Analysis of DNA Diversity by Spatial Autocorrelation

Giorgio Bertorelle* and Guido Barbujani†

**Dipartimento di Biologia, Università di Padova, 35121 Padova, Italy and †Dipartimento di Scienze Statistiche, Università di Bologna, 40126 Bologna, Italy*

Manuscript received November 21, 1994
Accepted for publication February 25, 1995

ABSTRACT

Two statistics are proposed for summarizing spatial patterns of DNA diversity. These autocorrelation indices for DNA analysis, or AIDAs, can be applied to RFLP and sequence data; the resulting set of autocorrelation coefficients, or *correlogram,* measures whether, and to what extent, individual DNA sequences or haplotypes resemble the haplotypes sampled at arbitrarily chosen spatial distances. Analyses of computer-generated sets of data, and of RFLP data from two natural populations, show that AIDAs allow one to objectively and simply identify basic patterns in the spatial distribution of haplotypes. These statistics, therefore, seem to be a useful tool both to explore the genetic structure of a population and to suggest hypotheses on the evolutionary processes that shaped the observed patterns.

POPULATIONS are said to show genetic structure whenever the distributions of their genes do not conform to panmictic expectations. Evolutionary processes may be inferred from data on the genetic structure of populations. Ideally, such inferences require two phases. Patterns of genetic variation are initially summarized by descriptive statistics, and then they are interpreted by comparison with the patterns predicted by theoretical models. The second phase includes the estimation of parameters quantifying the effects of mutation, selection, inbreeding, drift, or gene flow.

Well-established methods allow the analysis of population structure at the level of classical protein markers (see *e.g.*, WIJSMAN and CAVALLI-SFORZA 1984). Although a regular, clock-like rate of allele frequency divergence is sometimes assumed (CAVALLI-SFORZA *et al.* 1988; BARRANTES *et al.* 1990), the emphasis in most such studies has been on factors such as adaptation to variable habitats, patterns of gene flow, and isolation by distance. In other words, the genetic population structure is often analyzed with an eye to its geographical, rather than historical, determinants.

In contrast, for DNA data, the analysis generally leads to reconstruction of genealogies of alleles (reviewed in HUDSON 1990). In this way, under neutrality assumptions, the focus shifts to processes occurring in the course of the populations' history, typically branching events. There are, however, problems. As FELSENSTEIN (1982) first remarked, history and geography jointly determine patterns and levels of genetic diversity; in the absence of nongenetic information, their effects are difficult to disentangle. Equal levels of haplotype diversity, for example, may be caused by a stable regime of gene flow constrained by geography or by episodes punctuating the population's history, such as the colonization of a new habitat (see *e.g.*, TEMPLETON 1993).

Several attempts are being made to tackle this problem, by incorporating a geographic perspective into methods for the study of DNA variation. SLATKIN (1989, 1991) and SLATKIN and MADDISON (1989, 1990) worked out how to estimate the minimum numbers of migration events from gene genealogies, looked for empirical relationships linking such parameters with gene flow rates (m), inbreeding coefficients and effective population sizes (N_e), and extended their approach to nonequilibrium situations (SLATKIN 1993). EXCOFFIER *et al.* (1992) developed a method for the analysis of molecular variance, AMOVA. They proposed new statistics resembling WRIGHT's (1965) F, whose values depend on the correlations between haplotypes at various levels of subdivision, *i.e.*, within demes, between demes within regions, among regions, *etc.* TEMPLETON (1993) proposed to analyze evolutionary trees by a cladistic procedure (TEMPLETON *et al.* 1987) designed to identify, respectively, areas of population expansion, and areas where gene flow has been restricted.

In this paper, we present two indices of spatial autocorrelation, specifically developed for treatment of molecular data. In classical spatial autocorrelation analysis (SOKAL and ODEN 1978a,b), levels of genetic resemblance are calculated between pairs of localities, within arbitrary distance classes. Spatial patterns of allele frequencies are thus represented by sets of coefficients (MORAN's I or GEARY's c) calculated at different distances, or *correlograms.*

Correlograms give an easy-to-understand depiction of spatial patterns of variation. They can be compared across different geographical regions and genetic markers, thus suggesting obvious neutrality tests (SOKAL *et al.* 1987, 1989; PIGLIUCCI and BARBUJANI 1991; SOKAL

Corresponding author: Guido Barbujani, Dipartimento di Scienze Statistiche, Università di Bologna, via Belle Arti 41, 40126 Bologna, Italy. E-mail: g4bbovl5@icineca.cineca.It

and JACQUEZ 1991). As they are, spatial autocorrelation methods can be applied to DNA data if the data are transformed in such a way as to be treated as allele frequencies (see COSTA *et al.* 1992). This transformation, however, is possible only if several sequencies are sampled in each population, and it implies a loss of information because the haplotypes are classified as being the same or different, and the amount of sequence differences between them is not considered.

Ideally, an autocorrelation index for DNA analysis (we propose to call it AIDA) should measure whether, and to what extent, individual DNA sequences (or haplotypes) resemble the sequences sampled at different localities. In this way, this index could be employed for exploratory data analysis (SLATKIN and ARTER 1991). Few sequences sampled in different localities can be used to objectively recognize nonrandom distributions of haplotypes. Correlograms inspection may also help in testing hypotheses on the evolutionary history of the populations under study.

The two autocorrelation indices, or AIDAs, described in this paper, are a product-moment coefficient, analogous to MORAN's *I*, and a distance-like coefficient, analogous to GEARY's *c* (SOKAL and ODEN 1978a). Both are based on comparisons between individual sequences, rather than between frequencies of alleles or haplotypes. Because in this way genetic relatedness is inferred from sequence similarity, only parts of the genome should be considered that either are haploid or can be treated as such (*e.g.*, DNA segments where recombination can safely be ruled out). Indeed, a single crossing-over may result in large sequence differences between haplotypes. Under recombination, therefore, sequence differences are no longer a measure of evolutionary distance, and this may bias the inferences drawn from autocorrelation analysis. However, we shall also show how some sets of data where recombination has occurred may be liable to spatial autocorrelation analysis. A matrix is then calculated, of phenetic (namely counts of sequence differences between haplotypes) or evolutionary distances (estimates of the number of evolutionary events, mutation and possibly recombination, separating pairs of haplotypes); from either matrix is the correlogram evaluated.

REPRESENTATION OF DATA

Suppose there are *S* segregating or polymorphic sites in a DNA fragment, studied either by DNA sequencing or by restriction site analysis. In general, we shall assume that recombination is absent, and that only two different nucleotides can occur at each polymorphic site. The latter is a widely used assumption (see *e.g.*, KREITMAN 1983; EXCOFFIER *et al.* 1992; EXCOFFIER and SMOUSE 1994), consistent with the infinite-site model (KIMURA 1969; WATTERSON 1975), but not necessarily implying it. Following EXCOFFIER *et al.* (1992), an individual sequence, or haplotype, will be represented as a binary vector

$$\mathbf{p} = [p_1, p_2, p_3, \ldots, p_s]. \qquad (1)$$

In the case of DNA sequence data, p_i represents either of the possible nucleotides at each of the *S* sites, arbitrarily labelled as 0 or 1. Alternatively, for RFLP data, p_i represents the presence or absence of one of *S* possible restriction sites.

A modification is necessary when more than two different nucleotides occur at the same site. 0s and 1s will then indicate presence or absence of one specific nucleotide at a certain site. For example, A, C, G, and T may be coded as 1000, 0100, 0010, and 0001, respectively. In this way, each segregating site will be represented by a four-digit binary code. For the sake of simplicity, however, we shall refer to the simpler case of two alternative nucleotides.

Binary digits can also be used to represent presence and absence of a repeat in a microsatellite; however, the autocorrelation analysis of these data is complicated by the fact that recombination is probably a major force generating VNTR diversity (see HARDING 1993).

Representing haplotypes by binary vectors corresponds to the approach labelled as *phenetic* by EXCOFFIER *et al.* (1992). An alternative, *evolutionary*, approach exists; it will lead to different results if homoplasy is high. For the evolutionary approach, a minimum-spanning tree (PRIM 1957; ROHLF 1970) is constructed; the 1s in the **p** vector will then represent the occurrence of a mutational event along the branches of the tree that separate the sequence identified as ancestral (coded by a string of 0s) and the sequence of interest. Errors in the definition of the ancestral haplotype do not modify the results of the analysis, because they do not affect the number of mutational steps separating any pairs of haplotypes. Here we note that recombination, deletion and duplication events may sometimes be placed in the tree with a certain degree of confidence (see *e.g.*, COSTA *et al.* 1991). In this case, and only in this case, can AIDAs be employed to summarize geographic diversity even for DNA regions affected by recombination. Of course, only for the analysis of haploid DNA is there a one-to-one correspondence between individuals and haplotypes. Under both the phenetic and the evolutionary approach, each haplotype in the study will be represented by the spatial coordinates of its place of origin or sampling, and by its **p** vector.

CALCULATION OF AIDAs

Schematically, six steps are necessary: (1) calculation of a matrix of spatial distances between individuals; (2) choice of arbitrary distance classes; (3) evaluation of an average genetic vector; (4) calculation of an autocorrelation coefficient in each class; (5) estimation of confidence limits by a randomization procedure; and (6) significance testing.

Steps 1 and 2 deal with geographical information. They are straightforward and do not differ from the initial steps of classical autocorrelation analysis. A matrix of pairwise spatial distances between individuals is calculated, and arbitrary distance class boundaries are defined. Popular criteria for defining class boundaries are equal intervals (*i.e.*, point pairs are allocated to classes of equal width) and equal frequencies (*i.e.*, class limits are defined in such a way that equal numbers of point pairs fall in each class) (WARTENBERG 1989). Alternatively, specific class boundaries may be chosen. For each of these classes, however defined, an AIDA will be computed.

In step 3 one turns to molecular data. For each of the S sites, an average is calculated across all individuals. Such averages will be >0 and <1, because the monomorphic sites are excluded. The average vector $\bar{\mathbf{p}}$ is a mathematical artifact that does not represent any biological reality. The values of its individual components are simply the frequencies of either of the alternative nucleotides at each site (under the phenetic criterion), or the frequency at which each specific evolutionary change (each mutation event or, when possible, each deletion, duplication or recombination) is observed in the sample (under the evolutionary criterion).

For each class defined in step 2, the following AIDAs, respectively called II and cc, by analogy with MORAN's I and GEARY's c are thus calculated (step 4):

$$II = \frac{n\sum_{i=1}^{n-1}\sum_{j>i}^{n} w_{ij}\sum_{k=1}^{S}(p_{ik}-\bar{p}_k)(p_{jk}-\bar{p}_k)}{W\sum_{i=1}^{n}\sum_{k=1}^{S}(p_{ik}-\bar{p}_k)^2}, \qquad (2)$$

and

$$cc = \frac{(n-1)\sum_{i=1}^{n-1}\sum_{j>i}^{n} w_{ij}\sum_{k=1}^{S}(p_{ik}-p_{jk})^2}{2W\sum_{i=1}^{n}\sum_{k=1}^{S}(p_{ik}-\bar{p}_k)^2}, \qquad (3)$$

where n is the sample size, W is the number of pairwise comparisons in the distance class of interest, p_{ik} and p_{jk} are the haplotypes (*i.e.*, the $\mathbf{p}$ values) of the ith and jth individuals, respectively, at the kth site, $\bar{p}_k$ is the kth element of the average vector, and the weights w_{ij} are 1 if individuals i and j fall in the distance class of interest, otherwise they are 0. Summation is over the S sites, and for all n individuals in the sample. It is easy to assign different weights to sequence differences if necessary, *i.e.*, to attribute a different importance to transitions and transversions, to substitutions and deletions, or to synonymous and nonsynonymous changes. In Equation (3), note that the summation of the quantities $p_{ik} - p_{jk}$ is equal to the number of sites differing between haplotypes i and j under the phenetic approach, whereas it is the number of evolutionary steps between haplotypes i and j under the evolutionary approach.

The AIDAs thus defined share the statistical properties of related MORAN's I and GEARY's c (CLIFF and ORD 1981; RIPLEY 1981). For large samples, they are distributed in the range $-1 \leq II \leq 1$ and $0 \leq cc \leq \propto$.

Their expected values, under a randomization hypothesis, are as follows:

$$\mathrm{E}(II) = \mathrm{E}(I) = -1/(n-1) \qquad (4)$$

and

$$\mathrm{E}(cc) = \mathrm{E}(c) = 1. \qquad (5)$$

Similarity between haplotypes, or positive autocorrelation, is shown by positive II values and by cc values close to 0; haplotype dissimilarity, or negative autocorrelation, results in II and cc values at the other extreme of the range.

Haplotype divergence is constrained by the genealogy of the sample studied. Occurrence of different mutations in different lineages causes linkage disequilibrium (SLATKIN 1994), especially when recombination is scarce or absent. The different sites in a DNA region, therefore, do not vary independently, and the variances of II and cc, necessary for step 5, cannot be estimated on the assumption of independence of the segregating sites. The simplest alternative is a randomization approach (MANLY 1991); a null distribution of II and cc values is constructed, assuming that haplotypes are randomly distributed in the geographical space.

To this end, sequences are randomly assigned to the sampled localities, retaining for each locality the observed sample size, and AIDA's are recalculated. This operation is repeated for N_1 times, so that an empirical distribution of II and cc values is obtained, under the assumption of random haplotype distribution. From such an empirical distribution, confidence limits are estimated for each distance class, and the probability can be evaluated that a certain level of autocorrelation, observed in a distance class, may occur under the null hypothesis of no spatial structuring (step 6). By increasing the value of N_1, any desired level of significance can be tested. This procedure, very similar to that employed for assessing the significance of MANTEL's tests of matrix correlation (MANTEL 1967; SMOUSE *et al.* 1986), may require large amounts of computation: a faster permutation scheme is therefore reccomended for large samples. In this case, confidence limits are calculated only once, sampling N_2 random pairs of haplotypes regardless of their spatial location. As before, this operation is repeated N_1 times to found the empirical distributions of II and cc. The lower N_1, the more conservative will the test result. We suggest to use a value of N_1 equal to the lowest number of comparisons observed in the distance classes analyzed.

Testing for departure from random expectations of the entire correlogram requires one of the procedures reviewed by ODEN (1984), the Bonferroni test (SOKAL and ROHLF 1981) being the simplest, if not the most sensitive. On the contrary, no established procedure exists for comparison of pairs of significant correlograms, although approximate tests can be conceived. In particular, bootstrapping (EFRON 1982) could be appropriate, and has actually been employed for the analysis of data in space (see *e.g.*, DIACONIS and EFRON 1983), including genetic data (CAVALLI-SFORZA *et al.* 1988; TEMPLETON 1993). However, a general consensus has not been reached yet on whether spatial correlograms meet the conditions necessary for application of jackknife and bootstrapping (N. L. ODEN, personal communication). Therefore, we prefer not to deal with this issue in the present paper. The pc program AIDA is available on request.

APPLICATION OF AIDA TO ARTIFICIALLY GENERATED PATTERNS

In this section, the statistic II is applied to a set of computer-generated data. At this stage, we are mainly interested in studying whether this statistic can objectively recognize a random distribution of haplotypes from a simple spatial pattern; we also want to evaluate the sensitivity of AIDA when increasing levels of noise are added to a nonrandom spatial pattern.

Spatial distributions of sequences (referred to as *surfaces*) were generated by assigning to each node of a square 10 × 10 lattice a haploid individual, whose haplotype is represented by one **p** vector with nine polymorphic sites ($S = 9$). The nodes of the lattice are thus the spatial locations of the 100 individuals considered. There would be no additional difficulties if two or more haplotypes were placed at the same location, but this did not happen in the simulations we are presenting. Ten different haplotypes were considered. They can be regarded as the products of successive mutations in a nonrecombining fragment of DNA, with each site being allowed to undergo mutation only once. Note that, because recombination is not considered to occur, the relative position of the polymorphic sites on the DNA molecule is irrelevant for this type of analysis. Haplotypes were assigned to the individuals of the lattice according to one of the following models of spatial variation.

Random variation: Under this model, 100 individual haplotypes were extracted from a rectangular distribution of the 10 possible haplotypes, and randomly assigned to the nodes of the lattice.

Pure gradient: We define a gradient as a pattern where the highest levels of haplotype differentiation are between individuals at the extremes of the simulated range, and individuals at intermediate locations also have intermediate haplotypes. The easiest way to simulate this, was by constructing a surface where the frequency of ones (or zeros) in the **p** vector increases along one direction. In evolutionary terms, this may be regarded as an extreme consequence of adaptation to an environmental gradient or as a result of a series of founder effects (as observed in experimental studies by EASTEAL 1988; RAPACZ *et al.* 1991; BOILEAU *et al.* 1992). An example of such a distribution is in Figure 1A.

```
  000000000 000000000 000000000
A 000000001 000000001 000000001
  000000011 000000011 000000011
  000000111 000000111 000000111
  000001111 000001111 000001111
  000011111 000011111 000011111
  000111111 000111111 000111111
  001111111 001111111 001111111
  011111111 011111111 011111111
  111111111 111111111 111111111


  000001111 000000011 000000000
B 000000111 000001111 000001111
  000111111 000000011 000000011
  000000011 000000001 000111111
  001111111 000111111 000000000
  000011111 000000111 111111111
  000000011 000011111 000001111
  000001111 011111111 001111111
  000111111 000111111 000011111
  000111111 000011111 000011111
```

FIGURE 1.—Examples of spatial distributions of haplotypes (**p** vectors) in two 3 × 10 sections of the data matrices generated under different models of spatial variation. (A) Pure North-South gradient; (B) Modified gradient, neighborhood size $d = 4$.

Modified gradient: Under this model, a pure gradient was initially generated as described above; then it was modified by replacing the haplotype at each node with a neighbor, randomly chosen within a distance of d nodes of the lattice, $d = 1, 2, 3, 4, 5, 6$ ($d = 0$ is the pure gradient). A modified gradient surface is represented in Figure 1B. Its evolutionary causes include the movement of individuals along a cline.

The distance classes for the correlograms were chosen following an equal-frequency criterion: each class included approximately the same number of pairwise comparisons. Overall, 64 correlograms were calculated, describing respectively one purely clinal surface, and nine replicates each for the random surface and for the six modified gradients.

Results: The results of the analysis are given as median values of nine realizations of the same stochastic process, except for the pure gradient model, which was deterministic, and therefore was analyzed only once.

The spatial distribution of haplotypes generated under the first model showed no significant autocorrelation in any of the seven distance classes. As expected when variation is random, all II values are very close to 0, and they appear to randomly fluctuate about their

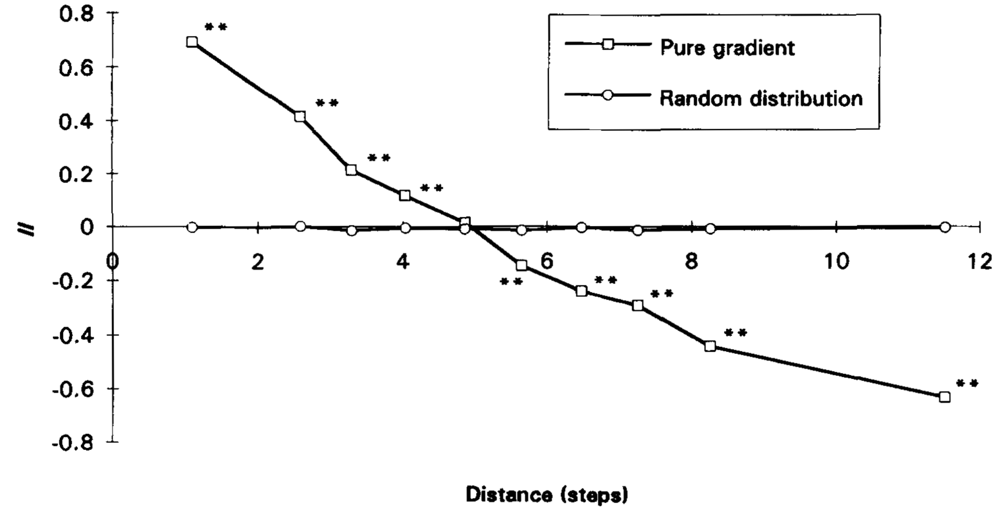

FIGURE 2.—Correlograms describing a pure gradient and a random distribution of haplotypes. For the random distribution of haplotypes, the correlogram is the median of nine independent realizations of the same process. Ordinate, *II*; abscissa, distance in steps, equal-frequency classes. * $P < 0.05$; ** $P < 0.01$.

expectation. The median correlogram summarizing nine random distributions is reported in Figure 2.

The pure gradient resulted in a monotonic decrease of autocorrelation coefficients, from highly significant positive at short distances, to highly significant negative at large distances (Figure 2). This autocorrelation profile is identical to the one observed for clines of allele frequencies both in simulation and in field studies (SOKAL 1979; SOKAL *et al.* 1989; SOKAL and JACQUEZ 1991).

When gradients are disturbed by random noise (Figure 3), the *II* values are closer to 0, and the autocorrelation profiles are less steep, as expected (SLATKIN and MARUYAMA 1975). However, even for $d = 5$, autocorrelation in the extreme distance classes retains significance.

APPLICATION OF AIDA TO mtDNA DATA

The data: As an example of application of AIDA to natural populations, we consider two RFLP data sets, which show similar overall levels of genetic diversity and are both subdivided in several subpopulations, arranged in a roughly linear manner along a transect.

The first data set comes from a study of the $2n = 60$ chromosome species of the mole-rat (*Spalax ehrenbergi*) in Israel. This is the second data set in the paper by NEVO *et al.* (1993), and it includes eight populations. Based on the analysis of restriction fragments generated by 10 endonucleases in 69 individuals, 21 polymorphic sites and 15 different mtDNA haplotypes were described.

The second data set refers to 12 North-American populations of the "*megarhyncha*" group of the fox sparrow (*Passerella iliaca*), analysed by ZINK (1994) using 18 endonucleases; 14 different mtDNA haplotypes were observed among 76 individuals, and 25 polymorphic sites were detected.

In both species, no haplotype was present in all the

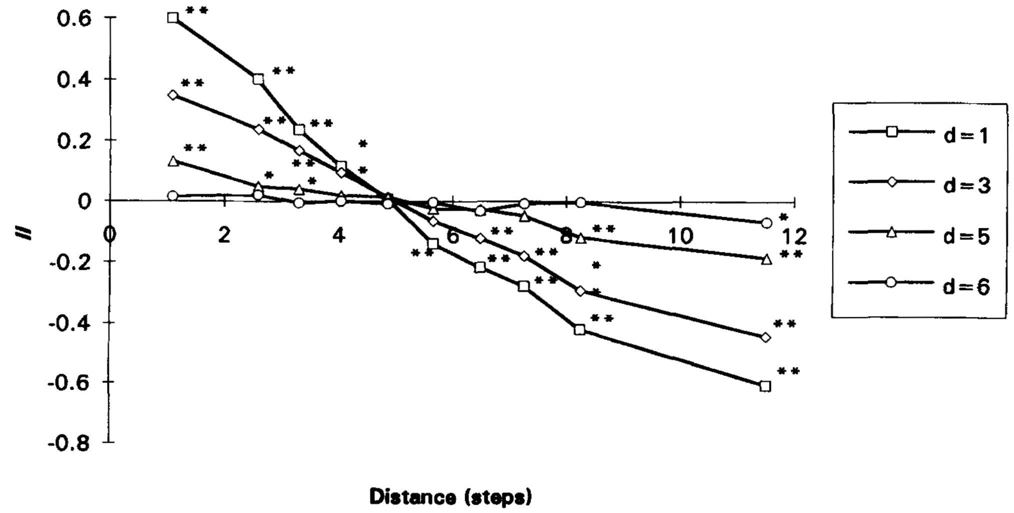

FIGURE 3.—Correlograms describing four modified gradients. Each correlogram is the median of nine independent realizations of the same process. Ordinate, *II*; abscissa, distance in steps, equal-frequency classes. * $P < 0.05$; ** $P < 0.01$.

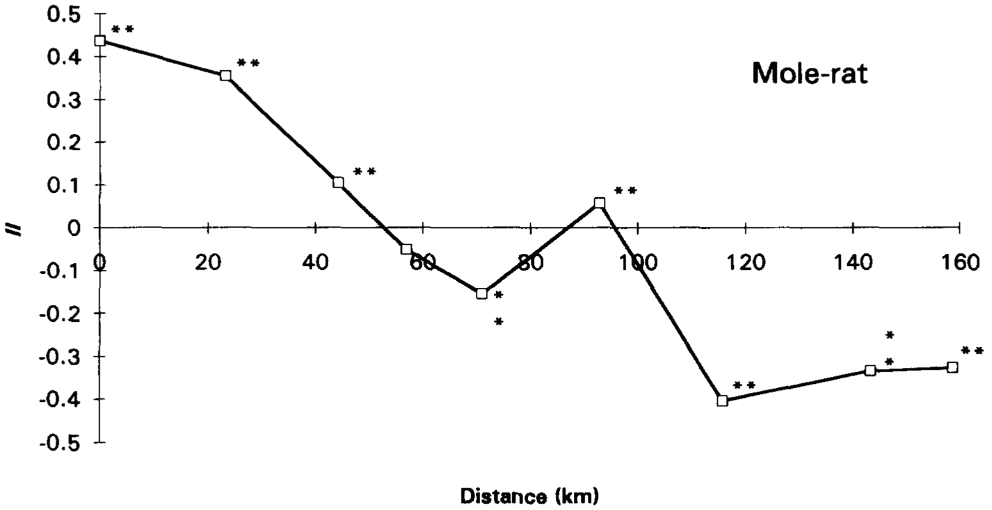

FIGURE 4.—Correlogram for the mtDNA data (NEVO *et al.* 1993) of the $2n = 60$ species of mole-rat (*Spalax ehrenbergi*) in Israel. Ordinate, *II*; abscissa, distance in kilometers. The first distance class (0 km) includes comparisons between haplotypes belonging to the same population. The other AIDA's are plotted in correspondence of eight class marks. * $P < 0.05$; ** $P < 0.01$.

subpopulations (in the mole-rat, no haplotype was present in more than three subpopulations) and the mean number of individuals per locality was <10. Therefore, the population structure could only approximately be described in terms of frequencies, because of the very large sampling errors. On the contrary, the AIDA statistics can be calculated even for such small samples.

Each individual was initially assigned the cartesian coordinates of its subpopulation. A matrix of distances between pairs of individuals was constructed; the maximum distance was 170 kilometers in the mole-rat data set and ~1500 km in the fox sparrow data set. Nine distance classes were chosen. The first of them corresponds to distance 0, *i.e.*, to within-population comparisons. The haplotypes were defined according to a phenetic approach.

Results: The autocorrelation indices show very different patterns of variation in the two data sets analysed (Figures 4 and 5).

In the mole-rat, the values of *II* (Figure 4) decrease almost continuously from positive significant to negative significant as the distance increases, resulting in an overall significant correlogram (Bonferroni's $P < 0.05$). This correlogram clearly resembles the ones summarizing the artificially generated gradients of the previous section. A significant negative peak ~70 km, followed by an upward fluctuation, shows that the basic clinal pattern was disturbed by additional evolutionary pressures. Strictly speaking, patterns of this kind are referred to as "long distance differentiation" in classical spatial autocorrelation studies (SOKAL *et al.* 1989; BARBUJANI *et al.* 1994); they are regarded as ancient clines on which the effects of successive gene flow, drift and/or adaptation to local environmental factors have been superimposed. In the present study, similar correlograms (not reported) have been observed for surfaces of modified gradients. As for the likely causes of such a significant spatial structure, after diverging from a presumably $2n = 58$ ancestral species, the $2n = 60$ mole-rat expanded southward in what is now Israel, colonizing new environments that became suitable in the Wurm pluvial period (NEVO 1989, 1991; NEVO *et al.* 1993). NEVO *et al.* (1993) describe qualitatively as a cline the pattern of haplotype variation they observed; they regard it as largely due to gene flow, *i.e.*, to colonization events occurring in the course of that range expansion. The AIDAs here calculated support quantitatively their conclusions, and are consistent with the effects of a population expansion accompanied by repeated founder effects (as in EASTEAL 1985). Levels of gene flow after the expansion need not have been extremely low; the quasi-continuous decline of *II* values at increasing distances shows that gene flow was sufficient to maintain a correlation between genetic differences and geographic distances of isolates. When we separately analysed by traditional autocorrelation the frequencies of the various morphs resulting from digestion by single enzymes (not given), the results were, by and large, consistent with AIDA, but inconclusive. For most morphs (*e.g.*, those generated by *Eco*RI and *Pst*I digestion), autocorrelation was positive and significant in the first distance class, and there was a declining trend. However, no correlogram achieved overall Bonferroni significance, probably because of the large errors due to the small number of comparisons in each distance class.

The "*megarhyncha*" group of the fox sparrow, on the other hand, shows non-significant values of *II* in all distance classes (Figure 5), and hence no geographical structure is apparent. Classification of this group as species or a subspecies is still controversial, but mtDNA data clearly suggest low levels of hybridization with other major groups of fox sparrow (ZINK 1994). The correlogram we evaluated is similar to the ones ob-

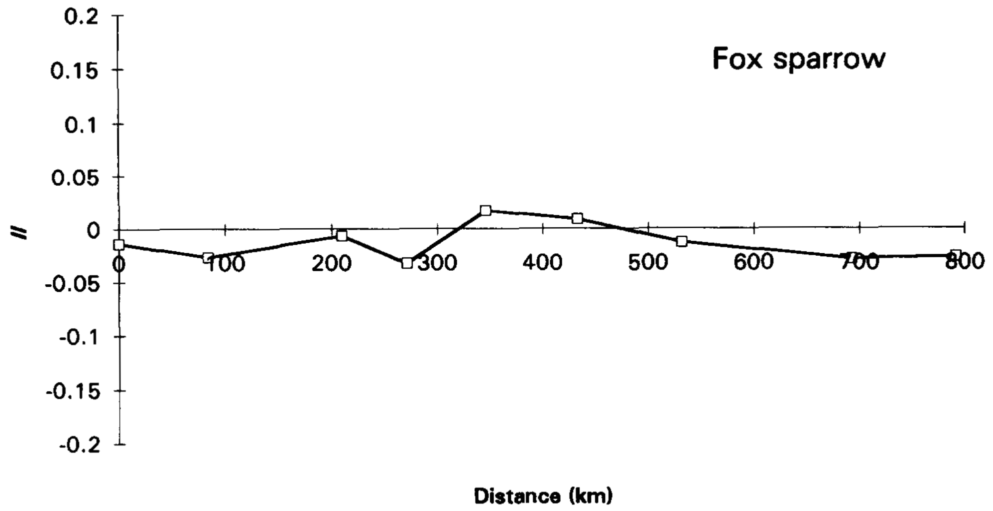

FIGURE 5.—Correlogram for the mtDNA data (ZINK 1994) of the "*megarhincha*" group of the fox sparrow (*Passerella iliaca*) in North America. Ordinate, *II*; abscissa, distance in kilometers. The first distance class (0 km) includes comparisons between haplotypes belonging to the same population. The other AIDA's are plotted in correspondence of eight class marks. * $P < 0.05$; ** $P < 0.01$.

served in the previous section when simulated haplotypes were randomly assigned to the spatial locations. Similar results were obtained by aplying AIDA's to the other three groups of fox sparrow studied by ZINK (1994). This finding supports the results found with other methods by ZINK (1994), and it is consistent with the hypothesis that gene flow has prevented geographic differentiation in this as well as in several passerine birds (*e.g.*, ZINK and DITTMANN 1993), where dispersal distances seem to be rather high.

DISCUSSION

Autocorrelation analysis has been used in population genetics to quantitatively describe geographic patterns of genetic variability, when for each population studied the frequencies of two or few alleles at different classical markers were known. The increasing amount of finer molecular data now available shows that the number of alleles can be much higher (even equal to the number of sampled individuals, when the sample size is small), and it allows sequence differences between alleles to be estimated. A classical approch to the study of the population structure seems thus improper when molecular data are concerned, both because the error in the frequency estimates is often large and because the information contained in the sequence would be neglected. The method described in this paper shares the general properties of previous autocorrelation statistics, but it is specifically designed for the analysis of molecular data.

When applied to artificially generated surfaces of DNA sequences or restriction haplotypes, AIDAs discriminated between random distribution of haplotypes and clinal patterns where genetic and geographic distances covaried; the sensitivity of AIDA is such that gradients were detected even when haplotypes traveled for substantial distances from their original position along the cline.

The results found when this method was applied to RFLP polymorphisms in mole-rat and fox sparrow populations gave quantitative support to earlier hypotheses on the evolutionary pressures affecting these species. Such hypotheses were based on the analysis of consensus trees or on the direct observation of the distribution of the haplotypes, but they had not been strictly tested. Compared with these analyses, AIDA seems to be more appropriate, both for its simplicity and because it allows significance testing.

A general limit of the sufaces artificially generated in this study is that the underlying genetic processes were not simulated. Their plausibility from the evolutionary standpoint will be briefly discussed later, and is being investigated using Monte Carlo simulations based on the coalescent process (HUDSON 1990). At any rate, computer simulations are unlikely to answer all the questions open; field studies and theoretical efforts are also necessary. Especially at the microgeographic level, descriptions of molecular variation are still scanty. As a consequence, the spatial patterns occurring in nature are still largely unknown. On the other hand, theory suggests that alleles generated in different realizations of the same genealogical process may be differently distributed in space, depending on the presence and on the impact of geographical factors; however, the role of barriers and distance has seldom been incorporated in comprehensive models so far (SLATKIN 1973; EPPERSON 1993).

Coming back to the artificial sufaces that we considered, random variation, as we simulated it, corresponds to a process whereby individual mobility is such that no spatial structure is retained, even where it may have existed. In natural populations a certain level of filopatry is generally present, and the distribution that we tested may prove exceedingly irregular. The analysis of the mtDNA data of four groups of North-American fox sparrows, however, suggests that high level of dispersion may lead to insignificant autocorrelation in all distance classes. That is actually what we observed for the simulated random surfaces.

Clines, on the other hand, may result from various evolutionary processes, which are reasonably understood only at the allele-frequency level (ENDLER 1977). It is still unclear whether the forces determining allele frequency clines also necessarily lead to clines in haplotype distributions. We expect that the genealogies of the genes considered may be a major determinant of spatial variation. As a consequence, the establishment of spatial patterns in the distribution of haplotypes may require long evolutionary times, much longer then the times leading to a significant structuring of allele or haplotype frequencies (SLATKIN 1987; BARBUJANI 1991). It seems, in other words, that patterns of genetic variation inferred from sequence and frequency data may reflect evolutionary phenomena ocurring at different time scales. The effects of recent demographic processes could be identified in the distribution of alleles or haplotypes, whereas sequence diversity accumulates depending on more remote events in the population history. There are exceptions to this rule. For instance, clines of haplotype frequencies generate clinal correlograms, in AIDA as well in traditional autocorrelation analysis, as is evident from the study of NEVO's *et al.* (1993) mole-rat data. At present, the results we obtained should be taken as a proof that certain clines in the distributions of sequences are actually recognized by one AIDA, even when they are not obvious at a visual inspection of data (Figure 2B). Nevertheless, some of them may prove to be improbable under more realistic assumptions on philogeny, selection and gene flow. Once again, the correlogram calculated for mole-rat mtDNA, are reassuring, in that *II* correctly identified a naturally occurring gradient and permitted us to assess its significance.

Finally, it does not seem unrealistic to assume that genetic similarity at short distance (*i.e.*, isolation by distance) is likely to produce asintotically decresing sets of autocorrelation indices, not only when frequencies are concerned (BARBUJANI 1987) but also when sequences are analysed by AIDAs. The surfaces expected under isolation by distance cannot be easily generated under the approch we adopted in this study. If this guess were true, however, the occurence of a distinct and recognizable isolation-by-distance autocorrelation pattern would increase the power of AIDA to support or reject various evolutionary hypotheses on the genetic structure of a population.

In conclusion, the autocorrelation indices we presented here objectively describe patterns of geographic variation and can therefore be used for exploratory data analysis. These patterns may point to the factors that have been important in the evolution of the genes of interest, and possibly of the population of interest. Tests of AIDAs against haplotype distributions generated by simulating a genealogic process (as in VALDES *et al.* 1993) may show to what extent AIDAs can also be employed to strictly test hypotheses on the evolution of polymorphism at the DNA level.

Our interest in spatial autocorrelation stems from the teaching of ROBERT SOKAL, to whom we are, as usual, indebted. Many thanks to MONTY SLATKIN, LAURENT EXCOFFIER, NEAL ODEN and an anonymous reviewer for critical reading of the manuscript, and to RICHARD NICHOLS, MAURIZIO BRIZZI and MIKE HENNESSY who gave useful suggestions. Financial support for this project was provided by the Italian Ministry of University (MURST) "60%" funds, and by the Italian CNR (grant 94-2270).

LITERATURE CITED

BARBUJANI, G., 1987 Autocorrelation of gene frequencies under isolation by distance. Genetics **117:** 777–782.

BARBUJANI, G., 1991 What do languages tell us on human microevolution? Trends Ecol. Evol. **6:** 151–155.

BARBUJANI, G., A. PILASTRO, S. DE DOMENICO and C. RENFREW, 1994 Genetic variation in North Africa and Eurasia: neolithic demic diffusion versus paleolithic colonization. Am. J. Phys. Anthropol. **95:** 137–154.

BARRANTES, R., P. E. SMOUSE, H. W. MOHRENWEISER, H. GERSHOWITZ, J. AZOFEIFA *et al.,* 1990 Microevolution in lower Central America: genetic characterization of the Chibcha-speaking groups of Costa Rica and Panama, and a consensus taxonomy based on genetic and linguistic affinity. Am. J. Hum. Genet. **46:** 63–84.

BOILEAU, M. G., P. D. N. HERBERT and S. S. SCHWARTZ, 1992 Non-equilibrium gene frequency divergence: persistent founder effects in natural populations. J. Evol. Biol. **5:** 25–40.

CAVALLI-SFORZA, L. L., A. PIAZZA, P. MENOZZI and J. MOUNTAIN, 1988 Reconstruction of human evolution. Bringing together genetic, archaeologic, and linguistic data. Proc. Natl. Acad. Sci. USA **85:** 6002–6006.

CLIFF, A. D., and J. K. ORD, 1981 *Spatial Processes.* Pion, London.

COSTA, R., A. A. PEIXOTO, J. T. THACKERAY, R. DALGLEISH and C. P. KYRIACOU, 1991 Length polymorphism in the Threonine-Glycine- encoding repeat region of the *period* gene in *Drosophila.* J. Mol. Evol. **32:** 238–246.

COSTA, R., A. A. PEIXOTO, G. BARBUJANI and C. P. KYRIACOU, 1992 A latitudinal cline in a Drosophila clock gene. Proc. R. Soc. Lond. B Biol. Sci. **250:** 43–49.

DIACONIS, P., and B. EFRON, 1983 Computer intensive methods in statistics. Sci. Am. **248:** 96–108.

EASTEAL, S., 1985 The ecological genetics of introduced populations of the giant toad, *Bufo marinus.* III. Geographical patterns of variation. Evolution **39:** 1065–1075.

EASTEAL, S., 1988 Range expansion and its genetic consequences in populations of the giant toad, *Bufo marinus.* Evol. Biol. **23:** 49–84.

EFRON, B., 1982 *The Jackknife, the Bootstrap and Other Resampling Plans.* Soc. Industr. Appl. Math., Philadelphia, PA.

ENDLER, J. A., 1977 *Geographic Variation, Speciation, and Clines.* Princeton University Press, Princeton, NJ.

EPPERSON, B., 1993 Spatial and space-time correlations in systems of subpopulations with genetic drift and mutation. Genetics **133:** 711–727.

EXCOFFIER, L., and P. E. SMOUSE, 1994 Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: molecular variance parsimony. Genetics **136:** 343–359.

EXCOFFIER, L., P. E. SMOUSE and J. M. QUATTRO, 1992 Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics **131:** 479–491.

Felsenstein, J., 1982 How can we infer history and geography from gene frequencies? J. Theor. Biol. **96:** 9–20.

Harding, R. M., 1993 VNTRs in review. Evol. Anthropol. **1:** 62–71.

Hudson, R. R., 1990 Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. **7:** 1–44.

Kimura, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics **61:** 893–903.

Kreitman, M., 1983 Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster.* Nature **304:** 412–417.

Manly, B. F. J., 1991 *Randomization and Monte Carlo Methods in Biology.* Chapman and Hall, London.

Mantel, N. 1967 The detection of disease clustering and a generalized regression approach. Cancer Res. **27:** 209–220.

Nevo, E., 1989 Modes of speciation: the nature and role of peripheral isolates in the origin of species, pp. 295–236 in *Genetics, Speciation and the Founder Principle,* edited by L. V. Giddings, K. Y. Kaneshiro and W. W. Anderson. Oxford University Press, NY.

Nevo, E., 1991 Evolutionary theory and processes of active speciation and adptive radiation in subterranean mole rats, *Spalax ehrenbergi* superspecies in Israel. Evol. Biol. **25:** 1–125.

Nevo, E., R. L. Honeycutt, H. Yonekawa, K. Nelson and N. Hanzawa, 1993 Mitochondrial DNA polymorphism in subterranean mole-rats of the *Spalax ehrenbergi* superspecies in Israel, and its peripheral isolates. Mol. Biol. Evol. **10:** 590–604.

Oden, N. L., 1984 Assessing the significance of a spatial correlogram. Geogr. Anal. **16:** 1–16.

Pigliucci, M., and G. Barbujani, 1991 Geographical patterns of gene frequencies in Italian populations of *Ornithogalum montanum* (Liliaceae). Genet. Res. **58:** 95–104.

Prim, R. C., 1957 Shortest connection networks and some generalizations. Bell Syst. Techn. J. **36:** 1389–1401.

Rapacz, J., L. Chen, E. Butler-Brunner, M.-J. Wu, J. O. Hasler-Rapacz *et al.,* 1991 Identification of the ancestral haplotype for apolipoprotein B suggests an African origin of *Homo sapiens sapiens* and traces their subsequent migration to Europe and the Pacific. Proc. Natl. Acad. Sci. USA **88:** 1403–1406.

Ripley, B., 1981 *Spatial Statistics.* John Wiley and Sons, New York.

Rohlf, F. J., 1970 Adaptive hierarchical clustering schemes. Syst. Zool. **19:** 58–82.

Slatkin, M., 1973 Gene flow and selection in a cline. Genetics **75:** 733–756.

Slatkin, M., 1987 Gene flow and the geographic structure of natural populations. Science **236:** 787–792.

Slatkin, M., 1989 Detecting small amounts of gene flow from phylogenies of alleles. Genetics **121:** 609–612.

Slatkin, M., 1991 Inbreeding coefficients and coalescence times. Genet. Res. **58:** 167–175.

Slatkin, M., 1993 Isolation by distance in equilibrium and nonequilibrium populations. Evolution **47:** 264–279.

Slatkin, M., 1994 Linkage disequilibrium in growing and stable populations. Genetics **137**: 331–336.

Slatkin, M., and H. E. Arter, 1991 Spatial autocorrelation methods in population genetics. Am. Nat. **138:** 499–517.

Slatkin, M., and W. P. Maddison, 1989 A cladistic measure of gene flow inferred from the phylogenies of alleles. Genetics **123:** 603–613.

Slatkin, M., and W. P. Maddison, 1990 Detecting isolation by distance using phylogenies of genes. Genetics **126:** 249–260.

Slatkin, M., and T. Maruyama, 1975 Genetic drift in a cline. Genetics **81:** 209–222.

Smouse, P. E., J. C. Long and R. R. Sokal, 1986 Multiple regression and autocorrelation extensions of the Mantel test of matrix correspondence. Syst. Zool. **35:** 627–632.

Sokal, R. R., 1979 Ecological parameters inferred from spatial correlograms, pp.167–196 in *Contemporary Quantitative Ecology and Related Ecometrics,* edited by G. P. Patil and M. Rosenzweig. International Co-operative Publishing House, Fairland, MD.

Sokal, R. R., R. H. Harding and N. L. Oden, 1989 Spatial patterns of human gene frequencies in Europe. Am. J. Phys. Anthropol. **80:** 267–294.

Sokal, R. R., and G. M. Jacquez, 1991 Testing inferences about microevolutionary processes by means of spatial autocorrelation analysis. Evolution **45:** 152–168.

Sokal, R. R., and N. L. Oden, 1978a Spatial autocorrelation analysis in biology. I. Methodology. Biol. J. Linn. Soc. **10:** 199–228.

Sokal, R. R., and N. L. Oden, 1978b Spatial autocorrelation analysis in biology. II. Some biological implications and four applications of evolutionary and ecological interest. Biol. J. Linn. Soc. **10:** 229–249

Sokal, R. R., and F. J. Rohlf, 1981 *Biometry,* Ed. 2. Freeman, San Francisco.

Sokal, R. R., N. L. Oden and J. S. F. Barker, 1987 Spatial structure in Drosophila buzzatii populations: simple and directional spatial autocorrelation. Am. Nat. **129:** 122–142.

Templeton, A. R., 1993 The "Eve" hypothesis: a genetic critique and reanalysis. Am. Anthropol. **95:** 51–72.

Templeton, A. R., E. Boerwinkle and C. F. Sing, 1987 A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in Drosophila. Genetics **117:** 343–351.

Valdes, A. M., M. Slatkin and N. B. Freimer, 1993 Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics **133:** 737–749.

Wartenberg, D. E., 1989 *SAAP, A Spatial Autocorrelation Analysis Program. Version 4.3.* Exeter Software, Setauket, NY.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

Wijsman, E. M., and L. L. Cavalli-Sforza, 1984 Migration and genetic population structure, with special reference to humans. Annu. Rev. Ecol. Syst. **15:** 279–301.

Wright, S., 1965 The interpretation of population structure by *F*-statistics with special regards to system of mating. Evolution **19**: 395–420.

Zink, R. M., 1994 The geography of mitochondrial DNA variation, population structure, hybridization, and species limits in the fox sparrow (*Passerella iliaca*). Evolution **48:** 96–111.

Zink, R. M., and D. L. Dittmann, 1993 Gene flow, refugia, and evolution of geographic variation in the song sparrow (*Melospiza melodia*). Evolution **47**: 717–729.

Communicating editor: G. B. Golding