# Points of View

*Syst. Biol.* 43(2):250–261, 1994

## Surveys of Gene Families Using Polymerase Chain Reaction: PCR Selection and PCR Drift

ANDREAS WAGNER,[1] NEIL BLACKSTONE, PAULYN CARTWRIGHT, MATTHEW DICK, BERNHARD MISOF, PETER SNOW, GÜNTER P. WAGNER, JANET BARTELS, MIKE MURTHA, AND JOHN PENDLETON

*Department of Biology, Center for Computational Ecology, Yale University, New Haven, Connecticut 06511, USA*

Studies of multigene families present both challenges and opportunities for evolutionary biologists. Gene families represent a rich and diverse source of characters with the potential to be phylogenetically informative (e.g., sequence variation, variation within gene clusters, and number of gene clusters [Koop et al., 1989; Bartels et al., 1993; Cartwright et al., 1993; Pendleton et al., 1993]). Nevertheless, this complexity is sometimes difficult to interpret in a phylogenetic context (e.g., orthologous and paralogous variation and gene conversion [Fitch, 1970; Hillis and Dixon, 1991; Sanderson and Doyle, 1992]). Analysis of the sometimes complex characteristics of multigene families has recently been greatly facilitated by use of the polymerase chain reaction (PCR). Using degenerate primers corresponding to highly conserved regions of homologous genes, PCR can be used to detect and identify members of these gene families in samples of genomic DNA or cDNA. The utility of such an approach has recently been demonstrated using degenerate primers to amplify a region of Antennapedia-class homeoboxes (Murtha et al., 1991; Pendleton et al., 1993). This technique, however, cannot be fully exploited until certain methodological issues have been explored. Specifically, how do degenerate primers targeted for two or more regions actually sample the genome? This question has clear relevance for the interpretation of PCR results in phylogenetic studies, for instance, concerning the phylogenetic significance of PCR surveys of the number of members of a gene family in different taxa.

The polymerase chain reaction has been described in detail (e.g., see Erlich, 1989). This reaction proceeds by dissociation of template DNA at high temperature, annealing of primers at low temperature, and extension of synthesized DNA for a given number of cycles (often 30–40). The product DNA may be inserted into a bacterial plasmid vector by DNA ligation, cloned in a bacterial host, and sequenced (Sambrook et al., 1989). A number of factors may affect this process of producing inserts from genomic DNA. In the context of sampling gene families, we are most interested in those factors that produce skewness in the distribution of inserts, i.e., an apparent excess of inserts of some members of a gene family relative to others. We suggest two major classes of processes leading to such excess, PCR selection and PCR drift. PCR selection occurs when the reaction favors certain members of a gene family (e.g., in this case, separate PCR reactions produce

[1] E-mail: waganda@doliolum.biology.yale.edu.

a distribution of inserts skewed toward the same genes). A major contributor to PCR selection probably is differential primer affinity due to differences in primary or secondary structure of DNA at potential target sites. PCR drift is the result of random events occurring in the early cycles of the reaction. In this case, the bias will not be repeatable, i.e., separate PCR experiments in general do not produce biases towards the same member of the gene family. In the context of phylogenetic studies of gene families, we wish to initiate an exploration of PCR selection and drift by (1) providing a simple mathematical model and stochastic simulation of these processes, (2) using this model to suggest procedures that may mitigate PCR selection and drift, (3) discussing data from PCR surveys of the genomes of several metazoan taxa in this context, and (4) proposing a variety of experiments to clarify these interpretations further.

## Probabilistic Model and Simulations

To explore general properties of PCR regarding amplification from target DNA regions that are conserved throughout gene families, we developed a simple probabilistic model describing as a stochastic process the amplification of DNA molecules corresponding to different members of a gene family. We define *template* as a DNA region that is a target for amplification in a given cycle of the PCR, whether it be part of a genomic fragment of DNA or an amplification product of a previous cycle. Templates are subdivided into template species, each species corresponding to a member of a gene family. We confine ourselves mainly to a model in which only two different species of templates, template species 1 and 2, are involved in the reaction (corresponding to a two-gene family). The generalization to $n$ templates is less graphic and adds technical difficulties. Further, under the assumptions of the model, the essential information to understand PCR bias for $n$ templates is contained in the two-template case.

We denote the number of templates of species $i$ at cycle $t$ of the process as $N_i^t$. It is assumed that at time $t = 0$ (before the reaction is started) all templates are parts of fragments of genomic DNA and that all parts of the genome are equally represented in the reaction mix, implying $N_1^0 = N_2^0$ as an initial condition. The first cycle of the reaction replicates the templates off the genomic fragments only. In all subsequent cycles, however, the population of template molecules of species $i$ is subdivided into $N_i^t - N_i^0$ templates, which are products of earlier replication events, and $N_i^0$ templates, which reside on the genomic fragments. We assign a replication probability, $p_i$, to each template of species $i$, independent of time $t$. Time independence implies that reaction parameters do not change throughout the time span for which we model the process. This includes, for example, a constant fraction of primer molecules bound to templates, ample amounts of nucleotides for chain elongation, and constant specific activity of *Taq* polymerase, i.e., negligible heat degradation of the enzyme. These assumptions seem reasonable, because events responsible for the bias occur in the early cycles of the amplification process. We do, however, distinguish between replication probabilities of templates on genomic fragments ($p_i^{gen}$) and those of templates that are products (copies) of genomic templates ($p_i^c$). Product templates that have been replicated off genomic fragments in cycle $t - 1$ will also be assumed to belong to the set of product templates in cycle $t$, although they will be slightly longer than product templates generated in cycle $t$ and earlier. In general, we will examine cases in which $p_i^{gen}$ is lower than $p_i^c$, because primers will in some cases not match exactly (as opposed to product templates), and the long genomic fragments may have a tendency to form higher order structures or to be associated with proteins, and thus they may interfere with primer encounter or attachment. We assume that each molecule replicates (stochastically) independently of each other molecule, i.e., that there are neither interactions between templates of the same species nor interactions of templates across species in the replication process. Also, interactions between primer molecules of the same or of different DNA sequence that can introduce such correlations are excluded by this assumption. The change in $N_i^t$, $\Delta N_i^t$, in cycle $t$ can then be described as the sum of two binomially distributed random variables, one for genomic templates and one for product templates. Defining

$$p_i^t = \frac{N_i^0 p_i^{gen} + (N_i^t - N_i^0)p_i^c}{N_i^t}, \qquad (1)$$

it can be seen that the probability of a change in $N_i$ of $\Delta N_i^t$ in cycle $t + 1$ can be described by

$$P(\Delta N_i^t | N_i^t) = \binom{N_i^t}{\Delta N_i^t}(p_i^t)^{\Delta N_i^t}(1 - p_i^t)^{N_i^t - \Delta N_i^t}. \qquad (2)$$

In other words, the PCR process in cycle $t + 1$ is viewed as a series of $N_i^t$ independent trials that determine the replication or nonreplication of each of the $N_i^t$ template molecules with a probability of success (replication) equal to $p_i^t$. The random variable $\Delta N_i^t$ represents the number of molecules replicated. The total number of molecules after $t + 1$ cycles is given by the random variable

$$N_i^{t+1} = N_i^0 + \sum_{t'=1}^{t} \Delta N_i^{t'}. \qquad (3)$$

The probability distribution of $N_i^t$ is completely determined by the parameters $p_i^c$, $p_i^{gen}$, and $N_i^0$ and the time $t$. The observable quantity relevant to the experimenter is, however, the relative frequency,

$$x(t) = \frac{N_1^t}{N_1^t + N_2^t}, \qquad (4)$$

because it corresponds to the relative frequency at which template 1 will occur in the pool comprising the two species after $t$ cycles.

Diffusion equations and the theory of branching processes (Karlin and Taylor, 1975) provide means to analyze the stochastic process defined by the sequence $\{N_i^t\}_{t=0,\ldots,T}$. Analytical solutions for the probability distributions of $x(t)$ can, however, have fairly complicated properties. Because our general conclusions given below do not depend on the analytical form of any particular solution, we restrict ourselves to presenting the results of Monte Carlo simulations of the process for different values of the above parameters together with some elementary probabilistic arguments. In this way, we facilitate an intuitive comprehension of the occurring phenomena without having to present an overwhelming amount of mathematical formalism.

We start out with $N_i^0$ molecules of species $i$, generating random numbers that are distributed as specified in Equation 2, adding them up according to Equation 3, and calculating $x(t)$ at each time step. This procedure is carried out several thousand times to obtain an estimate of the ensemble probability distribution, i.e., the probability density function $f(x)$ of $x$ for each $t$ (defined in [0, 1]) after infinitely many realizations of a PCR process with identical initial conditions. For reasons of computational convenience, we approximate the binomially distributed random variable in Equation 2 by a normally distributed random variable $\xi(t)$ with mean $N_i^t p_i^t$ and variance $N_i^t p_i^t (1 - p_i^t)$ as soon as $N_i^t$ exceeds 30 template molecules. Although this approximation makes it possible that $N_i^{t+1} = N_i^t + \xi(t)$ exceeds $2N_i^t$ or is smaller than $N_i^t$·(in which case $N_i^{t+1}$ is set to $2N_i^t$ and $N_i^t$, respectively, in the simulations), the parameters used in this study assure that these events occur only outside a radius of more than three standard deviations from the mean of $\xi$. Moreover, they occur at symmetrical rates for species 1 and 2 and will therefore not bias the result.

At this point, the model does not explicitly account for differential cloning efficiency of PCR products into plasmids. Such differences can contribute to the kind of selection processes among different template species that are discussed below. However, because of the large number of insert molecules involved, stochastic effects should be negligible.

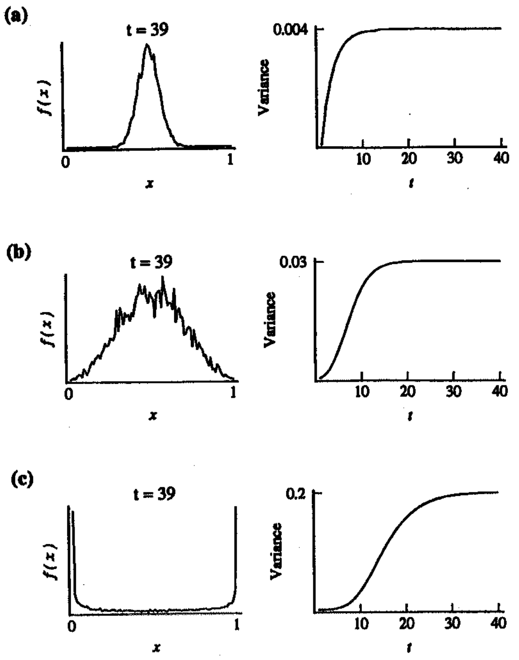

FIGURE 1. Comparison of amount of PCR drift generated by different replication probabilities of genomic templates, $p_i^{gen}$, for $N_0 = 10$. Left graphs show histograms of the estimated ensemble distribution for the relative frequency $x$ of template species 1 after 39 cycles of the PCR. Histograms divide the interval [0, 1] into 100 subintervals. Right graphs show the variance of the estimated ensemble distribution as a function of time ($t$) (cycle number). Ensemble sizes for Monte Carlo simulations are 5,000. Initial conditions: (a) $N_1^0 = N_2^0 = 10$, $p_1^{gen} = p_2^{gen} = 0.5$, $p_1^c = p_2^c = 0.5$; (b) $N_1^0 = N_2^0 = 10$, $p_1^{gen} = p_2^{gen} = 0.1$, $p_1^c = p_2^c = 0.5$; (c) $N_1^0 = N_2^0 = 10$, $p_1^{gen} = p_2^{gen} = 0.01$, $p_1^c = p_2^c = 0.5$.

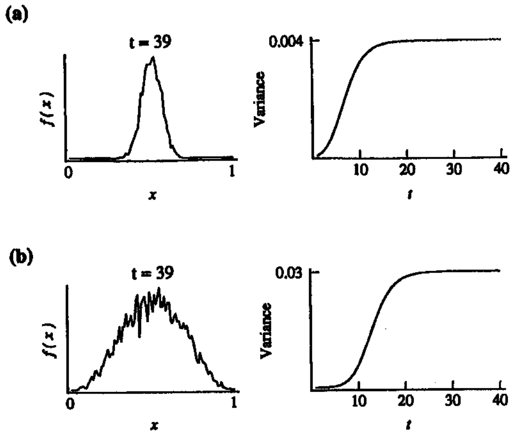

FIGURE 2. Comparison of amount of PCR drift generated by different replication probabilities of genomic templates, $p_i^{gen}$, for $N_0 = 100$. Left graphs show histograms of the estimated ensemble $f(x)$ distribution for the relative frequency $x$ of template species 1 after 39 cycles of the PCR. Histograms divide the interval [0, 1] into 100 subintervals. Right graphs show the variance of the estimated ensemble distribution as a function of time ($t$) (cycle number). Ensemble sizes for Monte Carlo simulations are 5,000. Initial conditions: (a) $N_1^0 = N_2^0 = 100$, $p_1^{gen} = p_2^{gen} = 0.1$, $p_1^c = p_2^c = 0.5$; (b) $N_1^0 = N_2^0 = 100$, $p_1^{gen} = p_2^{gen} = 0.01$, $p_1^c = p_2^c = 0.5$.

### *The Case $p_1^t = p_2^t$: PCR Drift*

A comparison of Figures 1b and 2a shows that with equal probabilities $p_i^t$ for the two template species the outcome is substantially different depending on $N_i^0$. The mean of the distribution is, as expected, in both cases close to 1/2, but the variance is much higher, i.e., the distribution is broader, if $N_i^0$ is lower. From the plots of variance against time in Figures 1 and 2, it seems that a buildup of variation occurs as long as $N_i^t$ is low. The shape of the distribution freezes as $N_i^t$ gets very large, conserving the variance that has been built up in early cycles. Starting at high $N_i^0$ means that the increase in $N_i$ will occur earlier and the distribution will be narrower. The coefficient of variation of the distribution of $\Delta N_i^t$ shows why this has to hold:

$$\frac{\sigma_{\Delta N_i^t}}{\mu_{\Delta N_i^t}} = \sqrt{\frac{1 - p_i^t}{N_i^t p_i^t}},$$

where $\sigma_{\Delta N_i^t}$ and $\mu_{\Delta N_i^t}$ are the standard deviation and the mean of $\Delta N_i^t$, respectively. The coefficient of variations scales as $\sqrt{N^{-1}}$. Thus, as $N_i^t$ gets larger, $\Delta N_i^t$ will be closer to the mean, thereby reducing the amount of variation generated for each of the two species. Because $N_i^t$ grows exponentially, the mean growth rate (averaged over the ensemble) of $N_i^t$ determines the dynamics of the process. A high growth rate implies earlier stabilization of the distribution.

From a comparison of Figures 1a, 1b, and 1c and Figures 2a and 2b, it can be seen that the ratio of $p_i^c$ to $p_i^{gen}$ is also crucial in this scenario. A very low $p_i^{gen}$ causes slow initial growth of $N_i^t$, implying that $p_i^t$ grows very slowly, too. It is not until a considerable number of new templates have been created that $p_i^t$ becomes large and the growth rates of $N_i^t$ increase, leading to faster growth of $N_i^t$ and stabilization of the distribution. The lower genomic replication probabilities keep the number of templates low and increase the period of time during which buildup of variance is possible. In Figure 1c, this period extends almost to the end of the monitored number of cycles. A large fraction of the ensemble drifts freely, getting concentrated at the boundaries before $N_i$ is sufficiently large to prevent further accumulation of variability.

In summary, the effects of stochastic buildup of variability (i.e., the effects of PCR drift) increase for decreasing $N_i^0$ and decreasing ratio $p_i^{gen}/p_i^c$. What are the implications of these results for an actual experiment? Although for technical reasons we used values for $N_i^0$ that may be orders of magnitude lower than those used in many of the actual applications of the technique, our results were not affected in a qualitative way by this choice of initial conditions. Furthermore, we are not aware of any estimates for $p_i^{gen}$ in an empirical model system; therefore actual values for $p_i^{gen}$ may be much lower than the ones considered here, thus compensating for higher $N_i^0$.

For Figure 1b, assume that the parameters for this simulation occur in an actual experimental setup. Each PCR experiment that is being carried out in this same setup represents a sample of size 1 from the distribution in Figure 1b. Picking a sample close to 1/2 (or close to $1/n$ in the corresponding $n$-template case) is the ideal scenario for the experimenter; the products have equal concentrations (molarities) after the PCR. It is, however, likely that the sample picked will be displaced from the mean and therefore that species 1 will have a molarity different from that of species 2 after the reaction. In the $n$-template case, this difference causes the phenomena observed in the data discussed below: the distributions are skewed, with some templates occurring very frequently and others occurring at very low frequency. The sampling distribution of template sequences amplified by the PCR will be a multinomial distribution with unequal probabilities assigned to templates of different species. This result is undesirable, because one will have to analyze larger numbers of PCR products in the skewed case than in the case with uniform probabilities to find the rare PCR products. How is it possible to improve the situation? Simple considerations for the two-template case show that carrying out several independent PCRs at identical initial conditions and pooling of the products will improve the properties of the sampling distribution. Formally, carrying out $k$ reactions and pooling the products corresponds to forming an average $Z_k = (X_1 + \ldots + X_k)/k$ of independent, identically distributed random variables $X_i$. Using the central limit theorem (Feller, 1968), it is a trivial result that $Z_k$ will be asymptotically normally distributed with mean 1/2 and variance $\sigma_x^2/k$ where $\sigma_x^2$ is the variance of $x$. The variance of this distribution scales as $1/k$, implying that the more reactions one pools, the more likely one is to get closer to the desired mean. For small $k$, where this approximation may not be applied, one can use Chebychev's inequality (Feller, 1968) to obtain

$$P\left(\left|Z_k - \frac{1}{2}\right| > \epsilon\right) < \frac{\sigma_x^2}{\epsilon^2 k},$$

where $\epsilon$ is some arbitrarily small positive real number, which shows that qualitatively the same result holds: given an arbitrarily small but constant $\epsilon$, the probability that $Z_k$ deviates from its expected value 1/2 by more than $\epsilon$ is smaller than $S^2/(\epsilon^2 k)$. This probability scales as the inverse of $k$ (the number of reactions pooled), implying that an increase in the number of reactions will lead to a decrease in this probability. In other words, the more reactions are carried out, the more likely it is that the amount of bias generated by PCR drift does not exceed a given threshold $\epsilon$.

Considering a case similar to that in Figure 1c with highly skewed distributions, it is possible to make simple generalizations for the case of $n$ template species. We show that the expected degree of improvement of the composition of one's reaction mix increases with the number of samples drawn out of one ensemble distribution and increases with the dimensionality of the problem. The $n$ template species define $(n - 1)$ frequencies $x_1, \ldots, x_{n-1}$ in an appropriate ensemble distribution. The appropriate domain of their density is an $(n - 1)$-dimensional simplex with $n$ vertices. We approximate the highly skewed distribution by the discrete limiting case, such that taking a sample out of the $n$-dimensional ensemble distribution on the simplex means picking a vertex of the simplex with probability $1/n$. This procedure corresponds to carrying out a PCR experiment with $n$ templates and obtaining one template species at a high concentration and the others at low concentration. Let the random variable $Y_k$, $k < n$, denote the number of samples of size 1 out of the ensemble distribution (i.e., the number of independent PCRs) that is necessary to obtain a new template species at a high concentration after having already obtained $k$ different template species, i.e., $k$ template species are available at appreciable and almost uniform frequency, whereas $n - k$ template species are rare. Because the probability that one has to wait for $i$ new PCRs until one obtains a new template species at high frequency is given by

$$P(Y_k = i) = \frac{n-k}{n}\left(\frac{k}{n}\right)^i;$$

it is obvious that $Y_k$ has a geometrical distribution (Feller, 1968). This yields for the expectation $E$ of $Y_k$

$$E(Y_k) = \frac{k}{n-k}.$$

With $k$ constant and $n$ increasing, the expected number of PCRs necessary for obtaining a new template species approaches zero. In terms of the experiment, this simple result implies that it is even more advantageous (in the case of highly skewed distributions) to pool individual reactions in higher dimensional cases than in lower dimensions, because pooling will more likely allow finding of a new template sequence at appreciable rates in a new PCR.

### *The Case $p_1^t \neq p_2^t$: PCR Selection*

Several factors may cause asymmetries in the replication rates across template species. Examples include different melting temperatures of different primers in the reaction mix, causing different binding constants of different primers at the elongation temperature, and secondary structure formation of templates in the annealing stage of each cycle, thereby causing steric hindrance for primer binding. Figure 3 shows for the example of $N_1^0 = N_2^0 = 10$ that even a moderate difference of 0.1 in the replication probabilities across template species can cause a substantial shift in the ensemble density, even after a small number of cycles. After 39 cycles, the distribution is shifted with a mean close to $x = 1$, reflecting the faster (mean) exponential growth of the species with higher replication probability, the species that is favored by PCR selection. Not unexpectedly, the shift of the mean occurs gradually and is decoupled from the buildup of variance in the distribution (which ceases very early in the process). Accompanying the shift of the mean, a rapid decrease in the variance of $x(t)$ occurs, as can be seen from the lower right graph in Figure 3 and the narrowing of the density in consecutive cycles. In an experimental context, this suggests that if there may be unequal replication probabilities across template species (e.g., by using primer binding sites with very different AT/GC ratios) it may not be advisable to carry out the PCR for a large number of cycles, unless very large amounts of PCR product are necessary for post-PCR processing. Instead, it may be advantageous to stop the reaction as early as possible to minimize loss of variability of PCR products or to use low annealing temperatures for early cycles, followed by a high annealing temperature for later cycles to increase the amount of products obtained. A complementary but costly strategy may in-

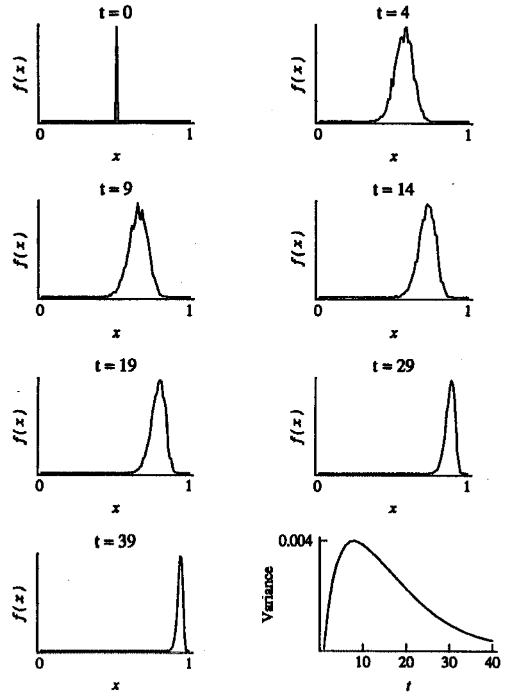

FIGURE 3. Effects of PCR selection caused by different replication probabilities of different template species. Lower right: Variance of estimated ensemble distribution for the relative frequency $x$ of template species 1 as a function of time ($t$) (cycle number). Initial conditions are $N_1^0 = N_2^0 = 10$, $p_1^{gen} = p_1^c = 0.5$, $p_2^{gen} = p_2^c = 0.4$. All other graphs: Histograms of estimated ensemble distribution for the relative frequency $x$ of template species 1 at different times (cycles) under the same initial conditions. Ensemble size for Monte Carlo simulations was 5,000. Histograms divide the interval [0, 1] into 100 subintervals.

volve carrying out different reactions with different mixes of primers for each reaction and pooling products to compensate for inequalities in replication rates across reactions. A further conceivable remedy might be starting reactions with a very small number of templates so that the strong stochastic buildup of variability (as in Fig. 1b) can override the effects of PCR selection.

## SURVEYS OF ANTENNAPEDIA-CLASS HOMEODOMAINS IN METAZOANS

The homeobox genes encode a family of DNA-binding regulatory proteins characterized by the helix-turn-helix conformation of the highly conserved homeodomain. Antennapedia (Antp)-class homeobox genes are of particular interest because of their conservation in chromosomal organization (Duboule and Dollé, 1989; Graham et al., 1989). Using primers targeted for Antp-class homeobox sequences, surveys of these homeobox genes have now been carried out on a number of metazoan taxa. Usually, 500 ng of genomic DNA was used as a template; depending on the genome size, this corresponds to roughly $10^3$–$10^4$ copies of the genome. After 35–40 PCR cycles, the resulting product from a variable number of PCR reactions (one to six) was cloned and a variable number of inserts were sequenced. These data can provide an empirical perspective on the general considerations discussed above.

These data are presented as frequency distributions of the number of members of a gene family found in frequency classes (Figs. 4–7); thus these classes correspond to the number of times that a gene family member was found as an insert. The frequency of the zero class (i.e., the number of gene family members that were not found as inserts) cannot be determined from a single PCR experiment. Repeated PCR experiments (or alternative lines of evidence) can contribute to estimates of this zero class. If the zero class is known, the distribution can be compared with known distributions, e.g., a Poisson distribution in which the mean equals the variance. A Poisson distribution would suggest that the PCR process is randomly scattering inserts on the members of this gene family, i.e., the probability for obtaining any one member out of the family is equal for all members of the family. In two cases, the zero class is known with confidence and sample statistics can be estimated; for Figure 4b, the mean and the variance of the distribution are nearly equal ($\bar{x} = 3.26$, variance = 4.66), whereas for Figure 5c the variance greatly exceeds the mean ($\bar{x} = 7.2$, variance = 77.4). Although more definitive calculations were not done because of small sample sizes, it seems that the distribution in Figure 4b is similar to a Poisson distribution and that of Figure 5c is not. Further,

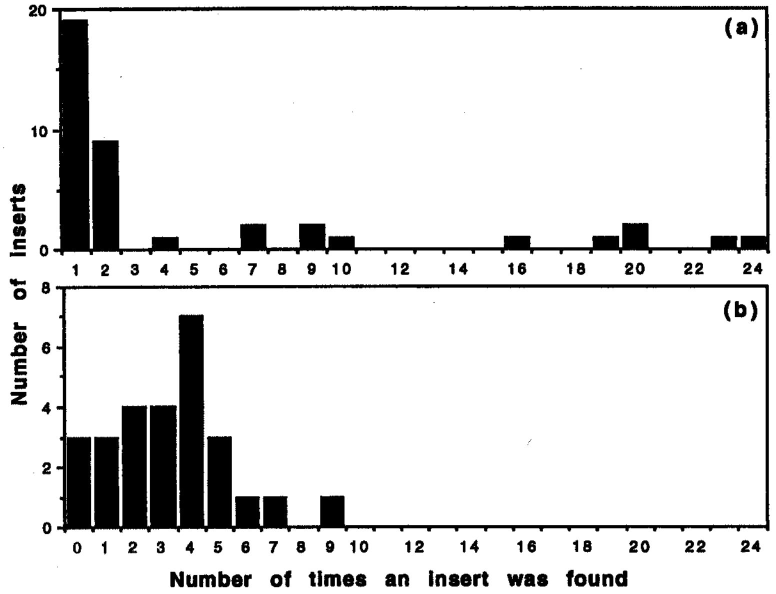

FIGURE 4. Frequency distributions of the number of members of a gene family (Antennapedia-class homeobox genes) found in frequency classes. These classes correspond to the number of times that a gene family member was found as an insert, e.g., in (a), 19 genes were found as an insert once and 1 gene was found 24 times. Distributions represent independent PCR experiments on different vertebrate taxa. (a) Product is from a single 40-cycle PCR experiment using the same primers as in Figure 5. (b) Product is from a single 35-cycle PCR experiment (Murtha et al., 1991) using the described primers. The zero class estimate was provided by other experiments.

qualitative examination of the remaining distributions suggests that the distribution in Figure 4b is exceptional. These data thus suggest that with the experimental conditions used, there is a tendency for some sequences to occur as inserts much more frequently than others and to a greater extent than expected under the null hypothesis of a Poisson distribution. Either PCR selection or PCR drift seems to be operating.

Distinguishing between PCR selection and drift hinges on the issue of repeatability. In one case, two PCR experiments were done on conspecifics (Table 1). There is a low and nonsignificant correlation between the frequencies at which the gene family members occur in the two experiments. Nevertheless, there may be indications of PCR selection, e.g., the same sequence is the second most common in both experiments, and the most common sequence in one experiment is the fourth most common in the other. Repeatability can also be judged from PCR experiments on closely related taxa, if the frequencies of cognate sequences are compared. Such a comparison (Table 1) shows a nonsignificant overall correlation, but the same cognate sequence is the most common in both experiments.

Indirect evidence for repeatability can be gathered by examining the sequences of the primers that bind to common and rare gene family members, e.g., do these suites of primers differ in their GC/AT ratios? Such data were assessed for a PCR experiment that produced one of the more skewed distributions (Fig. 4a). For the gene family member found at the highest frequency, the exact sequences of a sample of 10 3F primers (5′ end) were determined. For gene family members found only a single time, the exact sequences of a sample of six 3F primers were determined. The total GC/AT ratio in the variable region of the high frequency members (21 base pairs per primer) was 103/107 (0.96); for the same

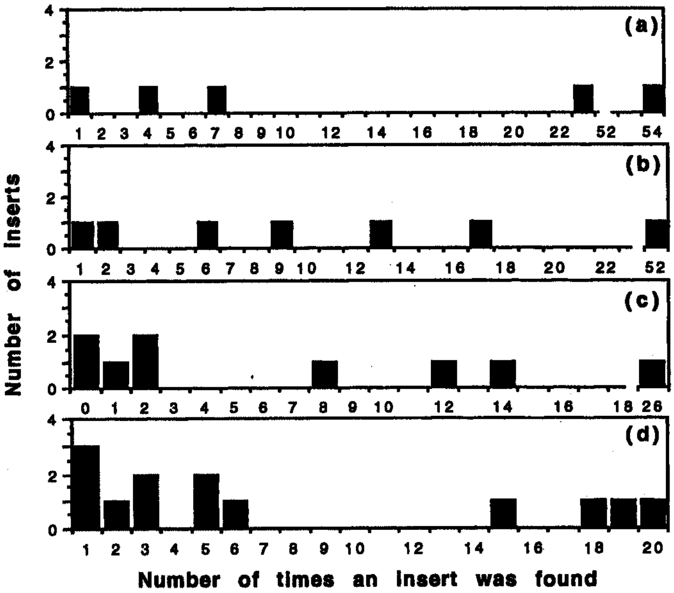

FIGURE 5. Frequency distributions of the number of members of a gene family (Antennapedia-class homeobox genes) found in frequency classes for independent PCR experiments on different metazoan taxa using the same primers. Each distribution is from product combined from one to four 40-cycle PCR experiments. In (c), the zero class estimate was provided by independent experiments. In cases where they occurred, single base-pair differences were attributed to PCR error and inserts were grouped accordingly (cf. Pendleton et al., 1993).

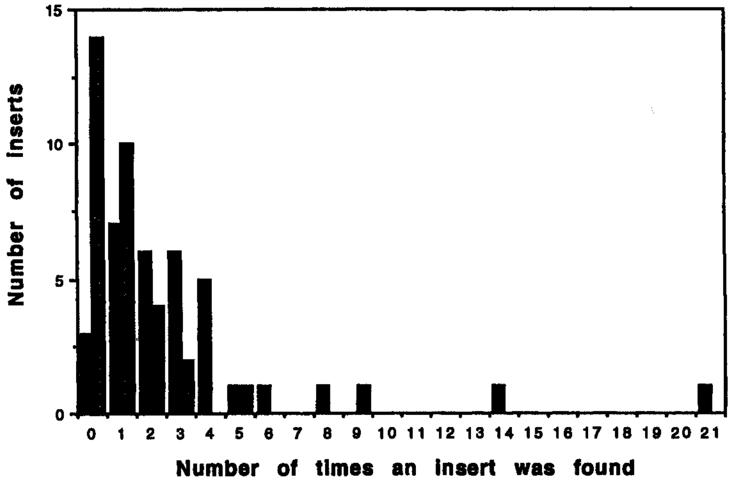

FIGURE 6. Frequency distributions of the number of members of a gene family (Antennapedia-class homeobox genes) found in frequency classes for two independent PCR experiments (hatched and solid bars) on the same (arthropod) taxon using the same primers. Estimates of the zero class are provided by the number of genes found in one PCR experiment but not the other. In each case, product is from a single PCR reaction for 40 cycles (hatched bars) and 80 cycles (solid bars).

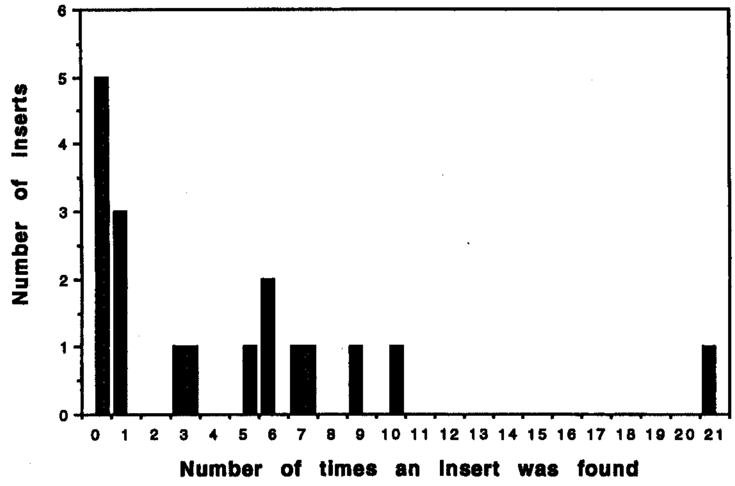

FIGURE 7. Frequency distributions of the number of members of a gene family (Antennapedia-class homeobox genes) found in frequency classes for two independent PCR experiments on the same (annelid) taxon using different primers. Counts shown by hatched bars were produced using primers of Murtha et al. (1991; see Fig. 4b) in a single 35-cycle PCR experiment; counts shown by solid bars were produced using the same primers as used in other experiments shown (Figs. 5, 6) and mixing product from six 35-cycle PCR experiments. The zero class estimate is provided by the number of genes found in one PCR experiment but not the other.

region of the members found only once, this ratio was 54/72 (0.75). Similarly, for the 5E primers (3′ end), a sample of eight primers that amplified the highest frequency sequence had a GC/AT ratio in an 18-base-pair variable region of 66/78 (0.85), whereas a sample of six primers that amplified sequences found only a single time had a ratio of 43/65 (0.66) for the same region. PCR selection based on the thermal stability of primer–genomic template duplexes may be indicated.

## FUTURE DIRECTIONS

The two proposed mechanisms for causing skewness of distributions of PCR products for different members of a gene family, PCR drift and PCR selection, are derived from a simple stochastic model of the PCR process. Unfortunately, the model is structurally unstable in the sense that small deviations of the condition $p_i^t = p_j^t$ for any $i$ and $j$ will produce substantially different concentrations of PCR products. Empirical data, although not definitive, suggest that both mechanisms may operate concurrently. Because of implications of the model, we propose a series of remedies that may be combined if there is uncertainty as to which of the mechanisms is important in an actual experimental system. First, in cases where only PCR drift is presumed to be important, carrying out several independent reactions and pooling the products should reduce the skewness of the distribution. Second, in cases where PCR selection is suspected, e.g., if there is wide variation in the GC/AT ratio of the set of degenerate primers used or of the target region of the gene family, or both, one may carry out the reaction only for the smallest necessary number of cycles or, alternatively, start the reaction with a small amount of DNA (small number of genomes) so as to override the effects of selection by the strong stochastic forces occurring in the first few cycles of the reaction.

These recommendations are supported by the data currently available. However, both additional data and further experimentation are clearly needed. For instance, additional data on the GC/AT ratios for the primers of gene family members obtained at high frequency versus low frequency

TABLE 1. Comparisons used for assessing biases toward particular gene family members in PCR surveys of the same arthropod taxon (comparison 1; two independent PCR experiments on conspecifics) and two different annelid taxa (comparison 2; cognate genes are paired). Single base-pair differences were attributed to PCR error, and sequences were grouped accordingly (cf. Pendleton et al., 1993).

| Gene | Comparison 1[a] | | Comparison 2[b] | |
|---|---|---|---|---|
| sequence | Animal 1 | Animal 2 | Taxon 1 | Taxon 2 |
| 1 | 21 | 3 | 20 | 21 |
| 2 | 14 | 5 | 18 | 6 |
| 3 | 8 | 0 | 15 | 6 |
| 4 | 6 | 9 | 6 | 1 |
| 5 | 5 | 1 | 5 | 7 |
| 6 | 4 | 1 | 5 | 7 |
| 7 | 4 | 1 | 3 | 3 |
| 8 | 4 | 1 | 3 | 2 |
| 9 | 4 | 1 | 1 | 7 |
| 10 | 4 | 0 | 1 | 1 |
| 11 | 3 | 2 | 1 | 3 |
| 12 | 3 | 2 | | |
| 13 | 3 | 0 | | |
| 14 | 3 | 0 | | |
| 15 | 3 | 0 | | |
| 16 | 3 | 0 | | |
| 17 | 2 | 3 | | |
| 18 | 2 | 1 | | |
| 19 | 2 | 1 | | |
| 20 | 2 | 0 | | |
| 21 | 2 | 0 | | |
| 22 | 2 | 0 | | |
| 23 | 1 | 2 | | |
| 24 | 1 | 1 | | |
| 25 | 1 | 0 | | |
| 26 | 1 | 0 | | |
| 27 | 1 | 0 | | |
| 28 | 1 | 0 | | |
| 29 | 1 | 0 | | |
| 30 | 0 | 2 | | |
| 31 | 0 | 1 | | |
| 32 | 0 | 1 | | |

[a] Nonsignificant correlation (Spearman's $R_S = 0.23$, $P > 0.20$; Kendall's $\tau = 0.18$, $P > 0.20$).

[b] Nonsignificant correlation ($R_S = 0.39$, $P > 0.20$; Kendall's $\tau = 0.30$, $P > 0.20$).

should be collected. The critical question of whether skewed GC/AT primer ratios result from PCR selection or drift should be assessed by sequencing the flanking regions of the genomic templates, e.g., by using RACE (rapid amplification of cDNA ends [Frohman et al., 1988]). The repeatability of biases toward certain gene family members also requires more investigation, perhaps by carrying out separate PCR experiments on conspecifics (as reported above) or on a cosmid or YAC (yeast artificial chromosomes [Burke et al., 1987]) library containing several gene family members on a single clone. The relationship between $p_i^{gen}$ and $p_i^c$ could be tested by carrying out a PCR experiment on a mixture of genomic DNA (containing a known number of targets) with a known number of PCR products of the same target. One could perhaps distinguish the resulting PCR products with a base-pair change outside the primer binding region of the introduced product. PCR selection would lead to differential amplification of the introduced target with respect to other genomic targets, which could be verified by restriction site analysis or sequence analysis of the reaction products.

## ACKNOWLEDGMENTS

Two anonymous reviewers provided useful comments and criticisms. Support was provided by the National Science Foundation (BSR-88-05691, IBN-91-19437). This paper is communication no. 9 from the Center for Computational Ecology of the Yale Institute for Biospherics Studies.

## REFERENCES

BARTELS, J. L., M. T. MURTHA, AND F. H. RUDDLE. 1993. Multiple *Hox/HOM*-class homeoboxes in Platyhelminthes. Mol. Phylogenet. Evol. 2:143–151.

BURKE, D. T., C. F. CARLE, AND M. V. OLSON. 1987. Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. Science 236:806–812.

CARTWRIGHT, P., M. DICK, AND L. W. BUSS. 1993. *HOM/Hox* type homeoboxes in the chelicerate *Limulus polyphemus.* Mol. Phylogenet. Evol. 2:185–192.

DUBOULE, D., AND P. DOLLÉ. 1989. The structural and functional organization of the murine HOX gene family resembles that of *Drosophila* homeotic genes. EMBO J. 8:1497–1505.

ERLICH, H. A. (ed.). 1989. PCR technology. Stockton Press, New York.

FELLER, W. 1968. An introduction to probability theory and its applications. Wiley, New York.

FITCH, W. 1970. Distinguishing homologous from analogous proteins. Syst. Zool. 19:99–113.

FROHMAN, M. A., M. K. DUSH, AND G. R. MARTIN. 1988. Rapid production of full-length cDNAs from rare transcripts: Amplification using a single gene-specific oligonucleotide primer. Proc. Natl. Acad. Sci. USA 85:8998–9002.

GRAHAM, A., N. PAPALOPULU, AND R. KRUMLAUF. 1989. The murine and *Drosophila* homeobox gene complexes have common features of organization and expression. Cell 57:367–378.

HILLIS, D. M., AND M. T. DIXON. 1991. Ribosomal

DNA: Molecular evolution and phylogenetic inference. Q. Rev. Biol. 66:411–453.

KARLIN, S., AND H. M. TAYLOR. 1975. A first course in stochastic processes. Academic Press, New York.

KOOP, B. F., D. L. TAGLE, M. GOODMAN, AND J. L. SLIGHTOM. 1989. A molecular view of primate phylogeny and important systematic and evolutionary questions. Mol. Biol. Evol. 6:580–612.

MURTHA, M. T., J. F. LECKMAN, AND F. R. RUDDLE. 1991. Detection of homeobox genes in development and evolution. Proc. Natl. Acad. Sci. USA 88: 10711–10715.

PENDLETON, J. T., B. K. NAGAI, M. T. MURTHA, AND F. R. RUDDLE. 1993. Expansion of the *Hox* gene family and the evolution of chordates. Proc. Natl. Acad. Sci. USA 90:6300–6301.

SAMBROOK, J., E. F. FRITSCH, AND T. MANIATIS. 1989. Molecular cloning. A laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

SANDERSON, M. J., AND J. J. DOYLE. 1992. Reconstruction of organismal and gene phylogenies from data on multigene families: Concerted evolution, homoplasy, and confidence. Syst. Biol. 41:4–17.

*Received 18 December 1992; accepted 13 October 1993*