# Global fractal dimension of human DNA sequences treated as pseudorandom walks

Cheryl L. Berthelsen
*Department of Medical Informatics, Genetic Epidemiology, University of Utah, 420 Chipeta Way, Suite 180, Salt Lake City, Utah 84108*

James A. Glazier*
*Research Institute of Electrical Communication, Tohoku University, Sendai 980, Japan*

Mark H. Skolnick
*Department of Medical Informatics, Genetic Epidemiology, University of Utah, 420 Chipeta Way, Suite 180, Salt Lake City, Utah 84108*

(Received 5 December 1991; revised manuscript received 19 February 1992)

We used a pseudorandom-walk representation in a four-dimensional embedding to estimate the global fractal dimension $D$ of 164 sequences from GenBank and generated length-matched control sequences of three types: random, matched in base content, and matched in dimer content. The mean $D$ of the sequences was $1.631 \pm 0.137$. This $D$ was significantly lower than the $D$'s for all three control types, indicating the presence of significant information content in DNA sequences not explained by base or dimer frequencies. This variation was due largely to nonuniform distribution of bases and dimers within DNA sequences. The $D$ of genomic DNA sequences was different from the $D$ of messenger RNA sequences.

PACS number(s): 87.10.+e, 05.45.+b, 02.70.+d

## INTRODUCTION

The volume of DNA sequence data available for analysis is increasing dramatically. One of the challenges of sequence analysis is to determine patterns in sequences that have no current explanation and therefore potentially point toward structures that require further exploration. It is also useful to distinguish coding from noncoding sequences. Fractal analysis is a relatively new analytical technique that has proven useful in revealing complex patterns in natural objects. This paper represents our initial application of fractal analysis to pseudorandom walks derived from genomic DNA and messenger RNA (mRNA) sequences.

The numerical relationship between two correlated properties of an object may be simple. For example, the area of a square increases with the square of the length of its sides. In this case the area is said to *scale* as the square of the length. The relationship may also be complicated. For example, an animal's weight may increase as a power of its age at certain periods and be independent of age at others. For an arbitrarily chosen pair of properties there will generally be no simple relationship. However, if one of them can be written as a simple power-law function of the other with a single uniform exponent, i.e., property $1 \approx (\text{property } 2)^{\alpha}$, then we say that the pair exhibits *scaling* with exponent $\alpha$.

Many complex spatial patterns possess well-defined scaling behaviors. One example is the global fractal dimension $D$, which is the exponent describing the increase of an object's mass with its size. A particular scaling exponent is often associated with a particular process that gave rise to the pattern. Thus, scaling can be used to define classes of patterns. The global fractal dimension is the most common of these scaling exponents and has been used to study patterns in biology, physics, economics, biochemistry, populations genetics, and epidemiology.

While $D$ may be calculated for an arbitrary object, it is most informative in treating structures generated by low-dimensional deterministic rules such as period doubling [1]. However, a deterministic structure is unlikely for DNA sequences, given the large amount of information needed to specify an organism. To a first approximation, we know that DNA behaves like a random sequence so any direct measurement of the $D$ of DNA sequences will yield arbitrarily large dimensions. Nevertheless, when a DNA sequence is treated as a list of pseudorandom numbers and used to generate a pseudorandom walk, deviations from typical random-walk behavior are immediately apparent as long periodic, correlated, and anticorrelated subsequences [2]. The pseudorandom walks derived from DNA sequences in Figs. 1(a)–1(f) show obvious qualitative deviations from the paired random walks shown in Figs. 2(a)–2(f). The large-scale structure of the walks reflects underlying correlations within the sequences. The $D$ of the walk should reflect the overall importance of these correlations, since it quantifies the average density of the clustering of data points in the walk and ignores localized behavior.

Recently, there have been efforts to apply the techniques of chaos theory to molecular biology. This effort has been made at all levels from protein folding [3–8] to three-dimensional structures of DNA [9] and RNA [10]. There also have been some limited fractal analysis of DNA sequences. (See *Note added in proof.*)

Gates [2] suggested that nucleic acid sequences could be represented as random walks in two-dimensional space

 

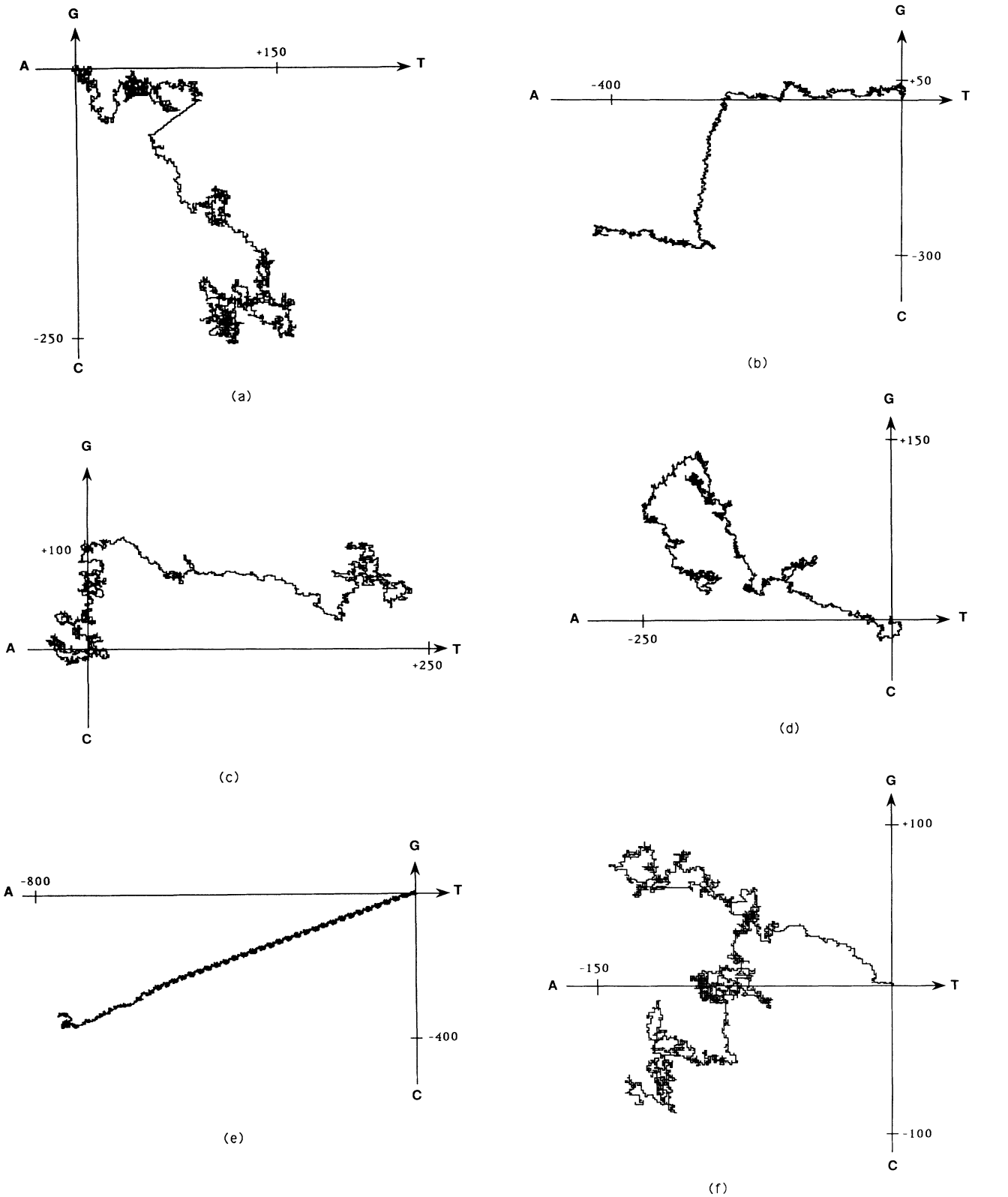


FIG. 1. Pseudorandom walks of six human DNA sequences from GenBank using the two-dimensional embedding scheme. They are qualitatively quite different from the random walks shown in Fig. 2. (a) Human opsin gene, accession no. K02281, 6953 base pairs (bps), $D=1.650$; (b) human factor V mRNA, accession no. M16967, 6909 bps, $D=1.490$; (c) human T-cell receptor germline beta-chain, accession no. M14158, 4913 bps, $D=1.541$; (d) human *PRH1* gene (*Hae* II–type subfamily), accession no. M13057, 4946 bps, $D=1.532$; (e) human mRNA for apolipoprotein (a), accession no. X06696 M17399, 13 938 bps, $D=1.547$; (f) human alpha-1-acid glycoprotein 2, accession no. M21540, 4944 bps, $D=1.671$.

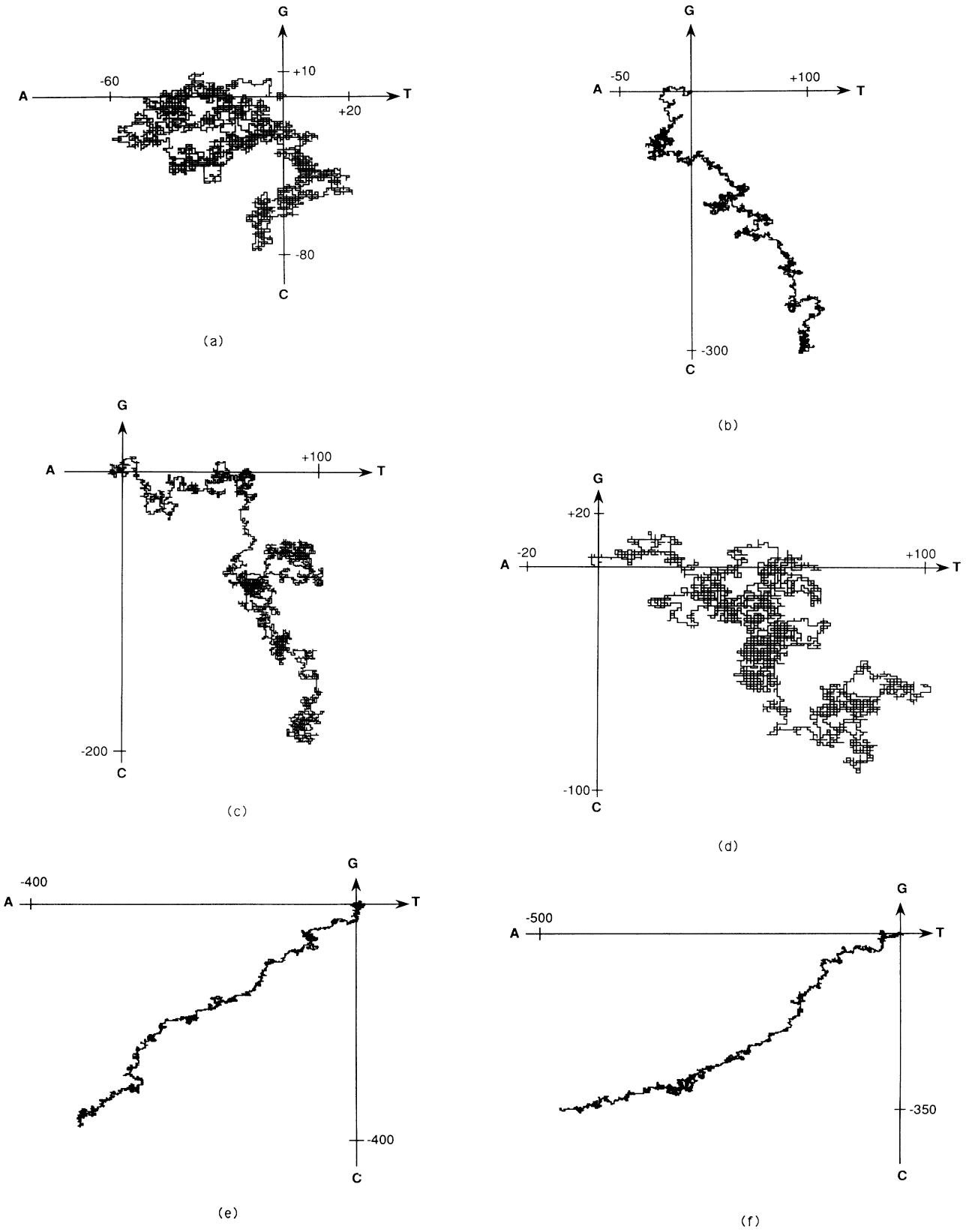


FIG. 2. Pseudorandom walks of random, base-matched, and dimer-matched control sequences for the sequences shown in Figs. 1(a) and 1(b). (a) Random control for human opsin gene, $D=1.891$. (b) base-matched control for human opsin gene, $D=1.701$; (c) dimer-matched control for human opsin gene, $D=1.744$; (d) random control for human factor V mRNA, $D=1.779$; (e) base-matched control for human factor V mRNA, $D=1.536$; (f) dimer-matched control for human factor V mRNA, $D=1.600$.

with the bases cytosine and guanine (C,G) on opposite ends of one axis and adenine and thymine (A,T) on the other. He suggested the following two methods to calculate $D$: (i) the logarithm of the number of bases divided by the logarithm of the Euclidean distance [11] between end points of the random walk, and (ii) the logarithm of the number of bases divided by the logarithm of the Manhattan distance [12] between the end points of the random walk.

Luo and Tsai [13] calculated the $D$ of nucleic acid sequences from 14 different organisms to study the relationship between $D$ and the evolutionary complexity of organisms. They represented nucleic acid sequences as random walks in a two-dimensional space using the same scheme as Gates and calculated the $D$ using the mean-square separation between end points of a segment of the sequence containing $N$ bases. The standard deviations for their estimates of $D$ were generally 20% of the mean value of $D$. They found that $D$ increased with organism complexity and that it correlated statistically with the entropy measure from information theory [14]. Their study assumed that the $D$ of a single, relatively short DNA sequence is representative of all the DNA of an organism.

Jeffrey [15] investigated a graphic representation of nucleic acid sequence using iterated function systems [16]. He represented DNA sequences by points within a square, with each base represented by a corner of the square. The first point, representing the first base in the sequence, is plotted halfway between the center of the square and the corner representing that base. Subsequent bases are plotted as points located halfway between the previous point and the corner representing the base. The result is a bit-mapped image with sparse areas representing rare subsequences and dense regions representing common subsequences. He found interesting visual patterns in nucleic acid sequences, but did not attempt mathematical characterization or estimation of $D$.

In a recent abstract, Lim [17] presented a fractal analysis of sequence data. He found that introns and exons are distinct and suggested that fractal techniques could be used to create a classification scheme.

These four fractal analyses of DNA sequence data have produced suggestive results on the utility of chaos techniques. However, they are limited to estimating $D$ for a small number of short sequences, generally in two dimensions.

We present a detailed exploration of fractal analysis including the estimation of $D$. We employ the technique of pseudorandom control sequences to evaluate the $D$'s of 164 relatively long human sequences (4500–15 000 bases). The issues of adequate sequence length, proper embedding dimension, and scaling ranges have been addressed in the design of our calculations and analyses.

## METHODS

### Methodological issues

An important rule in fractal theory involves the choice of the embedding dimension used to represent the data. The embedding dimension must be at least as high as the highest possible $D$ rounded up to the next whole number plus 1 [18,19]. Methods of estimating $D$ are known to be biased. The bias in the estimate of $D$ increases with embedding dimension, since higher embedding dimensions require longer sequences to reduce the effect of bias [20]. Minimizing the embedding dimension minimizes computation time as well as bias.

The choice of DNA sequences of sufficient length is equally important. It has been suggested that the number of points required to estimate the correlation dimension [21] within 5% of its true value is at least $42^M$ where $M$ is the largest integer less than the correlation dimension of the fractal [22]. However, Ramsey [20] found that for simple models, 5000 is a rough lower bound for the number of points needed to achieve reasonable results. The estimate obtained for deterministically generated data sets is known to be biased high, but this bias decreases with sequence length and estimates for random noise are actually biased low [20]. The effect of finite length may be reduced by applying the widest possible range of scaling. The resolution is limited by the area visited by the random walk, since the area is a function of the number of steps in the walk.

To evaluate the effect of finite sequence length on the estimate of $D$, we generated random sequences of equal base frequencies over a range of lengths, 25 of each length, and estimated $D$ for each as discussed below. Figure 3 demonstrates that the average estimate of $D$ increases with length and the standard deviation of the estimate of $D$ decreases with length. The mean $D$ of random sequences of length 50 000 is 1.93 with a standard deviation of 0.03. The standard deviation of $D$ is only 1.6% of the mean at this length. The mean $D$ for random sequences of length 5000 is 1.847 with a standard deviation of 0.101. The standard deviation of $D$ is 5.5% of the mean at length 5000. Although we do not converge to $D=2$ at these lengths, we can control for finite-length errors by always comparing length-matched sequences. DNA and control sequences of the same length will be

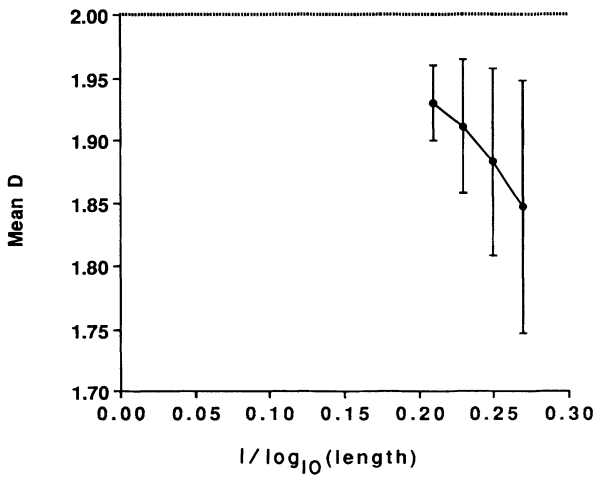


FIG. 3. $D$ vs length for random sequences. $D$ increases with the length of random sequences and its standard deviation decreases.

affected by finite-length errors in precisely the same way, which will not affect our statistical analyses. Therefore, reasonable estimates of $D$ are possible for random walks at least 4000 bases in length with a standard deviation of $D$ less than 6% of the mean. We have addressed these convergence issues in greater detail elsewhere [23]. To assure convergence to a $D$ of 1.995 we would need a sequence longer than 500 000, which would rule out analysis of DNA sequences, since they are typically 5000–50 000 bases in length.

We selected human sequences from GenBank version 55 for analysis based entirely on the length of the sequence. There were 164 human nucleic acid sequences of length 4500 to 15 000 (average length, 7178 bases) and all were included in the study. Although 57 mRNA sequences are included, we loosely refer to the entire set as DNA sequences. The mRNA sequences are composed primarily of coding segments. The remaining 107 came from genomic DNA and have coding segments separated by introns and other noncoding segments that comprise the majority of the bases in the sequence. The sequences we analyzed are not completely representative of the human genome. Several sequences are from gene families, some chromosomes are underrepresented, and the sample is severely deficient in noncoding DNA. Since the sequences came from GenBank, our sample includes mostly important or interesting genes rather than genomic sequences in general.

### Random-walk representation

DNA is composed of two strands that bind together by a specific base-pairing rule. Adenine (A) always pairs with thymine (T) and cytosine (C) always pairs with guanine (G). DNA replicates in a very specific manner. A new strand elongates by placing the next base on what is referred to as the 3′ end. DNA sequences are listed in 5′ to 3′ order by convention, 5′ being the "head" of the sequence and 3′ the "tail." The complementary strand runs in an antiparallel direction. The following example illustrates the antiparallel nature of the two strands of DNA and the base-pair bonding rule

```
5'->AACTGGGATATATTTGGG->3'
    ||||||||||||||||||
3'<-TTGACCCTATATAAACCC<-5'
```

so the subsequence 5′→AACTGG→3′ on one strand is biologically identical to 5′→CCAGTT→3′ on the complementary strand. In general, DNA with both strands bound together does not have direction, so neither stand has precedence. When DNA is being copied to make messenger RNA, however, the two strands are distinct. One strand is called the sense strand and the other the coding strand. A mRNA sequence is synthesized from 5′ to 3′ using the sense strand template read from 3′ to 5′. A protein is made from this mRNA by grouping the bases into triplets called codons. To summarize, DNA is double-stranded and has no direction. The two complementary strands are synonymous and biologically indistinguishable; mRNA, on the other hand, is single-stranded and its base sequence is biologically distinct from its reverse complement. Both direction and codon framing are important for mRNA.

We use the following natural method to convert a DNA sequence into a pseudorandom walk in $N$ dimensions. We represent a DNA sequence as a series of vectors $\mathbf{x}_j$ representing the four base types A, C, G, and T. The complementary base pairing of A with T and C with G suggests a natural embedding of a sequence into a two-dimensional space. We chose the axis assignments to specifically represent DNA, so the representation was strand independent. In two dimensions, any single-base axis assignment will produce a dimension that is the same for a sequence and its complement. The fractal dimension of a sequence is unchanged under a transformation if the transformation causes only a reflection or rotation of the pseudorandom-walk structure and does not take any nonzero length trajectories into zero length trajectories. Higher embedding dimensions will not yield this result for complements in all representations. The requirement that $D$ be unchanged for complements is the only biological constraint on our representation. However, a natural representation should also yield a $D$ unchanged by certain symmetry operations. For a true random walk, our assignment of the symbol types is arbitrary and the direction in which we read should be unimportant. Therefore, to preserve the symmetries of the random walk in our embedding structure, we require the following.

(i) *Complementarity.* The estimate of $D$ must be the same for both DNA strands; that is, a strand read 5′ to 3′ will produce the same $D$ for its reverse complement read 5′ to 3′.

(ii) *Reflection symmetry.* The estimate of $D$ must be the same for a single strand regardless of reading direction; that is, a strand read 5′ to 3′ will produce the same $D$ if read 3′ to 5′.

(iii) *Compatibility.* Representations of different embeddings must be compatible; that is, dimers that produce the same trajectory in a higher dimension do so in a lower-dimensional scheme.

(iv) *Substitution symmetry.* $D$ remains unchanged under the single exchange of either A↔T or G↔C.

Note that (i) is a special case of (ii) and (iv). (iv) is also suggested by the natural biological grouping of A and T as weak-bonding bases and G and C as strong-bonding bases.

In two dimensions, complementarity (i), reflection symmetry (ii), and compatibility (iii) are satisfied for any single-base representation. Substitution symmetry (iv), however, requires that $\{A\}=-\{T\}$ and $\{G\}=-\{C\}$. Otherwise, the sequence AG is a zero-step trajectory while AC is not, which violates substitution symmetry (iv). Therefore, we employ the following axis assignments.

*Axis 1.* $\{A\}=(-1,0)$ and $\{T\}=(1,0)$.
*Axis 2.* $\{C\}=(0,-1)$ and $\{G\}=(0,1)$.

In four and higher dimensions we begin a new $\mathbf{y}_i$ for each base in the sequence. Thus our representation is indepen-

dent of our reading frame and each base is used in two successive vectors. We determine our four-dimensional embedding as follows: Complementarity (i) requires that $\{AA\}=-\{TT\}$, $\{AC\}=-\{GT\}$, $\{AG\}=-\{CT\}$, $\{AT\}=-\{AT\}$, $\{CA\}=-\{TG\}$, $\{CC\}=-\{GG\}$, $\{CG\}=-\{CG\}$, $\{GA\}=-\{TC\}$, $\{GC\}=-\{GC\}$, and $\{TA\}=-\{TA\}$. Reflection symmetry (ii) requires that $\{AC\}=\{CA\}$, $\{AG\}=\{GA\}$, $\{AT\}=\{TA\}$, $\{CG\}=\{GC\}$, $\{CT\}=\{TC\}$, and $\{GT\}=\{TG\}$. Thus, we must group $\{AC\}$, $\{CA\}$, $\{GT\}$, and $\{TG\}$ on one axis and $\{AG\}$, $\{GA\}$, $\{CT\}$, and $\{TC\}$ on another axis. If we pair $\{AC\}=\{TG\}$, rather than $\{AC\}=\{CA\}$, we violate compatibility (iii) with our two-dimensional scheme. Similarly, we cannot set $\{AC\}=\{CA\}=\{GT\}=\{TG\}=0$ by (iii) or $\{AG\}=\{GA\}=\{CT\}=\{TC\}=0$ by substitution symmetry (iv). So our only grouping is $\{AC\}=\{CA\}=-\{GT\}=-\{TG\}\neq 0$ and $\{AG\}=\{GA\}=-\{CT\}=-\{TC\}\neq 0$. In four dimensions the only scheme that obeys these conditions is the following:

*Axis 1.* $\{AA\}=(-1,0,0,0)$ and $\{TT\}=(1,0,0,0)$.
*Axis 2.* $\{CC\}=(0,-1,0,0)$ and $\{GG\}=(0,1,0,0)$.
*Axis 3.* $\{AC\}=\{CA\}=(0,0,-1,0)$ and $\{GT\}=\{TG\}=(0,0,1,0)$.
*Axis 4.* $\{AG\}=\{GA\}=(0,0,0,-1)$ and $\{CT\}=\{TC\}=(0,0,0,1)$, $\{AT\}=\{TA\}=\{CG\}=\{GC\}=(0,0,0,0)$.

Axes 3 and 4 correspond to dimers that form 45° angle lines in two dimensions. Figure 4 gives a graphic representation of our two-dimensional and four-dimensional embedding schemes.

To expand to six dimensions, we can either split $\{AT\},\{TA\},\{CG\},\{GC\}$ or our other two quartets ($\{AC\},\{CA\},\{GT\},\{TG\}$ and $\{AG\},\{GA\},\{CT\},\{TC\}$) on axes 3 and 4. We choose the latter to preserve the compatibility of the assignments of the zero trajectories in lower-dimensional embeddings as much as possible. However, our choice here is arbitrary. (We do not use the six-dimensional embedding for any of the major calculations reported in this paper.) Our six-dimensional embedding is therefore the following.

*Axis 1.* $\{AA\}=(-1,0,0,0,0,0)$ and $\{TT\}=(1,0,0,0,0,0)$.
*Axis 2.* $\{CC\}=(0,-1,0,0,0,0)$ and $\{GG\}=(0,1,0,0,0,0)$.
*Axis 3.* $\{AC\}=(0,0,-1,0,0,0)$ and $\{CA\}=(0,0,1,0,0,0)$.
*Axis 4.* $\{GT\}=(0,0,0,-1,0,0)$ and $\{TG\}=(0,0,0,1,0,0)$.
*Axis 5.* $\{AG\}=(0,0,0,0,-1,0)$ and $\{GA\}=(0,0,0,0,1,0)$.
*Axis 6.* $\{CT\}=(0,0,0,0,0-1)$ and $\{TC\}=(0,0,0,0,0,1)$, $\{AT\}=\{TA\}=\{CG\}=\{GC\}=(0,0,0,0,0,0)$.

There is only one possible dimer-pair eight-dimensional embedding. We add two axes for the AT-TA and CG-GC pairs that were zero-step trajectories in lower dimensions.

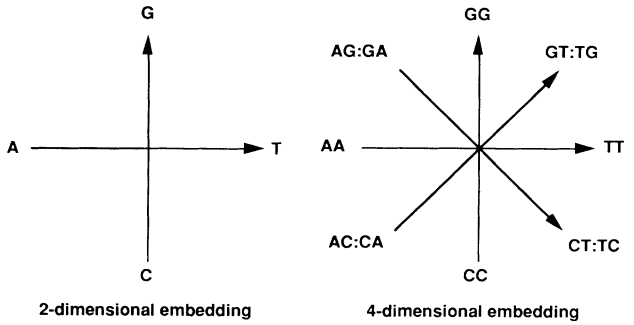


FIG. 4. Embedding schemes in two and four dimensions for pseudorandom-walk representations of DNA sequences.

*Axis 1.* $\{AA\}=(-1,0,0,0,0,0,0,0)$ and $\{TT\}=(1,0,0,0,0,0,0,0)$.
*Axis 2.* $\{CC\}=(0,-1,0,0,0,0,0,0)$ and $\{GG\}=(0,1,0,0,0,0,0,0)$.
*Axis 3.* $\{AC\}=(0,0,-1,0,0,0,0,0)$ and $\{CA\}=(0,0,1,0,0,0,0,0)$.
*Axis 4.* $\{GT\}=(0,0,0,-1,0,0,0,0)$ and $\{TG\}=(0,0,0,1,0,0,0,0)$.
*Axis 5.* $\{AG\}=(0,0,0,0,-1,0,0,0)$ and $\{GA\}=(0,0,0,0,1,0,0,0)$.
*Axis 6.* $\{CT\}=(0,0,0,0,0,-1,0,0)$ and $\{TC\}=(0,0,0,0,0,1,0,0)$.
*Axis 7.* $\{AT\}=(0,0,0,0,0,0,-1,0)$ and $\{TA\}=(0,0,0,0,0,0,1,0)$.
*Axis 8.* $\{CG\}=\{0,0,0,0,0,0,0,-1)$ and $\{GC\}=(0,0,0,0,0,0,0,1)$.

In future studies we can exploit the fact that our estimate of $D$ is not independent of representation by relaxing our symmetry conditions to allow other representations. DNA sequences that code for protein have a strong bias for content and arrangement of certain dimers. We can change our axis assignments to emphasize these dimers in the estimated fractal dimension and evaluate whether the global fractal dimension based on a strand-dependent scheme is useful in determining which strand is the coding strand and which is the sense strand.

The pseudorandom walk is defined as the sequence $\{\mathbf{y}_i\}$, where

$$\mathbf{y}_i \equiv \sum_{j=1}^{i} \mathbf{x}_j \ . \tag{1}$$

A true random walk of infinite length is space filling in a two-dimensional embedding and has $D=2$ for higher-dimensional embeddings [24,25]. However, our representation of a DNA sequence in dimensions between 2 and 8 is not a true random walk. In two dimensions, a subsequence such as ATATAT produces alternating steps between two points in the random walk. In four dimensions, ATATAT produces steps of size zero. Since we allow multiple visits to lattice sites in our calculations, this produces small localized regions with increased density, which decreases the estimated global fractal dimension slightly. The graph in Fig. 3 reflects this effect with $D$ converging to $D=1.93$ rather than $D=2$ for a random sequence as long as 50 000 bases. Some of this error is due to finite-length effects and some is due to our repre-

sentation, which is not a true random walk. We address convergence of a true random walk in detail in a separate paper [23]. As stated before, this lack of convergence does not affect the statistical results because finite-length effects are the same for DNA and length-matched controls. Our standard deviations are less than 7% of the mean for the shortest sequence evaluated, which allows reliable statistical analyses despite lack of true convergence.

### Control sequences

We generated controls matched to the length of each DNA sequence using a random number generator. Pairing the DNA sequences with control sequences of the same length allows us to control for the effects of finite length. To avoid introducing sequential correlations in our sequences, we used a linear congruential random number generator with a randomized shuffle with a period of at least 714 025 [26]. We used the following three types of control sequences: (i) random controls, where each base occurs with a probability of 0.25; (ii) base-matched controls, where the frequency of each base is determined from the DNA sequence and that frequency is used to generate the control sequences; and (iii) dimer-matched controls, where the frequency of each dinucleotide pair is determined and the control sequences are generated using the probability of what the next base will be, given the base selected previously.

### Estimating global fractal dimension

We calculated the global dimension known as the Hausdorff dimension using the sandbox method [27]. We used the sandbox method rather than the more widely used box-counting method of Grassberger and Procaccia [28,29] because it has been shown to be more accurate for fractals with known theoretical dimensions [27]. The sandbox method estimates $D$ by counting the number of data points that lie within a region of radius $R$ centered on a selected data point and measuring how the number of points within the radius changes over a range of radius lengths. Well-defined dimensions that are independent of local behavior are obtained by averaging the results over a number of randomly sampled points on the fractal [27]. $D$ is defined as

$$D=\lim_{R\to 0}\left[-\frac{\log\left[\frac{1}{N}\sum_{i=1}^{N}p_i^{-1}\right]}{\log R}\right], \tag{2}$$

where $R$ is the radius of the circle, $p_i$ is the number of points within the circle divided by the total number of points in the fractal, and $i$ indexes the $N$ circles around the randomly selected points. For our estimate of $D$ for a given sequence, we take the slope of the log-log plot of the sum of the fraction of the data points within radius $R$ centered at each sampled point versus the radius. The critical parameter for the sandbox method is the range of radii [27]. The largest radius should be significantly smaller than the size of the fractal. The smallest radius should be slightly larger than the smallest particle size. In a random walk in four dimensions with each step equal to one unit, the smallest particle size equals $(1+1+1+1)^{1/2}$ or 2 metric units. Within these constraints, we want the range of radii to be as large as possible, but we need to minimize the amount of probable overlap of radius regions. After extensive investigation of the scaling properties of random and pseudorandom walks, we determined that the optimum scaling range was 2–26 for our range of sequence lengths. In our calculations, we used a radius range of 2 to 26 with an increment of 2. Thus, all random walks were evaluated over the same range of scales, a factor of 13.

Two other parameters have an effect on the calculated estimate—the number of points to be sampled and how the sampling is done. Tel, Fulop, and Vicsek sampled about 1% of the points at random locations distributed over the points contained in the fractal [27]. The fractal analyzed in [27] visits each location only once. However, our random walks may visit a site multiple times, so a frequently visited site may be sampled more than once in a random sampling of data points. Thus there are three obvious sampling options: (i) we may allow a frequently visited site to be sampled more than once by randomly sampling data points, (ii) we may sample among unique sites rather than data points to avoid the possibility of examining the same radius region more than once, or (iii) we may sample uniformly every $i$th data point along the random walk. We chose to randomly sample 1% of the data points of each random walk. The probability of examining the same radius region more than once is roughly 2% at this sampling rate. Any duplicate sampling that does occur is a function of the frequency of site visitations and may help characterize the behavior of individual random walks. Thus, random point sampling incorporates density into the estimate of $D$.

## RESULTS

### Effects of sequence length and embedding dimension

We evaluated the effect of sequence length on our estimate of $D$ for the 164 DNA sequences. Although length clearly affects the error in the estimate of $D$ for sequences longer than 4000 bases, we found no correlation ($r^2=0.024$) between the length of a DNA sequence and its estimated $D$. We conclude that our minimum length cutoff of 4500 base pairs was adequate.

We randomly selected 10 of the 164 DNA sequences to evaluate the effect of embedding dimension. A dimer-matched random control was generated for each sequence and the $D$ calculated for each pair, embedded in 2, 4, 6, and 8 dimensions. $D$ increased between 2- and 4-dimensional embeddings but remained relatively constant between 4, 6, and 8 dimensions (Fig. 5). Both DNA and random sequences demonstrated this effect, confirming that it was appropriate to use the same embedding dimension for both types of sequences. Analysis of the 164 human sequences using two-dimensional embedding yielded a mean $D$ of 1.395 and a standard deviation of 0.146. The maximum $D$ was 2.348, so we must use an embedding dimension of at least 4 to satisfy the embed-

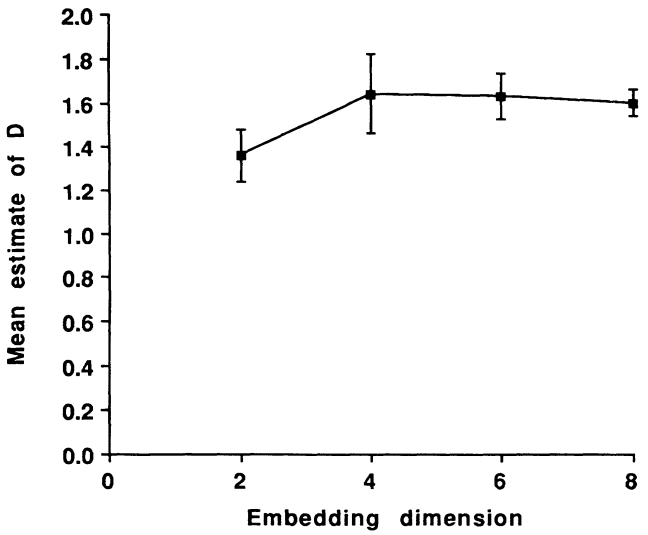


FIG. 5. Fractal dimension vs embedding dimension averaged for 10 DNA sequences. Increase in $D$ from two- to four-dimensional embeddings indicates that two-dimensional embedding is insufficient. The gradual decrease in $D$ for larger embedding dimensions demonstrates the effect of finite sequence length for higher embedding dimensions.

ding dimension rule ($D \geq 1+[2.348]=4$, where the square brackets mean next integer value). A four-dimensional embedding yielded a mean of 1.68 and standard deviation of 0.209. The increase in $D$ from two-dimensional embedding to four-dimensional embedding indicated that the two-dimensional embedding was insufficient and a higher embedding was required. Embeddings greater than 4 did not result in a significant increase in the estimate of $D$, indicating that the four-dimensional embedding was sufficient. The slight decrease in mean $D$ for embedding dimensions higher than 4 was the result of bias due to finite sequence length.

### Estimate of $D$ for DNA sequences and controls

We compared the estimated $D$'s for the 164 human sequences using the four-dimensional embedding scheme to three control types: (i) random, (ii) base-matched, and (iii) dimer-matched. All controls match the paired sequences in length. We generated 30 of each type of control for each DNA sequence to estimate the statistical distribution of $D$ for each. We then calculated a $z$ score for each DNA sequence

$$z=\frac{D_s-\bar{D}_c}{\sigma(D_c)}\ , \tag{3}$$

where $D$ is the estimated $D$ of a sequence, $\bar{D}_c$ is the mean $D$ for its matched controls, and $\sigma(D_c)$ is the standard deviation of the controls. The $z$ score describes the approximate position of $D$ for each DNA sequence within the distribution defined by its controls. To evaluate the group of sequences as a whole, a $t$ test was performed using the mean and standard deviation of the $z$ scores. Thus, $D$ for each DNA sequence is compared to the probability distributions of its controls.

The mean $D$ for our 164 human DNA sequences was 1.631 with a standard deviation of 0.137. The lowest $D$ was 1.300 and the highest $D$ was 2.253. The standard deviation was 8% of the mean $D$. The standard deviation for random controls was 1.8% of the mean $D$, for base-matched controls 7.5% of the mean $D$, and for dimer-matched controls, 6% of the mean $D$. Therefore, at least 75% of the variation found in $D$ for the DNA sequences is due to intrinsic properties of their random walks and not stochastic variation.

We present the aggregate results for all 164 DNA sequences evaluated together in Table I. The mean $D$ was significantly lower than for random controls ($t=-20.813$, $N=164$, $p<10^{-30}$), base-matched controls ($t=-6.111$, $N=164$, $p<10^{-8}$), and dimer-matched controls ($t=-10.280$, $N=164$, $p<10^{-18}$). The histogram in Fig. 6 shows a predominance of negative $z$ scores.

Nonparametric statistics using the rank of $D$ among its matched controls are also revealing. The rank of a sequence is one plus the number of matched control sequences that have a lower estimated $D$. The histogram in Fig. 7 indicates that over 50% of the 164 sequences have a rank below 8 when compared to 30 base-matched or 30 dimer-matched controls. The number of sequences with a $D$ rank of 1 was significant ($p=0.0001$ by a $\chi^2$ test) in both cases with 23% (38/164) ranking lowest among their base-matched controls and 25% of the sequences

TABLE I. Global fractal dimensions for DNA sequences and controls. Both the genomic DNA and mRNA subgroups show significant differences from random, base-matched, and dimer-matched controls.

| | Combined | Genomic DNA | mRNA |
|---|---|---|---|
| $N$ | 164 | 107 | 57 |
| DNA sequences | | | |
| Mean $D$ | 1.631 | 1.641 | 1.613 |
| $\sigma(D)$ | 0.137 | 0.140 | 0.130 |
| Random controls | | | |
| Mean $D$ | 1.863 | 1.865 | 1.859 |
| $\sigma(D)$ | 0.027 | 0.027 | 0.027 |
| Mean $z$ | −2.624 | −2.603 | −2.665 |
| $\sigma(z)$ | 1.615 | 1.665 | 1.530 |
| $t$ value | −20.813 | −16.171 | −13.152 |
| $p$ value | $<10^{-44}$ | $<10^{-29}$ | $<10^{-18}$ |
| Base-matched controls | | | |
| Mean $D$ | 1.702 | 1.709 | 1.690 |
| $\sigma(D)$ | 0.127 | 0.126 | 0.129 |
| Mean $z$ | −0.865 | −0.833 | −0.925 |
| $\sigma(z)$ | 1.812 | 2.004 | 1.397 |
| $t$ value | −6.111 | −4.298 | −5.000 |
| $p$ value | $<10^{-8}$ | $<10^{-4}$ | $<10^{-5}$ |
| Dimer-matched controls | | | |
| mean $D$ | 1.702 | 1.718 | 1.672 |
| $\sigma(D)$ | 0.109 | 0.105 | 0.111 |
| Mean $z$ | −1.068 | −1.193 | −0.834 |
| $\sigma(z)$ | 1.331 | 1.444 | 1.059 |
| $t$ value | −10.280 | −8.543 | −5.950 |
| $p$ value | $<10^{-18}$ | $<10^{-13}$ | $<10^{-7}$ |

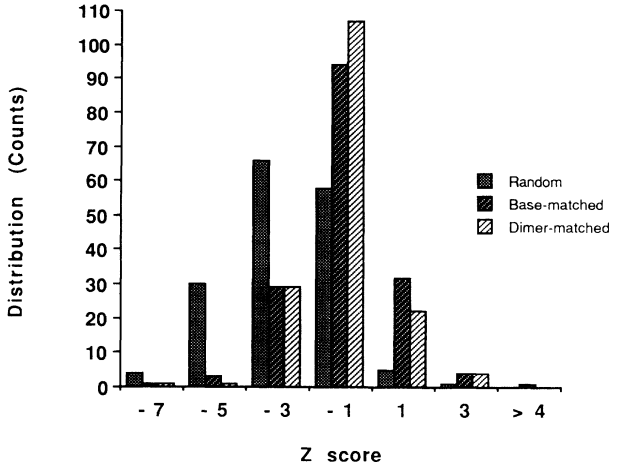


FIG. 6. Distribution of $z$ scores for $D$ of 164 DNA sequences compared to random, base-matched, and dimer-matched controls. The distribution is shifted in the negative direction indicating that $D$ of human DNA is significantly lower than for all controls.

(41/164) ranking lowest among their dimer-matched controls. 75% of the sequences had a rank lower than 16 for both base and dimer-matching.

We expected dimer-matched controls to match the DNA sequences better than base-matched controls. That dimer-matching increased the magnitude of the difference rather than decreasing it requires further explanation. The base-matched and the dimer-matched controls had bases and dimers distributed uniformly within each sequence. The increased differences for dimer-matched controls reflect nonuniform base and dimer distributions within the DNA sequences, with greater differences in dimer distributions than base distributions. We divided each DNA sequence into 500-base subsequences and compared the base and dimer content in these subsequences to the overall base and dimer content. We found wide fluctuations in base and dimer content within sequences. Table II summarizes the results of base distribution analysis. Only 8/164 (4.9%) of the sequences show uniform distribution of all four bases while 85/164 (51.8%) show significant nonuniformity of all four bases. Over 75% of the sequences show significant nonuniformity for each of the four bases. The distribution of dimers within sequences was even more divergent (See Fig. 8). Over 95% of the sequences showed significant nonuniformity of AA or its complement TT, or of CC or its complement GG. Two-thirds of the sequences showed significant nonuniformity of AT, CG, GA-TC, GC, or TA. There is also a symmetry in the frequency of nonuniformities between mirror image dimers. The frequency of CG nonuniformity is approximately equal to the frequency of GC nonuniformity and the frequency of AT nonuniformity is equal to the frequency of TA nonuniformity. AG-CT nonuniformity is as frequent as GA-TC nonuniformity. AC-GT nonuniformity and CA-TG nonuniformity are the least frequent. This symmetry between the nonuniformity of mirror image dimers provides a biological justification for our requirement of reflection

TABLE II. The distribution of bases within sequences. Bases within DNA sequences are not uniformly distributed.

| Base | Number of sequences nonuniform | % |
|---|---|---|
| A | 126 | 76.8 |
| C | 135 | 82.3 |
| G | 131 | 79.9 |
| T | 127 | 77.4 |
| No base | 8 | 4.9 |
| Only one base | 10 | 6.1 |
| Any two bases | 15 | 9.2 |
| Any three bases | 46 | 28.1 |
| All four bases | 85 | 51.8 |

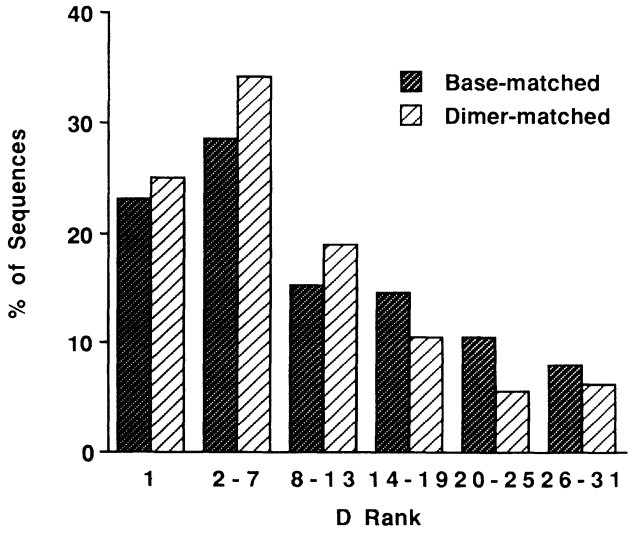


FIG. 7. Distribution of $D$ rankings for 164 human DNA sequences. Rank of $D$ indicates how many matched control sequences had a lower $D$.

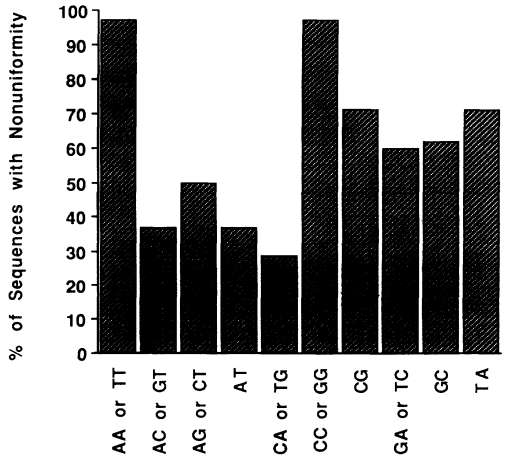


FIG. 8. Distribution of dimers within sequences. Distributions shows marked nonuniformities with symmetries between mirror image dimers.

symmetry (see above) in our embedding scheme.

The 164 nucleic acid sequences in our sample consist of two types: (i) mRNA sequences composed primarily of coding segments but with 5′ and 3′ untranslated segments, and (ii) genomic DNA sequences in which introns and other noncoding segments are predominant. The mean $D$ for the genomic DNA group was $1.641\pm0.14$ and for mRNA $1.613\pm0.13$. However, this difference was not significant ($p=0.20$ by an unpaired $t$ test). Therefore, we compared the $z$ scores of DNA and mRNA groups for random, base-matched, and dimer-matched controls to determine if there was any difference in $D$ between genomic DNA and mRNA. We found that both groups showed significantly lower $D$ estimates than random controls ($p<10^{-29}$ for genomic DNA and $p<10^{-18}$ for mRNA), base-matched controls ($p<10^{-4}$ for genomic DNA and $p<10^{-5}$ for mRNA), and dimer-matched controls ($p<10^{-13}$ for genomic DNA and $p<10^{-7}$ for mRNA). The mean $z$ score for genomic DNA ($-1.193\pm1.444$) was significantly ($p=0.05$) lower than the mean $z$ score for mRNA ($-0.834\pm1.059$).

## DISCUSSION

Matching for base frequencies and even dimer frequencies does not explain the nonrandomness of DNA sequences. The $D$ estimates for the DNA sequences are significantly lower than the $D$'s found for all three types of random controls (length-matched only, base-frequency matched, and dimer-frequency matched), indicating the presence of regions in the pseudorandom walks generated from DNA that are relatively more linear or less clustered than in the controls. It appears that much of the nonrandomness revealed by fractal analysis is due to nonuniform distributions of bases and dimers within sequences. Quasilinear segments may result from single-base runs, dimer runs of GT, CT, GA, or CA, and other oligo-$n$-mers. Runs of CA and other short tandem repeats in mammalian DNA are frequent, as are $n$-mers composed of periodic short runs of T or A that have been associated with nucleosome formation sites [30–33]. This finding correlates with the results of Markov chain analyses [34–38], which found strong nearest-neighbor effects in DNA sequences. There are also families of repetitive elements present in human DNA that often contain internal short repeats.

The $D$'s of sequences composed primarily of noncoding segments (genomic DNA) are different from those composed primarily of coding segments (mRNA). Using an unpaired $t$ test on the $z$ scores, which includes dimer-matched controls, distinguishes the populations at $p=0.05$. Using an unpaired $t$ test directly on the estimates of $D$ fails to distinguish the populations ($p=0.20$). Genomic DNA and mRNA are not totally distinct since their sequences contain both coding and noncoding segments, reducing our power to discriminate between coding and noncoding populations.

This difference in $D$ between genomic DNA and mRNA sequences agrees with the findings of Blaisdell [39] that coding sequences generally contain a significant excess of runs of length 1 or 2 of weak-bonding bases (A or T) and of strong-bonding bases (C or G). Noncoding sequences generally contain a significant excess of long runs of purine (A or G) and pyrimidine (C or T). Long runs produce linear regions in the random walk that decrease $D$. Short repeated sequences should decrease $D$ less than long repeated sequences. We found that both genomic DNA and mRNA have significantly lower estimates of $D$ than all three types of matched controls. However, sequences of genomic DNA have significantly lower dimer-matched $z$ scores than those of mRNA. Thus this difference cannot be due to differences in dimer frequencies [40]. Other methods of DNA sequence analysis (such as Markov chains) have failed to show much difference beyond strong nearest-neighbor influences.

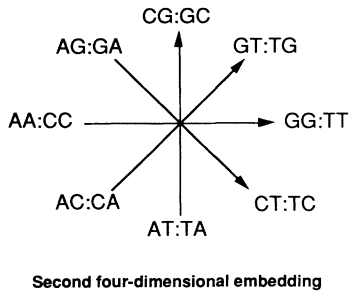


FIG. 9. Second four-dimensional embedding scheme in which all dimers step.

The results we obtained in this study were based on one of many possible axis assignments. Do our results and conclusions change when we use an alternate four-dimensional scheme? To address this issue, we studied a subset of 33 DNA sequences randomly selected from our original set and calculated $D$ using a second alternative embedding scheme in which all dimers step (Fig. 9).

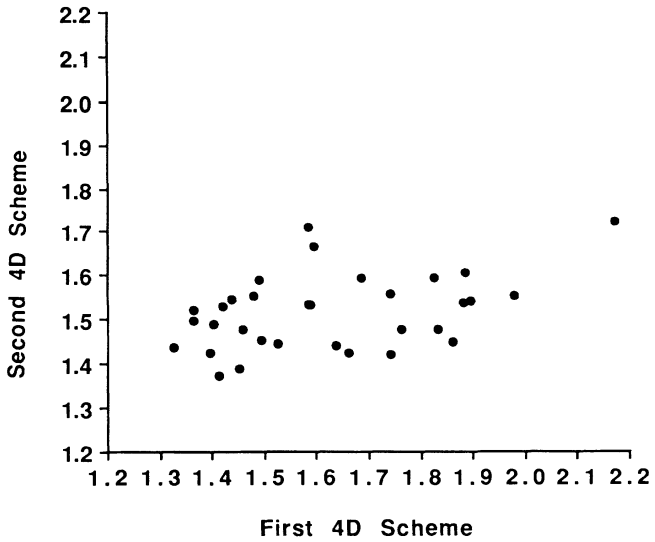


FIG. 10. Scatter plot for a subset of 33 DNA sequences. There is no correlation between estimates of $D$ using the first embedding scheme in Fig. 4 and the second embedding scheme in Fig. 9.

TABLE III. Global fractal dimensions for DNA sequences and controls using a second four-dimensional (4D) embedding. The estimated global fractal dimension for a 33-sequence subset of original data using the embedding scheme in Fig. 9 is significantly different from that obtained using the scheme in Fig. 4. However, it is still significantly lower than $D$ for base-matched controls.

| | First 4D embedding | Second 4D embedding |
|---|---|---|
| $N$ | 164 | 33 |
| DNA sequences | | |
| Mean $D$ | 1.631 | 1.519 |
| $\sigma(D)$ | 0.137 | 0.085 |
| Base-matched controls | | |
| Mean $D$ | 1.702 | 1.580 |
| $\sigma(D)$ | 0.127 | 0.128 |
| Mean $z$ | −0.865 | −1.338 |
| $\sigma(z)$ | 1.812 | 2.171 |
| $t$ value | −6.113 | −3.540 |
| $p$ value | $<10^{-8}$ | 0.0006 |

*Axis 1.* {AA}={CC}=(−1,0,0,0) and {GG}={TT}=(1,0,0,0).

*Axis 2.* {AT}={TA}=(0,−1,0,0) and {CG}={GC}=(0,1,0,0).

*Axis 3.* {AC}={CA}=(0,0,−1,0) and {GT}={TG}=(0,0,1,0).

*Axis 4.* {AG}={GA}=(0,0,0,−1) and {CT}={TC}=(0,0,0,1).

This embedding preserves $D$ for complements, reflections and substitutions, but it is not compatible with our two-dimensional representation. For example, the sequence CC and AA are equivalent in this representation but not in our two-dimensional representation. We found a significant difference ($p=0.005$ by paired $t$ test) in our estimate of $D$ for individual DNA sequences from that obtained using our first embedding scheme and there was no correlation ($r^2=0.159$) between the two estimates of $D$ (Fig. 10). We compared $D$ for DNA to base-matched controls using the second embedding scheme (Table III). The mean global fractal dimension of the DNA sequences (1.519±0.085) was significantly lower than for base-matched controls (1.580±0.128). The mean $D$ was also lower for both DNA and controls using the second embedding scheme and the standard deviation of $D$ for the DNA was smaller (±0.085 versus ±0.137). The mean $z$ score for the second embedding scheme was −1.338 with a $p$ value of $<10^{-3}$ compared to −0.865 with a $p$ value of $<10^{-8}$ for our first scheme. In other words, the second four-dimensional embedding scheme produced the same general result—that the average fractal dimension of DNA is significantly lower than that of base-matched controls (Fig. 11). It is impressive that despite the absence of correlation between the individual $D$ values for DNA sequences in the two schemes, the differences between the DNA sequences and their controls and their ensemble statistical properties are unaffected by the change in representation. While we do not propose that all axis assignments will produce the identical result (some may be more or less discriminating than the two we used), this equivalence is strong evidence that the qualitative differences between random controls and DNA will persist regardless of the embedding scheme.

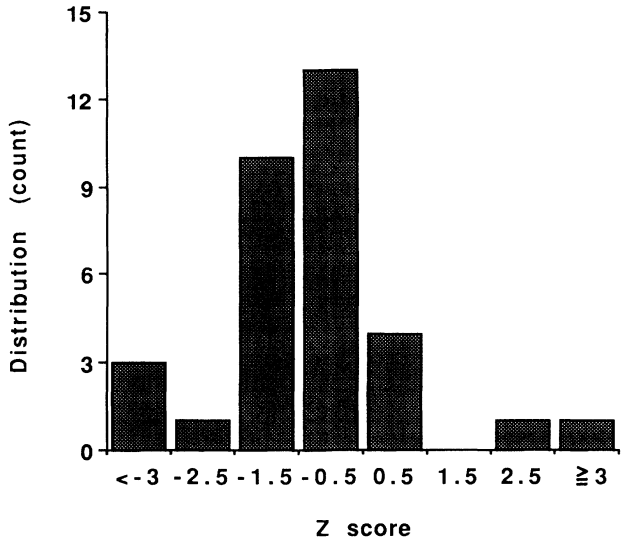


FIG. 11. Distribution of $z$ scores for $D$ of 33 DNA sequences compared to base-matched controls using the second four-dimensional embedding scheme. The distribution is shifted in the negative direction indicating that $D$ of human DNA sequences is significantly lower than controls, agreeing with the results in Fig. 6.

The global fractal dimension of DNA sequences clearly demonstrates that there is significant nonrandom organization within the sequences that is not explained by base or dimer frequencies. Future analyses of $D$ using controls that match for trimer and longer $n$-mer frequencies may be quite revealing, as may investigation of the effects of nonuniform distribution of oligomers within sequences. Investigation of $D$ for DNA of other species and organisms may reveal differences that have not been measurable by other methods of sequence analysis. Finally, studies that evaluate the multifractal spectrum [21,41] of DNA may reveal information of a more localized nature.

*Note added in proof.* Two papers have been published since we submitted this paper involving fractal analysis of DNA sequences [42,43]. Their results appear to support our conclusions.

## ACKNOWLEDGMENTS

We would like to thank David Goldgar for his assistance in the statistical components of this research and Arnold Oliphant for his guidance in molecular biology. This research was supported by NIH Grant Nos. CA-48711 and CA-36362.

*Permanant address: Department of Physics, University of Notre Dame, Notre Dame, IN 46556.

[1] J. A. Glazier and A. Libchaber, IEEE Trans. Circuits Syst. **35**, 790 (1988).
[2] M. A. Gates, J. Theor. Biol. **119**, 319 (1986).
[3] T. G. Dewey and M. M. Datta, Biophys. J. **56**, 415 (1989).
[4] J. S. Helman, A. Coniglio, and C. Tsallis, Phys. Rev. Lett. **53**, 1195 (1984).
[5] Y. Isogai and T. Itoh, J. Phys. Soc. Jpn. **53**, 2162 (1984).
[6] M. Lewis and D. C. Rees, Science **230**, 1163 (1985).
[7] H. J. Stapleton, J. P. Allen, C. P. Flynn, D. G. Stinson, and S. R. Kurtz, Phys. Rev. Lett. **45**, 1456 (1980).
[8] C. X. Wang and Y. Y. Shi, Phys. Rev. A **41**, 7043 (1990).
[9] M. Takahashi, J. Theor. Biol. **141**, 117 (1989).
[10] M. D. Purugganan, Naturwissenschaften **76**, 471 (1989).
[11] For a DNA sequence, the Euclidean distance equals $\{[n(\mathrm{G})-n(\mathrm{C})]^2+[n(\mathrm{T})-n(\mathrm{A})]^2\}^{1/2}$, where $n(\mathrm{A})$, $n(\mathrm{C})$, $n(\mathrm{G})$, and $n(\mathrm{T})$ are the number of each of the bases {A,C,G,T} in the sequence. For a random walk, Euclidean distance equals $[(x_b-x_e)^2+(y_b-y_e)^2]^{1/2}$, where $x_b$ and $y_b$ are the coordinates of the beginning of the random walk (usually zero) and $x_e$ and $y_e$ are the coordinates of the end of the walk.
[12] For a DNA sequence, the Manhattan distance equals $|n(\mathrm{G})-n(\mathrm{C})|+|n(\mathrm{T})-n(\mathrm{A})|$. For a random walk, Manhattan distance equals $|x_b-x_e|+|y_b-y_e|$.
[13] L. Luo and L. Tsai, Chin. Phys. Lett. **5**, 421 (1988).
[14] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).
[15] H. J. Jeffrey, Nucleic Acids Res. **18**, 2163 (1990).
[16] M. Barnsley, *Fractals Everywhere* (Academic, San Diego, CA, 1988), p. 86.
[17] V. V. Solovyev, S. V. Korolev, and H. A. Lim, Int. J. Genome Res. **1**, 108 (1992).
[18] H. S. Greenside, A. Wolf, J. Swift, and T. Pignataro, Phys. Rev. A **25**, 3453 (1982).
[19] F. Takens, in *Dynamical Systems and Turbulence,* edited by D. A. Rand and L. S. Young, Lecture Notes in Mathematics Vol. 898 (Springer-Verlag, Berlin, 1981).
[20] J. B. Ramsey and H.-J. Yuan, Phys. Lett. A **134**, 287 (1989).
[21] The correlation dimension is a scaling exponent related to the global fractal dimension. H. G. E. Hentschel and I. Procaccia, Physica D **8**, 435 (1983).
[22] L. A. Smith, Phys. Lett. A **133**, 283 (1988).
[23] C. L. Berthelsen and J. A. Glazier (unpublished).
[24] B. B. Mandelbrot, *The Fractal Geometry of Nature* (Freeman, New York, 1983), p. 232.
[25] J. Rudnick and G. Gaspari, Science **237**, 384 (1987).
[26] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing* (Cambridge University Press, Cambridge, England, 1988), p. 209.
[27] T. Tel, A. Fulop, and T. Vicsek, Physica A **59**, 155 (1989).
[28] P. Grassberger and I. Procaccia, Phys. Rev. Lett. **50**, 346 (1983).
[29] P. Grassberger, Phys. Lett. A **148**, 63 (1990).
[30] T. Kimura, T. Takeya, and M. Takanami, Biochim. Biophys. Acta. **1007**, 318 (1989).
[31] S. Pennings, S. Muyldermans, G. Meersseman, and L. Wyns, J. Mol. Biol. **207**, 183 (1989).
[32] T. E. Shrader and D. M. Crothers, Proc. Natl. Acad. Sci. U.S.A. **86**, 7418 (1989).
[33] E. C. Uberbacher, J. M. Harp, and G. J. Bunick, J. Biomol. Struct. Dyn. **6**, 105 (1988).
[34] H. Almagor, J. Theor. Biol. **104**, 633 (1983).
[35] B. E. Blaisdell, J. Mol. Evol. **21**, 278 (1985).
[36] P. W. Garden, J. Theor. Biol. **82**, 679 (1980).
[37] J. Kleffe and U. Langbecker, Comput. Appl. Biosci. **6**, 347 (1990).
[38] S. Tavare and B. W. Giddings, in *Mathematical Methods for DNA Sequences,* edited by Michael S. Waterman (CRC, Boca Raton, 1989), p. 117.
[39] B. E. Blaisdell, J. Mol. Evol. **19**, 122 (1983).
[40] R. A. Elton, J. Mol. Evol. **4**, 323 (1975).
[41] T. C. Halsey, M. H. Jensen, L. P. Kadanoff, I. Procaccia, and B. I. Shraiman, Phys. Rev. A **33**, 1141 (1986).
[42] E. C. Uberbacher and R. J. Mural, Proc. Natl. Acad. Sci. U.S.A. **88**, 11 261 (1991).
[43] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Nature **356**, 168 (1992).

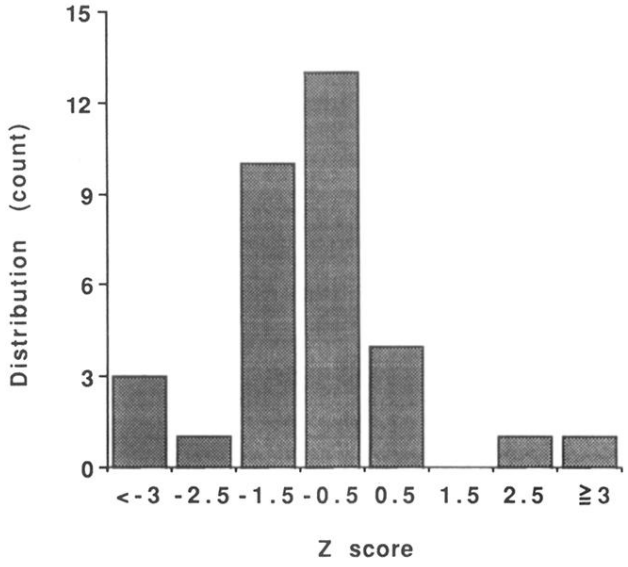


FIG. 11. Distribution of $z$ scores for $D$ of 33 DNA sequences compared to base-matched controls using the second four-dimensional embedding scheme. The distribution is shifted in the negative direction indicating that $D$ of human DNA sequences is significantly lower than controls, agreeing with the results in Fig. 6.

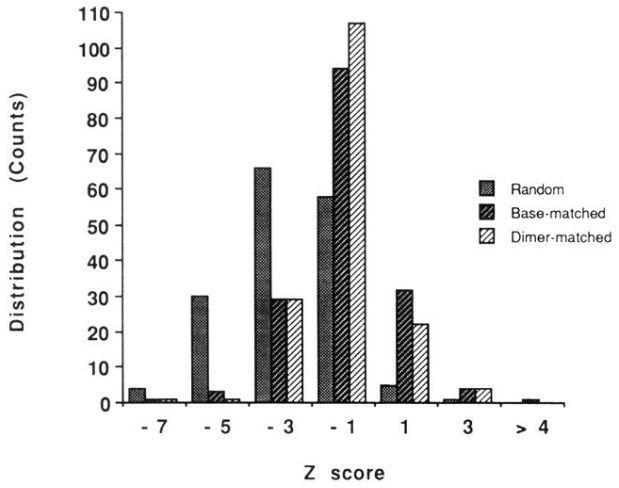


FIG. 6. Distribution of $z$ scores for $D$ of 164 DNA sequences compared to random, base-matched, and dimer-matched controls. The distribution is shifted in the negative direction indicating that $D$ of human DNA is significantly lower than for all controls.