Effective multifractal spectrum of a random walk

Cheryl L. Berthelsen,[1] James A. Glazier,[2,*] and S. Raghavachari[2]
[1]*University of Mississippi Medical Center, School of Health Related Professions, Department of Health Information Management, 2500 North State Street, Jackson, Mississippi 39216*
[2]*Department of Physics, University of Notre Dame, Notre Dame, Indiana 46556*
(Received 8 November 1993)

An infinite uniform random walk on a lattice is a monofractal, but any finite-length random walk contains both bulk and surface, which cannot be easily distinguished, but which have intrinsically different fractal dimensions. When we examine both the bulk and the surface using the sandbox method, we obtain a reproducible effective multifractal spectrum that is independent of the length of the random walk. This technique allows us to calculate the multifractal spectrum for short (~10 000 base pairs) DNA-sequence data sets. We show that human β globin DNA is significantly different from its length-, base-frequency-, and dimer-frequency-matched controls.

PACS number(s): 05.45.+b, 02.70.−c

INTRODUCTION

The analysis of DNA-sequence data requires techniques that allow reliable calculation of the properties of data sequences as short as a few thousand base pairs. Earlier work has shown that the global fractal dimension and related global measures can provide useful but limited information concerning DNA sequences [1–3]. More detailed measures of the distance between DNA sequences are needed to distinguish between introns (regions that code for proteins) and exons (noncoding regions) [4,5] or to reconstruct phylogenetic trees for cladistic analysis. However, global calculations neglect that DNA sequences are highly inhomogeneous.

Multifractal formalism is a useful way to characterize the spatial inhomogeneity of both theoretical and experimental fractal patterns [6–9]. The best known multifractal is the exactly soluble two-scale Cantor set, which is representative of a large class of pointlike deterministic attractors. In many cases, however, fractals are stochastic and higher dimensional. The simplest example is the random walk on a lattice, which finds many applications in polymer physics, biology, and economics. Calculation of multifractal spectra for random walks requires longer data sets than are available from experiments. However, for many purposes we are interested less in the exact values of the multifractal spectrum than in comparing a data set to a model which generates a set of matched controls. In this case, the consistency of the analysis between data sets is important rather than the degree of convergence for a single data set and it is desirable to use a scaling radius that includes both the bulk and the surface of the walk so as to maximize the discrimination of the method.

In this paper we show that, even for short random walks, we can obtain useful information from the effective length-independent multifractal spectrum. Using this technique, we show that we can distinguish an individual DNA sequence from a variety of models which generate matched controls.

METHOD

We calculate multifractal spectra using the sandbox method of Tel, Fulop, and Vicsek, which converges substantially faster than box-counting for the two-scale Cantor set [9–11]. The multifractal spectrum by the sandbox algorithm [10] is defined to be

$$D_q = \lim_{R\to 0} \frac{1}{q-1} \frac{\ln\left[\frac{1}{N}\sum_{i=1}^{N} p_i^{q-1}\right]}{\ln R} , \tag{1}$$

where the square brackets represent an average over a number of randomly sampled points on the fractal, p_i the number of points within a circle of radius R around the ith sampled point divided by the total number of points in the fractal, and N the total number of points sampled on the fractal. For an ideal fractal

$$D_q = \lim_{R\to 0} D_q(R) . \tag{2}$$

In practice, we perform a linear fit on $D_q(R)$ over a range of R, $[R_{\min}, R_{\max}]$, with $R_{\min}$ representing the size of a lattice point and $R_{\max}$ chosen to maximize the linearity of the fit. We discuss the choice of $R_{\max}$ below. An effective fractal dimension exists if D_q is monotonic, nonincreasing within error, and insensitive to small changes in the scaling range.

Alternatively we may characterize the fractal by a local scaling exponent α. The dimension of the subset of points that have the given value of α, $f(\alpha)$, gives all the scaling properties of the set [7]. Since scaling and dimension are equivalent concepts, the $f(\alpha)$ spectrum and the

*Corresponding author.

D_q curves [or more correctly $(q-1)D_q$] are related. $f(\alpha)$ is the Legendre transform of $(q-1)D_q$ given by

$$\alpha=\frac{d\tau(q)}{dq}\ , \tag{3}$$

$$f(\alpha)=\alpha q-\tau(q)\ . \tag{4}$$

Positive q values correspond to the low-α side of the $f(\alpha)$ curve and the negative q values correspond to the high-α part of the curve. In practice, we can first calculate the D_q spectrum and obtain the $f(\alpha)$ curve by means of the above relations.

SIERPINSKI CARPET

The edges of a fractal, which are usually ill-defined, affect the scaling exponents characterizing the distribution of points. Thus our measurements of the multifractal spectrum depend on the fractal's size or extent. To understand the effect of edges on the D_q of pseudorandom walks, we computed D_q for the square Sierpinski carpet. The ideal, infinitely subdivided carpet is a monofractal ($D_q=\ln 8/\ln 3=1.893$). A finite-sized carpet with a finite number of subdivisions will have an effective multifractal spectrum due to the inclusion of the edges of the carpet.

Figure 1 shows the results for a 150×150 Sierpinski carpet generated by the iterated function system method of Barnsley [12]. We computed the D_q curves using the sandbox method with the sandbox centers distributed randomly over the carpet. The spectrum thus produced is monotonic, nonincreasing for all values of $R_{\max}$ used. For most centers, circles of radius $R_{\max}$ avoid the edge of the fractal completely, resulting in a regime of constant scaling up to some value of R. As $R_{\max}$ increases, at each center, the sandbox will cross an edge of the fractal at a typical radius R_{crit}. That part of the sandbox lying outside the fractal will not contribute additional points, so

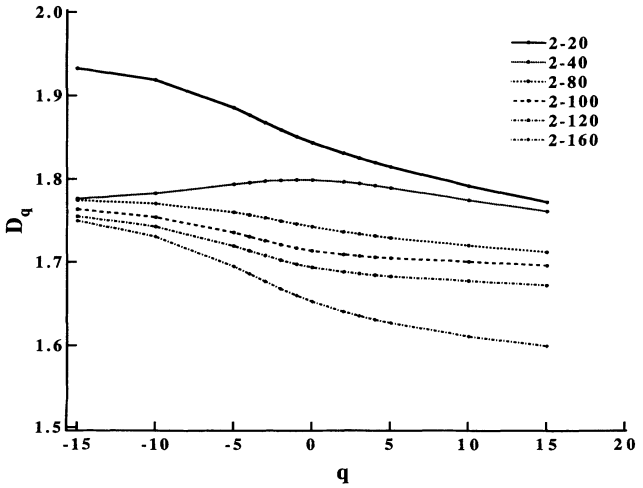

FIG. 1. Multifractal spectra for a Sierpinski carpet of size 150×150. Monotonic nonincreasing spectra occur for all $R_{\max}$ (within error). The D_q values decrease for any given q as $R_{\max}$ increases.

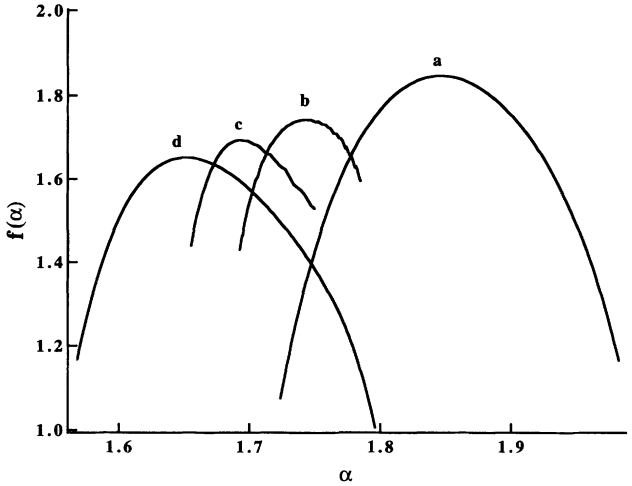

FIG. 2. $f(\alpha)$ curves for a Sierpinski carpet of size 150×150 for $R_{\max}$ equal to (a) 20, (b) 80, (c) 120, and (d) 160. The $f(\alpha)$ curves shift to lower α values with increasing $R_{\max}$, resulting in the lowering of the D_q curves.

the rate of increase of points with the radius will drop. Thus for $R>R_{\text{crit}}$, the effective scaling is less than that of the bulk. Thus as $R_{\max}$ increases, the effective α decreases. Therefore the computed $f(\alpha)$ curve shifts to the left.

We observed that initially an increase in $R_{\max}$ reduced the high-α part of the curve more than the low-α part (Fig. 2), resulting in a flatter D_q curve shifted to lower q values. Further increase in $R_{\max}$ reduced both the high- and low-α parts of the curve almost equally, shifting the D_q curve lower, but not flattening it any further. For $R_{\max}$ greater than the size of the fractal (150), the low-α part of the $f(\alpha)$ curve was decreased more than the high-α part, broadening the D_q curve. Any additional increase in $R_{\max}$ resulted in both the high- and low-α parts of the curve being reduced almost equally. We can now apply these ideas to calculate the multifractal spectra for random walks.

RANDOM WALKS

A random walk is defined as the sequence $\{\mathbf{y}_i\}$, where

$$\mathbf{y}_i=\sum_{j=1}^{i}\mathbf{x}_j \tag{5}$$

and $\mathbf{x}_j$ is either $(\pm1,0,0,0)$, $(0,\pm1,0,0)$, $(0,0,\pm1,0)$, or $(0,0,0,\pm1)$. We chose the direction of each step randomly with uniform probability. We can define quantities that characterize the spatial extent of random walks such as the average span

$$P=\frac{1}{d}\sum_{i=1}^{d}(i_{\max}-i_{\min})\ , \tag{6}$$

where d is the embedding dimension, $i_{\max}$ and $i_{\min}$ are the most distant points, maximum and minimum values, visited on the ith axis. The mean-square displacement of

a walk after n steps is

$$\langle R_n^2 \rangle = nb^2 , \tag{7}$$

where b is a constant.

A true random walk of infinite length is space filling in two dimensions, i.e., $D=2$. Higher-dimensional embeddings also result in fractal dimension [13] $D=2$. Local fluctuations can result in regions with fractal dimensions greater than 2. Since we must work in a space that is at least one dimension higher than the dimension of the structure [14], we used a four-dimensional embedding for our calculations.

Sampling rates above 2% in the sandbox method resulted in D_q values differing only by about 1–2%. Therefore we sampled 2% of the points in the random walk in each case. For standard box-counting techniques the D_q curve does not converge for any of the parameter ranges discussed [15,16].

We calculated D_q curves for ten random walks each of length 50 000, 100 000, and 250 000 over a limited scaling range [2,40]. We expected the resulting curves to be relatively flat since the sampling was local. The D_q curves were approximately monofractal with the average $D_q = 2 \pm 0.1$ for all q. The standard deviations were smallest at $q=0$ for all lengths and largest at the extremes, $q=-15$ and $q=+15$, with all standard deviations less than 5% [9,11]. We did not observe a monofractal behavior for the 10 000-step walk for the same scaling range. The shortest walk to yield a monofractal spectrum was 4000 steps with standard deviations of about 10%, using a scaling range [2,4] with all line fits based on two-point fits between radii of 2 and 4. Thus for short walks the region of monofractality is extremely small and hard to determine in a consistent manner. In addition, to characterize the global structure of an unknown object, we must look for correlations at all length scales, which precludes such a small scaling range.

A larger sampling region inevitably includes the edge of the walk. An infinite-length random walk will have edges at infinity which have no effect on its fractal dimension. For finite-length walks, however, there is no way to totally avoid the edge while computing D_q. Though there has been some work on adjusting measures of fractal dimensions for edge effects [17], these adjustments apply to graphs of continuous functions when using the box-counting method to compute D_q and not for finite unit-step random walks. Unlike the Sierpinski carpet, the random walk does not have regions within the fractal which have distinct edges. However, random walks typically have holes within the bulk of the walk of varying shapes and sizes. These spaces are difficult to predict, though we expect the same flattening and lowering of the D_q curve. Thus the only way to calculate the multifractal spectrum of a random walk is to randomly sample points over the walk using a scaling range that can be applied consistently.

We generated 30 random walks each of various lengths and calculated their mean-square displacements and average spans. To avoid sequential correlations while choosing the steps of the walk we used a random-number generator based on a subtractive method [18]. The average span and the mean-square displacement both scaled by the square root of the walk length, with $b=0.868n^{1/2}$ and $P=0.764n^{1/2}$, differing only by a constant with average span slightly smaller than the mean-square displacement. For real fractals like DNA, which is nonuniform in extent along different axes, the mean-square displacement is an indicator of the maximum extent while the average span is more representative of the true scaling range. This difference would not be an issue under conditions of ideal scaling, since D_q is invariant under conformal transformations. For effective dimensions we must pick a scaling range in a systematic manner. The average span is less sensitive to the aspect ratio than the maximum radius.

Figure 3 reveals that there is some change in the average multifractal spectrum over a range of $R_{\max}$ values for the 50 000-step random walks with sampling radii set to $0.8R$, R, and $1.2R$, where R is the average span of the walk along the axes. The average D_q values for $0.8R$ and $1.2R$ differ from the D_q values for R by only about ± 0.04 for all q values used (i.e., a 20% change in the scaling range gives only a 2% change in D_q).

The edge effect on the multifractal spectrum of the random walk is similar to the effect on the spectrum of the Sierpinski carpet. As expected from the Sierpinski carpet results, increasing $R_{\max}$ decreased the D_q curve. An increase in $R_{\max}$ from $0.8R$ to R caused a greater shift in the positive q values of the D_q curve. An increase from R to $1.2R$ had the same effect but to a lesser degree, consistent with the behavior of the D_q curves of the Sierpinski carpet. However, since the $R_{\max}$ used were rather close to the average span, the amount of edge included was not significantly different at each scaling range. Therefore small changes in the scaling range around the

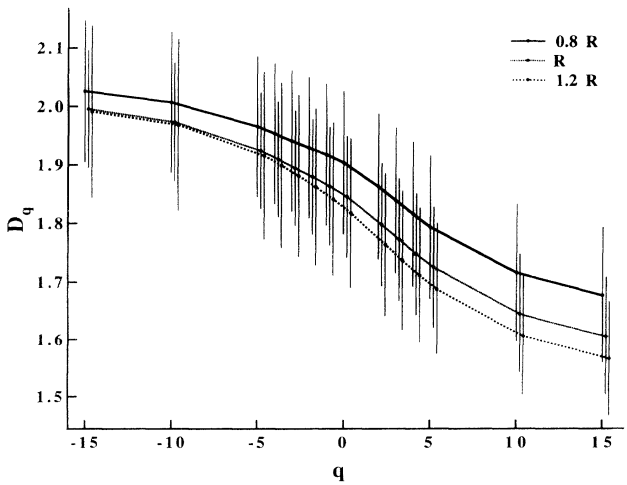

FIG. 3. The average D_q curves for 20 random walks of 50 000 steps with $R_{\max}$ set to $0.8R$, R, and $1.2R$, where R is the average span of the walk. The D_q values decreases for a given q as $R_{\max}$ increases in a manner similar to the D_q curves for the Sierpinski carpet due to the inclusion of the edges of the walk. Error bars offset for clarity.

average span had only a small effect on the spectrum, so we could consistently analyze the walks.

When characterizing the information content of DNA sequences we rarely have long sequences of DNA available. Typical DNA sequences are 5000–50 000 bases in length. Therefore the measures used to distinguish DNA from its controls should be independent of the length of the DNA sequence. We calculated D_q spectra for 20 random walks each of lengths 10 000, 50 000, 100 000, and 250 000 with the maximum scaling radius R_{max} set to the average span of each walk (Fig. 4). The spectra clearly exhibited multifractal behavior. The standard deviations decreased with the length of the walk (from 8% for the 10 000-step walks to 3.5% for the 250 000-step walks). The D_q curves were essentially similar and the differences between them were well within one standard deviation. We did a two-dimensional Kolmogorov-Smirnov (KS) test [19] comparing the D_q curves for 20 random walks each of length 10 000, 50 000, and 100 000 steps to the D_q curves for the 250 000-step walk. The two-dimensional KS test allows us to estimate whether distributions characterized by two variables differ. Small values of the KS significance indicate that the distributions are different. The significance levels for the 100 000-, 50 000-, and the 10 000-step walks were 0.778, 0.542, and 0.438, respectively, indicating that the distributions were not significantly different. Thus the multifractal spectrum calculated by setting R_{max} as the average span does yield consistent results from walk to walk and is essentially independent of the length of the walk.

Thus the average span of the walk along its axes provides a consistent scaling range that can be applied to calculate an effective, length-independent multifractal spectrum, even for short walks. We can use these spectra to compare different sequences or sequences to their controls.

MULTIFRACTAL SPECTRUM OF DNA

We mapped DNA into four-dimensional pseudorandom walks according to the mapping procedure and embedding scheme discussed in Ref. [1]. We then computed the multifractal spectra of these walks using the sandbox algorithm with R_{max} set equal to the average span of the walk. We employed three control sequences for comparison: (1) random, where each base has an equal probability of occurring; (2) base-matched, where the control sequences are generated according to the proportion of bases in the DNA sequence; and (3) dimer-matched, where the control sequences are generated using the probability of each dimer pair occurring according to the proportion of dimers in the DNA. The multifractal spectra of these control sequences were calculated. Figure 5 shows a clear difference between the human β globin gene and all three control sequences. We compared the length-, base-, and dimer-matched controls with the actual sequence using the two-dimensional KS test. The tests yielded significance levels on the order of $\sim 10^{-5}$ indicating a clear distinction between the DNA and its controls, showing the existence of long-range correlations in the DNA than is present in random control sequences. The base-matched controls had a higher KS significance level (10^{-5}) than the dimer-matched controls (10^{-7}), consistent with the result in Ref. [1]. We report a cladistic study of the species relations of mitochondrial DNA using these techniques elsewhere [20].

CONCLUSION

A finite unit-step random walk on a lattice is monofractal over a severely limited scaling range (a maximum radius of 10% of the average span on the axes). Larger scaling ranges yield an effective multifractal spectrum that characterizes the perimeter of the walk in addition to the bulk. The sandbox algorithm gives converged

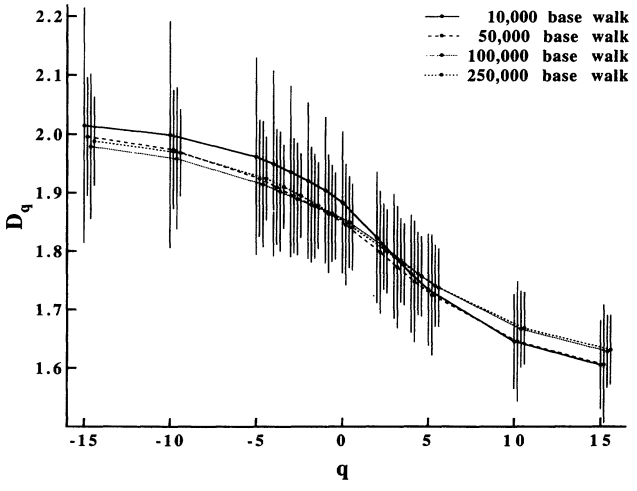

FIG. 4. The average D_q curves for 20 random walks each of length 10 000, 50 000, 100 000, and 250 000 over a scaling range equal to the average span of the walk are essentially similar and independent of the length of the walk. Error bars offset for clarity.

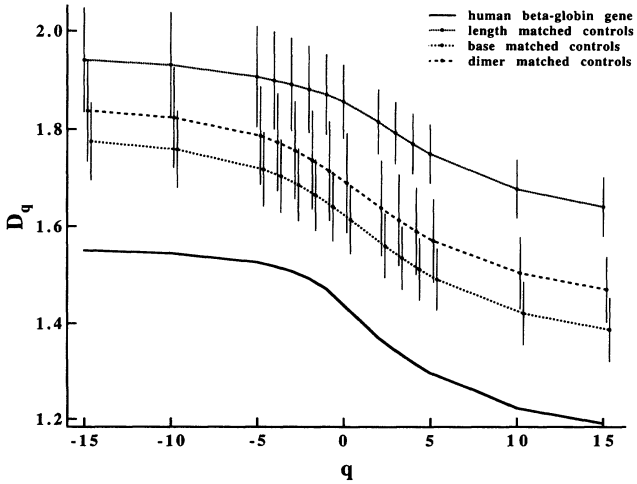

FIG. 5. The multifractal spectrum of the human β globin gene (length 73 326 base paris). Significant differences exist between its D_q curve and the average curves of the random, base-matched, and dimer-matched control sequences.

spectra for much shorter random walks than does box counting. The converged effective multifractal spectrum allows the analysis of short experimental data sets such as DNA sequences, and the characterization of DNA sequences in a length-independent, robust manner. With this method we can distinguish the human β globin DNA from length-, base-, and dimer-matched controls.

ACKNOWLEDGMENTS

We would like to thank M. Skolnick for continuing support of this project. This research was partly supported by J.S.P.S., Monbusho, NSF Grants Nos. DMR 92-57011 and INT 91-01345, Ford Motor Company, and the Exxon Educational Foundation.

[1] C. L. Berthelsen, J. A. Glazier, and M. H. Skolnick, Phys. Rev. A **45**, 8902 (1992).
[2] C. K. Peng, S. V. Buldryev, A. L. Goldberger, S. Halvin, F. Sciortino, M. Simons, and H. E. Stanley, Nature **356**, 168 (1992).
[3] P. B. DePretillo (unpublished).
[4] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).
[5] W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992).
[6] P. Grassberger and I. Procaccia, Phys. Rev. Lett. **50**, 346 (1983).
[7] T. C. Halsey, M. H. Jansen, L. P. Kadanoff, I. Procaccia, and B. I. Shraiman, Phys. Rev. A **33**, 1141 (1986).
[8] H. G. E. Hentschel and I. Procaccia, Physica D **8**, 435 (1983).
[9] H. Gould and J. Tobochnik, Comput. Phys. **4**, 202 (1990).
[10] T. Tel, A. Fulop, and T. Vicsek, Physica A **59**, 155 (1989).
[11] T. Vicsek, F. Family, and P. Meakin, Europhys. Lett. **12**, 217 (1990).
[12] M. F. Barnsley, *Fractals Everywhere* (Academic, San Diego, CA, 1988), p. 91.
[13] J. Rudnick and G. Gaspari, Science **237**, 384 (1987).
[14] F. Takens, in *Dynamical Systems and Turbulence,* edited by D. A. Rand and L. S. Young (Springer, Berlin, 1981).
[15] J. B. Ramsey and H.-J. Yuan, Phys. Lett. A **134**, 287 (1989).
[16] L. A. Smith, Phys. Lett. A **133**, 283 (1988).
[17] C. C. Taylor and S. J. Taylor, J. R. Stat. Soc. B **53**, 353 (1991).
[18] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in Fortran: The Art of Scientific Computing* (Cambridge University Press, Cambridge, 1988), p. 273.
[19] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in Fortran: The Art of Scientific Computing* (Ref. [18]), p. 640.
[20] C. L. Berthelsen, J. A. Glazier, and S. Raghavachari (unpublished).