the number of markers for UUU (Figure 4A) and NNN (Figure 4B) cases, for different numbers of rearrangements. We find that we need 700 markers or fewer (depending on the number of rearrangements) to obtain estimates of the number of conserved segments (and, thus, the number of rearrangements) within 5% of the true value, as long as the assumptions are UUU. If the assumptions are NNN (extreme nonuniform case), then the number of markers needed for a maximum error margin of 5% is >1000 (but <1200). For divergence levels greater than shown the error margins are fairly high, even for large numbers of markers, underscoring the problems with estimation of the number of conserved segments when the observed distribution of CSAMs is L-shaped (see Figures 3 and 6). As mentioned earlier, the simulation of chromosomal breakpoints may not be appropriate, creating extremely low α-values for larger numbers of rearrangements. If the real data fall between these two extreme scenarios (UUU and NNN), we find that we need a maximum of 1200 markers for good results, as long as the two species being compared are fairly closely related. For instance, for NUN we found that ~1000 markers were needed for an error margin of <5%, even for 200 rearrangements. SCHOEN (2000) used computer simulation to estimate the number of breakpoints with differing numbers of chromosomes, markers, and rearrangement events and reported a declining error in the estimate with increase in the number of markers. He found that for lower number of markers the error rates increased slightly with the number of rearrangements and that the type of rearrangement (translocation *vs.* inversion) did not affect the error rate.

DISCUSSION

In this article we have introduced the CSAM approach and compared it to the approach of SANKOFF and NADEAU (1996). We have shown that the enumeration of conserved segments by using conserved syntenies as a unit has limitations (*e.g.*, Figure 2). While such limitations have been pointed out in the literature (SANKOFF and NADEAU 1996), the severity of the problem has not been clear previously. We have demonstrated that the CSAM approach remedies many of these problems while employing the SANKOFF and NADEAU (1996) concept of estimating the number of unobserved segments from the distribution of observed segments. However, we have proposed a more flexible distribution to account for the effect of nonuniformity in the marker and breakpoint distributions—a biologically more realistic scenario.

As the CSAM approach requires relatively few markers for accurate estimation of the number of conserved segments, we now discuss its utility in establishing the extent of chromosomal homology between human, mouse, and rat genomes. In the first comparison, we use the mouse genome as the primary and the human genome as the secondary genome. This is because of the finer relative position information available for mouse genes (BLAKE *et al.* 2000). In this comparison, 310 CSAMs are directly observable (Figure 5). For the same set of markers, the conserved synteny analysis exposes only 143 conserved segments (conserved syntenies). The SANKOFF and NADEAU (1996) method of estimating the number of unobserved syntenies predicts 8 additional conserved syntenies. Thus, the total number of conserved segments is predicted to be only 151. This and other similar estimates have been previously obtained and used to indicate the minimum number of conserved segments between human and mouse (NADEAU and TAYLOR 1984; COPELAND *et al.* 1993; SANKOFF and NADEAU 1996; EHRLICH *et al.* 1997; NADEAU and SANKOFF 1998). This number is less than half the number of conserved segments identified through CSAMs.

The high genomic divergence between human and mouse genomes is also evident in the observed CSAMs (Figure 5). Many chromosomes show areas of high breakpoint frequency (probable hotspots for rearrangements). In these areas many conserved segments of small size appear to have been produced by in-place

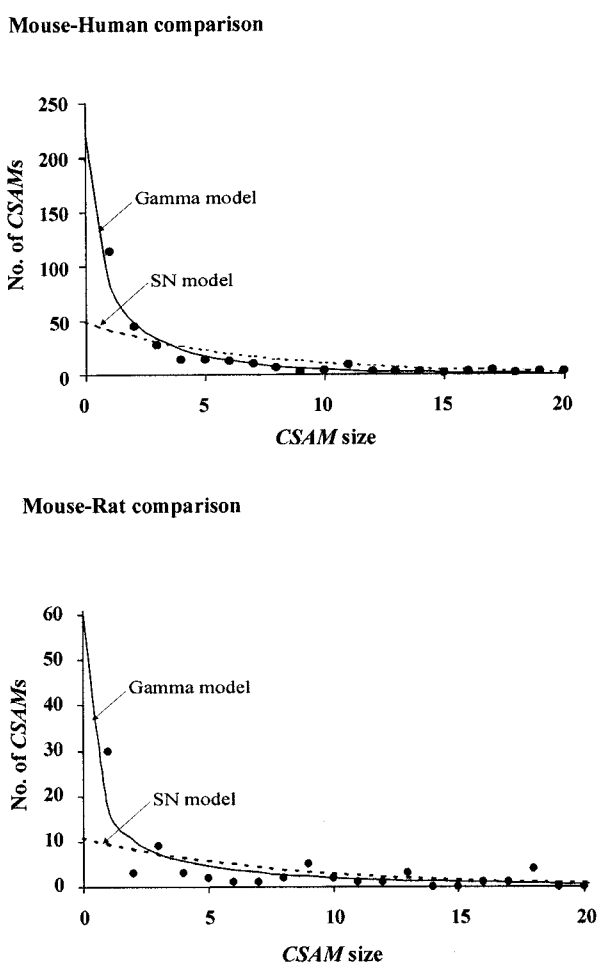

FIGURE 6.—Distribution of the CSAM size frequencies in (A) mouse-human comparison and (B) mouse-rat comparison. The mouse genome was used as the reference in both cases. Fit of the gamma distribution (solid line) and SN model (broken line) is shown, with the gamma model providing a better fit in both cases. For example, in the mouse-rat comparison, the chi-square fit of the gamma model is 21.55 (6 d.f.) whereas the chi-square fit for the SN model is 41.36. Note that, for computing the chi-square values, we considered only the observed frequency distribution of CSAMs of size 1–20, as the expected frequencies of larger CSAMs under the SN and gamma models were very small. Even in this range, we needed to pool consecutive CSAM sizes to ensure an expected frequency of at least five. Furthermore, we smoothed the observed segment size distribution in the case of the mouse-human comparison (by taking a moving average with a window size of three), due to the extreme L shape of the data, and truncated the long tail (which often contains segments with many functionally linked genes).

TABLE 3

Estimates of the number of conserved segments (CSAMs) from different releases of the data, with differing number of markers

Release date	Usable markers	Number of CSAMs: Total (observed + estimated)
	Mouse-human	
July 6, 1999	2239	529 (310 + 219)
July 13, 2000	2566	528 (331 + 197)
	Mouse-rat	
July 8, 1999	460	93 (56 + 37)
July 14, 2000	621	139 (78 + 61)

In each pair, the first species was used as the primary and the second as the secondary genome.

inversions, as evident from multiple adjacent conserved segments with alternating colors. It is, of course, possible that many such short segments are simply artifacts of map ambiguity. This would inflate the overall estimate of the number of conserved segments. However, our estimate is clearly not affected by such ambiguities in the relative order of markers in the human genome because the CSAM approach requires the use of marker order in one of the genomes only (mouse in the present case). In the mouse genome, whenever multiple genes were mapped to the same position, we ordered the markers so as to minimize the number of CSAMs.

The size distribution of the observed CSAMs between human and mouse genomes is given in Figure 6A. The fit of the gamma and SN models to the observed CSAM distribution shows that the gamma model fits the data better (χ^2 value = 4.74 for the gamma model, as opposed to 19.49 for the SN model; d.f. = 11). The CSAM analysis predicts 219 unobserved segments, with the total number of conserved segments between human and mouse genomes adding up to 529 ± 84. This result is not surprising considering that the relative genomic locations of <3% of all the genes for these two species were used (see Table 3, mouse-human comparison, July 1999 data release). However, it is almost three times the estimates based on conserved-synteny analysis (*e.g.*, SANKOFF and NADEAU 1996). Indeed, this almost threefold increase is consistently seen in the analysis of the July 2000 data, with >300 more usable genes (Table 3), as well as in the reciprocal analysis (primary, human; secondary, mouse; results not shown).

The autosomes of the mouse genome consist of a total of ~3000 Mbp and have a combined length of ~1500 cM (NUSBAUM *et al.* 1999; BLAKE *et al.* 2000). Therefore, the average length of a conserved segment is 2.84 ± 0.45 cM (5.67 ± 0.90 Mbp), which is <40% of the previous estimates of 8 cM or higher (for, *e.g.*, NADEAU and TAYLOR 1984; NADEAU and SANKOFF 1998). Interestingly, even the original NADEAU and TAYLOR (1984) approach using CSAM length data in centimorgan units (using CSAMs containing two or more markers) produced an estimate of 2.54 cM. This is somewhat surprising because the original method appears to be rarely used.

An approximate estimate of the rate of chromosomal rearrangement, averaged over evolutionary time, is obtained by assuming that each rearrangement produces 2 conserved segments (see Table 1). Table 1 shows the differential effects of the relative contributions of the

various rearrangement types to the number of conserved segments created by a rearrangement event. On average, a working figure of 2 segments/rearrangement appears conservative. For the mouse-human pair the estimate of the rate of chromosomal rearrangements is 1.15 ± 0.19 per million years, as these two species diverged ~110 million years ago (Kumar and Hedges 1998). This rate of rearrangement is approximate because the true relative contribution of different types of rearrangements that have led to the generation of the conserved segments remains unknown. Also, this is an average rate of divergence between human and mouse genomes; rates of evolution in the independent lineages leading to human and mouse may be unequal and differ substantially from each other.

We also used the CSAM approach to compare the mouse and rat (*Rattus norvegicus*) genomes (Table 3), and estimated 139 ± 25 conserved segments between their genomes (July 2000 data release). This translates to 0.74 ± 0.16 rearrangements per million years, assuming that the two lineages split 40 million years ago (Kumar and Hedges 1998). As mentioned above, the analysis of the mouse-human data yields a rate of 1.15 ± 0.19 rearrangements per million years. A reciprocal analysis of the mouse-rat data (*i.e.*, mouse as secondary and rat as primary) revealed a reduction in the number of useful markers to less than one-third of that in the previous analysis because the rat genome is not mapped as extensively as is the mouse genome. Figure 4 shows that the standard error of the estimate is large when the number of markers is low. This instability is reflected in the estimation of the number of CSAMs in the reciprocal analysis when compared to the corresponding estimates when mouse is used as the primary genome. This is in contrast to the mouse-human analysis, where the number of useful markers is large and similar for the reciprocal analysis (human used as the primary genome). Therefore, when the number of markers is reasonably large, the results obtained are quite robust (results not shown).

Our CSAM approach provides a simple way to compare extensively mapped model organism genomes with other genomes for which precise maps are not likely to become available in the near future. This facilitates quantification of the nature and tempo of macroevolutionary forces that have been instrumental in generating the current diversity of genomic organization in mammals (Burt *et al.* 1999; O'Brien *et al.* 1999).

We thank Tom Dowling and Mark Miller for discussions, Anne Beausang for some artwork, G. Valente for some database searches, and the Mouse Genome Database staff for their excellent user support. This research was supported by the National Institutes of Health, the National Science Foundation, and Burroughs Wellcome Fund grants to S.K.

LITERATURE CITED

Ayme, S., J. F. Mattei, M. G. Mattei, Y. Aurran and F. Giraud, 1976 Nonrandom distribution of chromosome breaks in cultured lymphocytes of normal subjects. Hum. Genet. **31:** 161–175.

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

Blake, J. A., J. T. Eppig, J. E. Richardson, M. T. Davisson and T. M. G. D. Group, 2000 The Mouse Genome Database (MGD): expanding genetic and genomic resources for the laboratory mouse. Nucleic Acids Res. **28:** 108–111.

Burt, D. W., C. Bruley, I. C. Dunn, C. T. Jones, A. Ramage *et al.*, 1999 The dynamics of chromosome evolution in birds and mammals. Nature **402:** 411–413.

Copeland, N. G., N. A. Jenkins, D. J. Gilbert, J. T. Eppig, L. J. Maltais *et al.*, 1993 A genetic linkage map of the mouse: current applications and future prospects. Science **262:** 57–66.

Efron, B., and R. J. Tibshirani, 1993 *An Introduction to the Bootstrap.* Chapman & Hall, New York.

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

Ewing, B., and P. Green, 2000 Analysis of expressed sequence tags indicates 35,000 human genes. Nat. Genet. **25:** 232–234.

Kumar, S., and S. B. Hedges, 1998 A molecular timescale for vertebrate evolution. Nature **392:** 917–920.

Liang, F., I. Holt, G. Pertea, S. Karamycheva, S. L. Salzberg *et al.*, 2000 Gene index analysis of the human genome estimates approximately 120,000 genes. Nat. Genet. **25:** 239–240.

Nadeau, J. H., and B. A. Taylor, 1984 Lengths of chromosomal segments conserved since divergence of man and mouse. Proc. Natl. Acad. Sci. USA **81:** 814–818.

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

Nusbaum, C., D. K. Slonim, K. L. Harris, B. W. Birren, R. G. Steen *et al.*, 1999 A YAC-based physical map of the mouse genome. Nat. Genet. **22:** 388–393.

O'Brien, S. J., M. Menotti-Raymond, W. J. Murphy, W. G. Nash, J. Wienberg *et al.*, 1999 The promise of comparative genomics in mammals. Science **286:** 458–462, 479–481.

Roest Crollius, H., O. Jaillon, A. Bernot, C. Dasilva, L. Bouneau *et al.*, 2000 Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. Nat. Genet. **25:** 235–238.

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

Schoen, D. J., 2000 Comparative genomics, marker density and statistical analysis of chromosome rearrangements. Genetics **154:** 943–952.

Schuler, G. D., M. S. Boguski, E. A. Stewart, L. D. Stein, G. Gyapay *et al.*, 1996 A gene map of the human genome. Science **274:** 540–546.

Scott, R., 1999 The future in understanding the molecular basis of life. The Institute for Genomics Research (TIGR) 11th International Genome Sequencing and Analysis Conference, Miami, FL (webcast at http://www.incyte.com/company/news/genes.shtml).

Waddington, D., A. J. Springbett and D. W. Burt, 2000 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., V. S. Nikiforov and E. V. Stepanyuk, 1995 Interval estimates of the combinatorial measures of similarity for orders of homologous genes. Genetika **31:** 1163–1167.

Communicating editor: H. Ochman