FRACTAL ANALYSIS OF DNA SEQUENCE DATA

by

Cheryl Lynn Berthelsen

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Medical Informatics

The University of Utah

March 1993

ABSTRACT

DNA sequence databases are growing at an almost exponential rate. New analysis methods are needed to extract knowledge about the organization of nucleotides from this vast amount of data. Fractal analysis is a new scientific paradigm that has been used successfully in many domains including the biological and physical sciences. Biological growth is a nonlinear dynamic process and some have suggested that to consider fractal geometry as a biological design principle may be most productive.

This research is an exploratory study of the application of fractal analysis to DNA sequence data. A simple random fractal, the random walk, is used to represent DNA sequences. The fractal dimension of these walks is then estimated using the "sandbox method." Analysis of 164 human DNA sequences compared to three types of control sequences (random, base-content matched, and dimer-content matched) reveals that long-range correlations are present in DNA that are not explained by base or dimer frequencies. The study also revealed that the fractal dimension of coding sequences was significantly lower than sequences that were primarily noncoding, indicating the presence of longer-range correlations in functional sequences.

The multifractal spectrum is used to analyze fractals that are heterogeneous and have a different fractal dimension for subsets with different scalings. The multifractal spectrum of the random walks of twelve mitochondrial genome sequences was estimated. Eight vertebrate mtDNA sequences had uniformly lower spectra values than did four invertebrate mtDNA sequences. Thus, vertebrate mitochondria show significantly longer-range correlations than do invertebrate mitochondria. The higher multifractal spectra values for invertebrate mitochondria suggest a more random organization of the sequences.

This research also includes considerable theoretical work on the effects of finite size, embedding dimension, and scaling ranges.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

ACKNOWLEDGMENTS

As with any major endeavor in life, many individuals contribute to a successful outcome. I hereby wish to acknowledge those who played a critical role in this project.

Mark Skolnick, my dissertation committee chair and mentor: You dedicated much time and support to this effort for which I am most grateful. You provided adequate equipment and space to accomplish this research. You have been a great scientific role model and patient mentor.

James Glazier, my fractal consultant and research collaborator: Thank you for providing the vital information and education in chaos theory. I could not have done this research without your tender guidance.

Arnold Oliphant, my molecular biology consultant and committee member: You helped me keep the biology of DNA in focus and provided opportunities to expand my perspective of fractal dimension.

David Goldgar, my statistical consultant and committee member: Your assistance in developing ways to analyze the statistical significance of my fractal dimensions was very helpful.

Simon Tavare, my mathematics consultant and committee member: Amid the theories of fractals and biology, you brought a mathematical perspective that was enlightening.

Reed M. Gardner, my other committee member: Your participation on my committee and helpful comments on the manuscript were very helpful.

I also wish to thank the entire Medical Informatics faculty for the exceptional education they provided. In particular I wish to acknowledge Stan Huff who supervised the first year and a half of my research in the Medical Informatics program. Our association provided invaluable experience in data structures and knowledge representation.

I also acknowledge the Graduate Research Committee of the University of Utah who provided grant support to complete this research.

Chapter 3 was published as "Global fractal dimension of human DNA sequences treated as pseudorandom walks," *PHYSICAL REVIEW A*, Vol. 45, no. 12, pp. 8902-8913. Permission was granted to include it in this dissertation.

Chapter 4 has been submitted for publication in *Physical Review A.* Permission was granted to include it in this dissertation.

1. INTRODUCTION

Deoxyribonucleic acid (DNA), the genetic blueprint of most living organisms, was discovered only 40 years ago by James D. Watson and F. H. C. Crick. Just 20 years ago, F. Sanger and W. Gilbert independently developed techniques to discover the nucleotide sequence that defines a living organism. Since then, DNA has been studied with unprecedented fervor, leading to important discoveries in genetics, molecular biology, and evolution. The Human Genome Project has the goal of sequencing the entire human genome, and the United States Congress has budgeted $200 million to this endeavor. The task of discovering the sequence of nucleotides that defines man is fairly simple, but to understand the meaning of the sequence of bases is a task of tremendous complexity. The discovery of abnormal sequences that cause many human diseases will revolutionize the practice of medicine. Of equal importance are the discoveries that will explain how four simple nucleotides control the growth, development, and differentiation of organisms. Can the language and grammar of life be elucidated from the patterns found in DNA? This is the goal of DNA sequence analysis.

1.1 The Structure of DNA

DNA is composed of two strands that bond by a specific base-pair bonding 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 subsequent bases 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 base-pair bonding rule:

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

The subsequence 5'->AACTGG->3' on one strand is complementary to 5'->CCAGTT->3' on the opposite strand. In general, DNA with both strands bound together does not have direction and neither strand has precedence.

It has been estimated that less than 5% of DNA actually codes for protein. The purpose and function of the remaining 95% are presently unknown.

When DNA is copied to make messenger RNA (mRNA), the two strands are distinct. One strand is the sense strand and the other is the coding strand. An mRNA sequence is synthesized from 5' to 3' using the sense strand template read from 3' to 5'. A protein is made from mRNA by reading it in a 3' to 5' manner, grouping the bases into triplets called codons that code the specific amino acids to be used to construct the protein.

To summarize, DNA is double-stranded and has no direction. The two strands in noncoding DNA are complementary and of unknown biological function; mRNA, on the other hand, is single-stranded, and its base sequence is biologically distinct from its reverse complement. The mRNA nucleotide sequence defines a specific sequence of amino acids composing a protein. Nucleotides are translated in groups of three (codons) starting with a specific base, defining the three-base frame. If translation begins with a different nucleotide (a frame-shift), a different protein is produced. If the mRNA was translated in the reverse direction, it would also result in a different protein. Therefore, both frame position and direction are important for translation of mRNA sequences into proteins.

It is well known that the genetic code is redundant and that some amino acids are represented by more than one codon. In fact, only 20 different amino acids are represented by the 64 possible codons. The four bases composing the DNA of an organism do not occur uniformly. Table 1.1 shows the base compositions of selected gene and genome sequences that demonstrates this nonuniformity. It is also interesting that synonymous codons are not used equally, and some are rarely used. Some of this codon preference or bias is explained by base and two-base dimer frequencies, but there appear to be other influences such as speed and accuracy of translation (Dix and Thompson 1989).

1.2 DNA Sequence Analysis Methodologies

The goal of DNA sequence analysis is to discover important patterns in the organization of DNA. (I distinguish these analysis methods from homology search, alignment algorithms, and

Table 1.1

Base composition of selected sequences shows nonuniform frequency patterns (Weir 1985).

Sequence	No. of bases	A	C	G	T
human mitochondrion	16,569	31%	31%	13%	25%
hepatitis B virus	3,182	28%	22%	27%	23%
human fetal globin exons	882	24%	25%	28%	22%
human fetal globin introns	1,996	27%	17%	27%	29%

sequence comparisons.) These patterns are then used to predict structure and function. For example, codon preferences have been used to determine whether a sequence codes for protein by analyzing the different framing possibilities. Important information about patterns identifying mRNA splice sites has been deduced from analysis of sequences. Finding an open reading frame is important in specifying whether a sequence codes for a protein.

Sequence analysis has also been used to identify the position of a nucleosome along a sequence and to predict the secondary folding structure of a protein. Other important patterns identify mutation patterns and regions of DNA that have a high probability of being polymorphic. These discoveries help explain how proteins function and how genes are expressed, help identify and map genes through the use of polymorphic genetic markers, and allow identification of mutations in defective proteins that cause diseases. DNA sequence analysis is very important for future discoveries and greater understanding of this molecule of heredity.

1.2.1 Oligonucleotide frequencies

Some of the first DNA sequence analyses involved frequency measures of short oligonucleotides, particularly dinucleotides. Oligonucleotide frequencies allow researchers to estimate the frequencies of longer oligos in a genome and the frequency of restriction enzyme cutting sites (Bishop, Williamson, and Skolnick 1983). The molecular weight of a strand of DNA and the melting point or hybridization temperature may also be estimated using oligonucleotide frequencies.

Nussinov (1981) found consistent asymmetries in the frequency patterns of dinucleotides in prokaryotes and eukaryotes--AT was more frequent than TA, CT more frequent than TC, TG more frequent than GT, and GC more frequent than CG. After analyzing 44 sequences comprising more than 70,000 bases, she concluded that there was "irrefutable evidence for the existence of nearest neighbor preferences and asymmetries in DNA sequences" (Nussinov 1980, p. 4545).

Later, Nussinov (1984a) found distinct patterns in DNA dimer frequencies. Nonvertebrate sequences showed more uniform dimer distributions than vertebrate sequences. Eukaryotic sequences showed almost equal frequencies of complementary dimers. The occurrence of CG in eukaryotes was rare but was common in prokaryotes. Bacterial phages that were primarily single-stranded genomes had different frequencies for complementary dimers. Eukaryotic viral sequences tended to show the same dimer preferences as eukaryotic organism sequences. In general, dinucleotide frequencies were different for evolutionarily distinct groups. In a follow-up paper she showed that doublet preferences were primarily due to conformation constraints of the DNA double helix (Nussinov 1984b). In eukaryotes, purine-purine and pyrimidine-pyrimidine dimers are preferred over purine-pyrimidine and pyrimidine-purine dimers. The tight packaging of DNA in nucleosomes disfavors these in eukaryotes primarily due to steric repulsion of opposite-chain nearest-neighbor purine clashes.

Blaisdell (1983) evaluated the occurrence of base runs in coding and noncoding DNA. He found that coding sequences generally contained a significant excess of dimers and trimers composed of A or T (AA, AT, TA, TT, AAA, AAT, ATA, ATT, TAA, TAT, TTA, TTT) and of C or G (CC, CG, GC, GG, CCC, CCG, CGC, CGG, GCC, GCG, GGC, GGG). Noncoding regions, in contrast, generally contained a significant deficit of dimers or trimers of purines (AA, AG, GA, GG, AAA, AAG, AGA, AGG, GAA, GAG, GGA, GGG) and pyrimidines (CC, CT, TC, TT, CCC, CCT, CTC, CTT, TCC, TCT, TTC, TTT) and a significant excess of long runs of purines and pyrimidines. The differences between coding and noncoding DNA composition were significant enough to distinguish them, and the runs results were not explained by nearest-neighbor preferences.

Volinia and colleagues (Volinia et al. 1989) studied the frequencies of k-tuples (oligonucleotide of length k) in mammalian DNA up to length 6. Some of the patterns found were that the frequencies of the 2-tuples, AA, AG, AT, GT, TA, and TT, were lower than expected in coding DNA and higher than expected for CA, CC, CG GC, and TG. CG is suppressed in noncoding DNA and TA is suppressed in coding DNA. In 3-tuples, there is a high incidence of triplets containing only A and T in noncoding DNA and low incidence in exons. (This is opposite of Blaisdell's findings.) This same pattern was also present for the 3-tuples, AGT, TAG, TTG, TTC, TGT, TCT, GTA, GTT, and GGG. The 3-tuples, NCG and CGN (where N is any base) occur more frequently in exons than noncoding regions. Noteworthy among 4-tuples was that TATA occurred almost 10 times more frequently in noncoding DNA than coding DNA. These occurrences were independent from the TATA box that occurs upstream from an open reading frame of a sequence encoding a protein. 5-tuples composed of A and G combinations were more frequent in exons than noncoding DNA. TATA-containing 6-tuples were rare and AAAAAA and TTTTTT were six times more frequent than expected on base composition in DNA examined. In general, k-tuples containing A and T, except for TATA, are prevalent in noncoding DNA and k-tuples of G and C are more frequent in coding DNA.

Hong (1990) used the results of nearest-neighbor analyses to predict the frequencies of hexanucleotides in DNA. He concluded that hexanucleotide frequency predictions based on dinucleotide frequencies alone were as satisfactory as those by third-order Markov chains (See discussion under Markov chains) that require tetranucleotide and trinucleotide frequencies as raw data.

1.2.2 Entropy measures

DNA is not just a random sequence of bases. The fact that it provides the genetic blueprint for a living organism implies that there is inherent order or nonrandomness in the sequences of bases. Information theory techniques attempt to quantify the amount of information contained in a sequence. Shannon's information entropy measure (1948) reaches its maximum for completely random sequence and is lower for sequences with nonuniform base frequencies. Gatlin (1972) refers to Shannon's entropy as "information potential" and introduced a measure that quantifies "stored information." (See Equation 2.10 for a formal definition of entropy.)

One of the first studies of information content in DNA was calculated from viral DNA sequences (Rowe 1983). Viral genes show significant information content only on levels of single bases, dimers, and trimers with little or no information content for quadruplets and quintuplets. The noncoding regions in the viral genomes studied showed some order on all levels tested. These results suggest that noncoding regions have more order than coding regions that cannot be explained on a functional basis. Genes coding for structural proteins tended to show stronger triplet correlations than other genes, an effect primarily due to codon preference or bias.

Information content from an evolutionary view was evaluated by Subba Rao and colleagues (Subba Rao, Hamid, and Subba Rao 1979; Subba Rao, Geevan, and Subba Rao 1982). They concluded that information content of DNA tends to increase with evolution. However, Konopka (1984) tested Rao's hypothesis with an entropy measure that adjusted for codon preference and found numerous contradictory examples. Entropy measurements for globin genes, which are fast-evolvers, were actually lower than those for histone genes, which are slow-evolvers. Entropy measurements for human nuclear genes were lower than human mitochondrial genes even though mitochondrial genomes are known to mutate seven times faster. He concluded that entropy was not a good indicator of evolutionary differences.

Entropy measures have been used to identify sets of homogeneous nucleotide sequences that define a class of DNA sequences (Ragosta et al. 1992). Genes from *Escherichia coli* were classified according to level of expressivity using a homogeneity index based on entropy. After identifying and eliminating genes that appeared to be a transitional element between classes, 67% of the genes were correctly classified using the homogeneity index.

A recent paper measures entropy in coding and noncoding regions of *Escherichia coli* genome DNA (Lauc, Ilic, and Heffer-Lauc 1992). The calculated entropies were not sufficiently different to distinguish coding from noncoding DNA on a nucleotide level. However, the entropies for nucleotides of reading frames were significantly lower than those of frame-shifted sequences, indicating the presence of information content for functional nucleotide sequences. Unexpectedly, protein sequences showed significantly ***higher*** entropies than sequences translated into protein using frame-shifted sequences.

1.2.3 Markov chains

A number of papers have been published on Markov chain analysis of DNA sequences. Markov chain methods attempt to evaluate the predictability of sequence occurrences based on the frequencies of shorter sequences. They are also useful in evaluating differences in the statistically expected frequencies with actual observed frequencies to discover unexpected patterns in DNA.

A process has the Markov property if the outcomes of the preceding *n* states are all that are required to predict the outcome of the next state (Kemeny and Snell 1976). A zero-order Markov chain defines an independent or random process where knowledge of previous outcomes does not have any predictive value for the next state. A Markov process is said be an *n*-order process if the previous *n* states are required to predict the next state. If a DNA sequence is viewed as a Markov chain of first-order, the preceding base in the sequence is all that is required to predict the next base in the sequence. Likewise, a second-order chain requires the preceding two bases to predict the next and a third-order chain requires the preceding three bases.

One of the first papers using Markov chains to analyze DNA sequences (Elton 1975) reported a study of dimer frequencies. Interesting results were found for the dimer CG in vertebrate DNA. He found that the extreme shortage of CG in bulk vertebrate DNA does not occur in ribosomal and transfer RNA genes. The Markov chain model assumes that the probabilistic structure is the same regardless of the position in the sequence being considered.

Interestingly, the sequences he analyzed from *Escherichia coli* did not show significant heterogeneity whereas viral sequences did. Also, significant doublet heterogeneity was found between translated sequences in bacteriophage data and untranslated sequences.

Garden (1980) analyzed viral DNA and RNA sequence using Markov chains. He found that a third-order model fit the ϕX174 virus, a second-order model fit the early and late regions of the SV40 virus, and a zero-order model fit the replicase gene of MS2. In general, the short-range order for MS2 contrasted with the longer-range order for SV40, and ϕX174 showed the lack of a common layout of genetic information for all species.

Fuchs (1980) found a disconcerting result in Markov chain analyses. The selected order of the Markov model tended to increase with the length of the sequence analyzed. The majority of nucleotides of length 500 were well-fitted by a order zero or one model as expected for short sequences. This result brought into question the conclusions on the appropriate order of the models from previous DNA analyses.

Blaisdell (1985) again found strong nearest-neighbor influence by Markov chain analysis. Most of the sequences he analyzed required at least a second-order Markov chain for their representation, and some required chains of third-order. This was true for both coding and noncoding sequences.

Phillips et al. analyzed the *Escherichia coli* genome using Markov chains. He studied the accuracy of Markov chains in predicting the frequencies of nucleotide sequences of length one to eight bases. Immediate neighbors could be predicted with high accuracy. A third-order Markov chain was relatively accurate in predicting most pentanucleotide frequencies. In general, a third-order Markov chain was a reasonable predictor of nucleotide frequencies up to eight bases in length. Interestingly, the coding strand of the genome was enriched for oligomers in high abundance sequences, and the noncoding strand was enriched for low abundance sequences.

Almagor (1983) used Markov chains to study two viral DNA sequences. In general, first-order correlations determined the triplet frequencies and the correlations between nearest neighbors seem to be the primary influence in nonrandomness found.

Tavare (Tavare and Song 1989) analyzed coding regions using Markov chains. In general, homogeneous Markov chains were not adequate to describe codon preference and amino acid usage. A simple spatially heterogeneous Markov model reflected these coding features more accurately.

1.2.4 Signal processing methods

Signal processing techniques search for recurrent periodicities in DNA sequences and have been used to find homologies between nucleotide or amino acid sequences.

Veljkovic (Veljkovic et al. 1985) represented DNA sequences as numeric values representing the potential electron-ion interaction value of each nucleotide or amino acid. This representation of DNA was then analyzed using Discrete Fourier

Transform, a common tool in signal processing, to extract any information present corresponding to biological function. They evaluated over 200 different protein sequences and found consistent results. Functionally related sequences exhibited significant frequency peaks, whereas unrelated sequences did not show these peaks. Different peaks were present for different biological functions. The importance of this result is the ability to detect functional relatedness among proteins without significant similarity in the underlying DNA sequences.

Silverman and Linsker (1986) used a tetrahedron representation for DNA sequences and a Fourier transform to detect periodicities. Their base-independent representation of DNA successfully identified tandem repeats of TG in the human somatostatin I gene, a GGCGGCGGC repeat of length 320 flanking the T24 human bladder carcinoma oncogene, and several five- and six-base repeats involving either three or four contiguous guanine bases. Another periodicity found in the T24 oncogene was a recurrence of guanine in the third codon position of a number of contiguous codons.

Benson (1990) introduced some novel variations of the classical Fourier transform for DNA sequence analysis. These enhancements enable the detection of clusters of matching bases, facilitate the insertion of gaps to enhance sequence similarity, and accommodate varying densities of bases in the input sequences.

Signal processing techniques have been used to predict nucleosome formation sites. Satchwell (Satchwell, Drew, and

Travers 1986) found a strong 10.2 base-pair periodicity for the dimers AA and TT and a phased 10 base-pair periodicity of the dimer GC about 5.1 bases away from AA/TT pattern. These periodicities correlate with one turn of the double helix with the AA/TT pattern forming the minor groove facing inward and the GC pattern facing outward. Uberbacher (Uberbacher, Harp, and Bunick 1988) using Fourier transform detected a strong 10 base-pair repeat pattern for the dinucleotides AA (or TT) as well as 21, 6.4, and 7.1 base-pair periodicity that correlate with phases in rotational conformation for major and minor grooves of the DNA helix.

1.2.5 Heterogeneity detection

Staden (1984) characterized heterogeneity by scanning the sequence with a fixed-size window and computing summary statistics of local composition. This involves an arbitrary choice of window size but provides a powerful tool to show local properties graphically. However, it does not provide a method of determining significant departures from homogeneity.

An ad hoc test for heterogeneity divides a sequence into k segments of equal length. Then a chi-square test is applied to test for differences in the proportions of single bases or dimers. This approach uses an arbitrary number of segments and window sizes. Although it provides a test statistic for overall heterogeneity, it fails to identify regions of difference and does not estimate local properties. Churchill (1989) developed a stochastic technique to evaluate heterogeneous DNA sequences using a state-space model as a hidden Markov chain. It finds local departures from homogeneity, estimates change points, and uses likelihood estimates to find the best fitting state-space model.

None of the above analysis methods adequately describe or model DNA sequence patterns. Markov chains show only low-order dependence. Signal processing techniques detect recurrent periodicities. Heterogeneities show transitions of base or dimer content across arbitrary-sized windows. Oligonucleotide frequency counts show departures from uniformity or expected patterns based on base content. However, none have shown sufficient power to differentiate coding from noncoding sequences in anonymous DNA. Dimer contents are different for evolutionarily distinct organisms but cannot be used to differentiate organisms per se. Entropy measures have not differentiated evolutionary groups and are not sufficiently different to distinguish coding and noncoding regions. All these methods seem to find fairly localized properties successfully but have not found large-scale correlations that intuitively should be present in DNA. Other tools are needed to extract long-range information in DNA sequences.

1.3 Fractal Analysis, a New Paradigm

Chaos theory is in an embryonic state of development yet is sufficiently developed to use as a scientific tool. Thomas S. Kuhn, in his classic, The Structure of Scientific Revolutions, defines a paradigm as an achievement:

> . . . was sufficiently unprecedented to attract an enduring group of adherents away from competing modes of scientific activity. Simultaneously, it was sufficiently open-ended to leave all sorts of problems for the redefined groups of practitioners to resolve. (1970, p. 10)

Physicists were the first to apply chaos theory and have used it successfully to model a variety of problems including turbulence and dynamic systems that were difficult by other methodologies. The number of publications now available involving fractal analysis in virtually all scientific disciplines attests to the fact that fractal analysis has reached the state of development to be called a paradigm. The discussions and disagreements among scientists (Pool 1990) confirm that much is still unknown in the applicability and methodology of chaos theory. However, this fact has not discouraged them from experimenting with it:

> Chaos has opened new horizons in science and it is already considered by many the third most important discovery in the twentieth century, after relativity and quantum mechanics. Philosophically speaking, chaos has brought some pessimism since it imposes limits on prediction. At the same time, however, it has offered a new forum for the understanding and description of irregularity, complexity and unpredictability in Nature. (Tsonis and Tsonis 1989, p. 31)

Kuhn noted an important fact about new paradigms in scientific research:

> To be accepted as a paradigm, a theory must seem better than its competitors, but it need not, and in fact never does, explain all the facts with which it can be confronted. (1970, pp. 17-18)

Mandelbrot, the father of fractal geometry, notes:

> While the diversity of nature appears to be without bound, the number of techniques one can use to grasp nature is extremely small and increases very rarely. Therefore, the enthusiasm usually generated by the birth of a new technique and the desire to test it more widely is healthy, and must not be disparaged. (1989, p. 11)

1.4 What Is Fractal Analysis?

The word 'fractal' was coined by Mandelbrot (1977) from the Latin *fractus*, meaning broken, to describe objects that were too irregular to fit into a traditional geometrical setting. He defines a fractal as follows:

> Broadly speaking, mathematical and natural fractals are shapes whose roughness and fragmentation neither tend to vanish, nor fluctuate up and down, but remain essentially unchanged as one zooms in continually and examination is refined. Hence, the structure of every piece holds the key to the whole structure. (Mandelbrot 1989, p. 4)

Falconer (1990) suggested that a fractal be defined in the manner that biologists define "life"--using a list of characteristics that are usually present but exceptions exist. He lists five characteristics of a set F that define a fractal:

1. F has a fine structure, i.e., detail on arbitrarily small scales.

2. F is too irregular to be described in traditional geometrical language, both locally and globally.

3. Often F has some form of self-similarity, perhaps approximate or statistical.

4. Usually, the 'fractal dimension' of F (defined in some way) is greater than its topological dimension.

5. In most cases of interest F is defined in a very simple way, perhaps recursively.

Fractals may possess certain inherent characteristics. A self-similar fractal may be decomposed into smaller, reduced copies of the whole. The reductions are linear and identical for all directions. The Sierpinski triangle gasket in Figure 1.1 is a standard strictly self-similar fractal. A self-affine fractal is a self-similar object in which the reductions are still linear but the reduction ratios in different directions are different.

The primary task of fractal analysis is to estimate the fractal dimension of an object, which describes its spatial complexity. In general terms, it describes how the mass of the object changes with scale. (The term fractal dimension is formally defined in Chapter 2.) In Euclidean geometry, dimensions take integer values. A point has a dimension of 0, a line has a dimension of 1, a surface such as a square has a dimension of 2, and a solid sphere or cube has a dimension of 3. In fractal geometry, dimensions are fractional and describe how much the space is filled by a fractal. The Sierpinski gasket in Figure 1.1 is a partially filled triangle and has a fractal dimension of 1.58, which is less than 2 (the dimension of a solid square) but greater than 1 (the dimension of a line).

It is important to note that the estimate of fractal dimension depends on the scale used to measure the object. If a

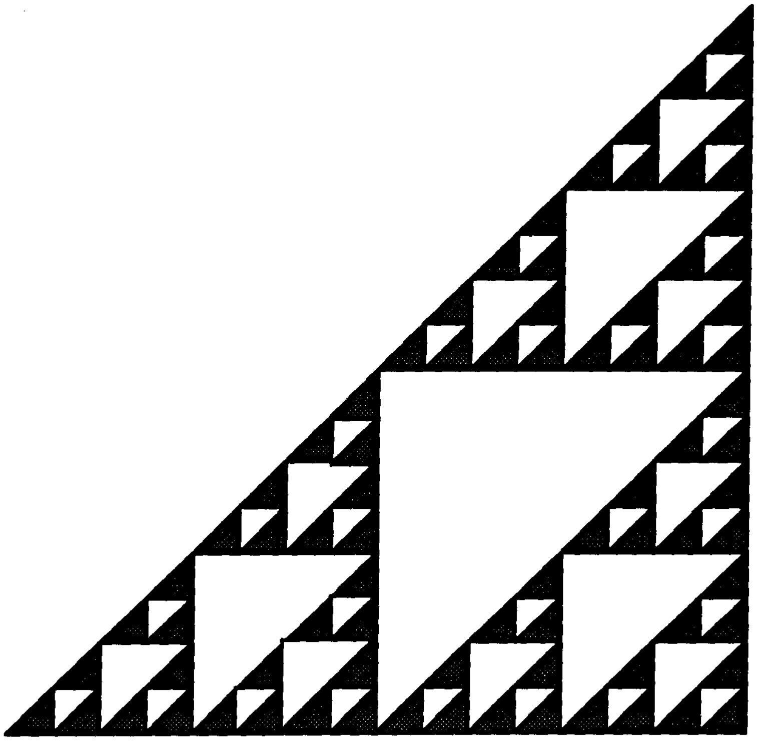


Figure 1.1 Sierpinski triangle gasket, a partially filled triangle, has a fractal dimension of 1.58.

pine tree is viewed from a great distance, it has a dimension between 1 and 2--it looks like more than just a line but does not quite fill a surface. If the pine tree is viewed from a closer perspective, it has a dimension between 2 and 3--it does not totally fill the three-dimensional space but is clearly more than a planar object with a dimension of 2. If the pine tree is viewed extremely close up so all that is distinguishable is the end of one pine needle, it has a dimension of 0. (It looks like a point.) If viewed less closely, the needle would look like a line with a dimension of 1. If the pine tree is viewed just inches from its trunk, it has a dimension of 2. (It looks like a surface.) Therefore, depending on the perspective, a pine tree may have a dimension of 0, 1, between 1 and 2, 2, or between 2 and 3. It should be obvious that none of these dimension estimates are necessarily ***the*** fractal dimension of a pine tree. They are estimates of fractal dimensionality at a specific scale.

Complex fractals that do not cover the embedding space uniformly may not be adequately described by a single fractal dimension. A spectrum of fractal dimensions that emphasize different scaling densities may be necessary. Multifractals have a different fractal dimension for subsets of points with different scalings. The multifractal spectrum may provide important information about an object that is lost or averaged out when a single fractal dimension is used to describe it. Two fractal objects may have the same fractal dimension at one scale but have very different multifractal spectrums. Therefore, an object's multifractal spectrum may be an important distinguishing property.

1.5 Why Fractal Analysis of DNA Sequences?

One of the intriguing characteristics of fractals is the fact that very simple mathematical equations can generate infinitely complex and beautiful patterns. The Julia and Mandelbrot sets show ever-changing variety of color and intricate detail with recurring shapes and patterns as iterations of the generating equations progress. Mandelbrot (1989) describes this process:

> The process of iteration effectively builds up an increasingly complicated transform, whose effects the mind can follow less and less easily. Eventually, one reaches something that is 'qualitatively' different from the original building block. (p. 6)

DNA behaves similarly. The almost infinite variety of living organisms on the earth is based on just four simple nucleotides. The DNA sequence of an organism defines it completely through every stage of its existence. The vast differences between organisms are traced to differences in their DNA sequences.

It has been proposed that the DNA molecule has evolved by multiple transformations that have been iteratively repeated over millions of years. There are consistent, recurring patterns in DNA that are universal, reflecting this possible origin as well as a certain element of seeming randomness. Ohno (1988) made an interesting observation that the so-called codon preference is just a mere reflection of the the construction principle of coding sequences. He showed how two coding sequences that demonstrated classic codon preference were derived from a repeating, mutating heptameric unit that defined the sequence rather than any selection process. Nussinov (1984a) noted that the symmetries present in complementary dimer frequencies probably reflect early evolutionary events such as simple and inverted duplication. This repetitive and iterative process of mutation, transformation, and replication results in an end product that is vastly different from the four nucleotide building blocks--a living, functional organism.

The observation that growth in biology is a nonlinear dynamic process and living tissues demonstrate fractal properties led one scientist to conclude:

> To consider fractal geometry as a biological design principle is heuristically most [*sic*] productive and pro-vides insights into possibilities of efficient genetic programming of biological form (Weibel 1991, p. L361).

It is the goal of this research to use fractal analysis as an exploratory tool. Other DNA sequence analysis methods have successfully revealed patterns and structure of a localized nature. It is important to remember that an estimate of fractal dimension depends on the scale used to measure it and the way the data is represented. The fractal dimensions estimated in this research are not ***the*** fractal dimensions of the sequences in an absolute sense; rather it is a heuristic measure of complexity and the departure from randomness of a sequence. What can be discovered about DNA using this new paradigm? Does it reveal patterns in DNA not readily discoverable by other methods? What long-range correlations in DNA are revealed by fractal analysis of sequence data?

2. METHODS OF ESTIMATING FRACTAL DIMENSION

2.1 Introduction

The most common fractal dimension is the "Hausdorff dimension" or "global fractal dimension." This is a single real value that characterizes how the density of an object varies with length scale. It is generally adequate for very simple objects. More complicated objects can be described by a continuous spectrum of fractal dimensions called "multifractal spectrum of generalized fractal dimensions." The multifractal spectrum may be helpful in distinguishing two objects that are inherently different but have identical global fractal dimensions.

A number of different methods of estimating a fractal dimension have been developed. Some are applicable only to a random walk representation, whereas others can be applied to any time series representation. Some can be used to find the multifractal spectrum, whereas others will only calculate the global fractal dimension.

In this chapter, the various algorithms are presented and experimentally applied to fractal analysis of DNA sequence data represented as pseudorandom walk. The random walk is defined in Equation 2.1

Equation 2.1 A random walk where $\{\vec{x}_j\}$ is either $(\pm1,0,0,0)$, $(0,\pm1,0,0)$, $(0,0,\pm1,0)$ or $(0,0,0,\pm1)$ chosen at random with equal probability.

$$\vec{y}_i \equiv \sum_{j=1}^{i} \vec{x}_j$$

(See section 3.2.2 for method of mapping DNA sequence into a random walk.) The estimates of fractal dimension for DNA sequences are calculated using four different algorithms and the results are compared. A conclusion is made about the appropriateness of each of these algorithms and the approach of choice for this research.

2.1.1 Gate's Manhattan distance

An early paper (Gates 1986) on ways to represent DNA sequences data graphically also presented one of the first methods of estimating fractal dimension. Gates suggested a fractal dimension measure for large scale structure in random walks (Equation 2.2). It is the logarithm of the number of bases in the sequence divided by the logarithm of the Manhattan distance of the endpoint from the origin. Another method he proposed was the logarithm of the number of bases in the sequence divided by the logarithm of the Euclidean distance of the endpoint from the origin. (Note: For a DNA sequence using the two-dimensional embedding scheme, the Euclidean distance equals $\sqrt{[\ n(G)\text{-}n(C)\ ]^2 + [\ n(T)\text{-}n(A)\ ]^2}$ where n(A), n(C), n(G), and n(T) are number of each of the bases {A,C,G,T} in the sequence. The

Manhattan distance equals $|n(G) - n(C)| + |n(T) - n(A)|$.) Manhattan and Euclidean distances are formally defined in Equations 2.3 and 2.4.

Equation 2.2 Gate's global fractal dimension based on the Manhattan distance between the endpoints of a random walk. N is the number of bases and $|\vec{x}|$ is the Manhattan distance defined in Equation 2.3.

$$D = \frac{\log N}{\log |\vec{x}|}$$

Equation 2.3 The Manhattan distance of the endpoint of a random walk from the origin. D_E is the embedding dimension.

$$||\vec{x} - \vec{y}|| = \sum_{i=1}^{D_E} | x_i - y_i |$$

Equation 2.4 The Euclidean distance of the endpoint of a random walk from the origin.

$$||\vec{x} - \vec{y}|| = \sqrt{\sum_{i=1}^{D_E} | x_i - y_i |^2}$$

2.1.2 Asphericity measure

Random walks are known to be asymmetrical in the distance travelled along each axis. Rudnick and Gaspari developed a method of estimating fractal dimension that quantifies the amount of asymmetry or asphericity of the trail left by a random walker (Rudnick and Gaspari 1987). Although this approach may produce useful results, my research did not apply this technique to DNA sequences.

2.1.3 Maximum radial distance

Another way to estimate fractal dimension uses the maximum distance the walk travels from the origin (Equation 2.5).

Equation 2.5 The global fractal dimension of a random walk based on the maximum radial distance from the origin. R is the maximum radial distance travelled from the origin and N is the number bases required to reach that distance.

$$D = \frac{\log N}{\log R}$$

2.1.4 Grassberger-Procaccia algorithm

The Grassberger-Procaccia algorithm (GPA) uses box-counting (Grassberger and Procaccia 1983). The number of uniformly sized boxes required to cover the trajectory of the sequence is determined for a range of box sizes. The exponent relating the increase in box size to the decrease in number of the boxes required is the fractal dimension (Equation 2.6).

Equation 2.6 The global fractal dimension by the Grassberger-Proccacia algorithm (GPA). N is the number of boxes required and L is the box size.

$$D = \lim_{L \to \infty} - \frac{\log N}{\log L}$$

GPA presupposes an optimal covering using the fewest possible non-overlapping boxes. However, finding an optimal covering is computationally intractable, so box counting is done using non-overlapping boxes aligned on a grid as a practical approximation. The steps in the random walk are exactly one unit in length so the coordinates are always integer values. To avoid ambiguity of box assignment for points falling on box boundaries, the grid is shifted by 0.5. This guarantees that every point in the graph belongs to one and only one box in the grid. The choice of box size range has a critical effect on the estimate of the fractal dimension. The most linear region of the log/log plot is generally used to select the range. My experiments with fractals of known theoretical fractal dimensions (data not given) indicate that good approximations are produced using a minimum box size of 2.5% and maximum box size of 30% of the maximum span of the fractal. Maximum span is defined as the maximum difference between the most negative and most positive coordinate of each axis.

2.1.5 Tel's sandbox algorithm

Tel's sandbox algorithm estimates the global fractal dimension by counting how many data points are within a region of radius R centered on a selected data point and measuring how the mass changes over a range of radius lengths (Equation 2.7).

Equation 2.7 The global fractal dimension by Tel's sandbox algorithm. R is the radius, p_i is the number of points within the circle around radius R divided by the total number of points in the fractal, and i indexes the N circles around the randomly selected points.

$$D_0 = \lim_{R \to 0} - \frac{\log \left[\frac{1}{N} \sum_{i=1}^{N} p_i^{-1} \right]}{\log R}$$

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 (Tel, Fulop, and Vicsek 1989).

The critical parameter for the sandbox algorithm is the choice of radius lengths. The largest radius length 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 this equals a step of one unit. However, the range of radii needs to be as large as possible. There are two other parameters that have an effect on the calculated estimates--the number of points to be sampled and the how the sampling is done.

Tel sampled about 1% of the points randomly distributed over the fractal (Tel, Fulop, and Vicsek 1989). However, the multifractal he analyzed visits a coordinate only once whereas the random walks may visit a site multiple times and a frequently visited site may be sampled more than once using random sampling of data points. Sites could be sampled rather than data points to avoid the possibility of examining the same radius region more than once. Another sampling option uniformly samples every ith data point along the walk.

2.1.6 Estimating generalized fractal dimensions

The Hausdorff dimension is only one of an infinite number of different generalized dimensions that characterize a fractal (Hentschel and Procaccia 1983). D_q provides information about the fractal at different levels of density in the random walk. The generalized dimension spectrum is defined in Equation 2.8.

Equation 2.8 Generalized dimension spectrum by the Grassberger-Proccacia algorithm. L is box size, b is the number of boxes, and p_i is the number of points in the ith box divided by the total number of points. q is varied over a range of negative and positive values.

$$D_q = \lim_{L \to 0} \frac{1}{q-1} \frac{\log \left[\sum_{i=1}^{b} p_i^q \right]}{\log L} \qquad q \neq 1$$

The theoretical D_q curve of a multifractal is smooth and monotonically decreasing (Halsey et al. 1986). Within the spectrum of D_q are several commonly used fractal dimensions: the Hausdorff dimension, $D_{q=0}$; the information dimension $D_{q=1}$; and the correlation dimension, $D_{q=2}$ (Halsey et al. 1986). Thus, the information dimension ($D_{q=1}$) is less than the Hausdorff dimension ($D_{q=0}$) but greater than the correlation dimension ($D_{q=2}$).

A simple modification of GPA allows calculation of D_q. The number of points contained in each box is accumulated and then divided by the total number of points. The range of box sizes is again an important parameter in D_q estimates. This range is usually specific to the fractal and found by experimentation.

The box counting method of calculating D_q curves has a serious drawback. The calculation of D_q for negative q is a well-documented problem yielding spectra that may be completely irrelevant (Hakansson and Russberg 1990). The boxes are not necessarily centered on the data points and the resolution is necessarily finite so some boxes will contain very few data points resulting in a disproportionate contribution when raised to a negative power (Hakansson and Russberg 1990). The sandbox algorithm has been successful in estimating the fractal spectrum to within 3% of the theoretical value for a known multifractal for q as low as -8 (Tel, Fulop, and Vicsek 1989).

There have been a number of modifications made to GPA to make the algorithm more efficient for calculating the multifractal spectrum. The original box-counting algorithm for a sequence of length N requires a computation time of order N^2. Block (Block, von Bloh, and Schellnhuber 1990) found a way to reduce the CPU time to order N $\log_2$ N, which requires less computer memory to do the computation. Grassberger (1990) also came out with an optimized box-counting algorithm. Dvorak and Klaschka (1990) modified GPA to handle high embedding dimensions.

The sandbox algorithm for estimating the generalized fractal dimensions overcomes the problem of nonoptimal box-covering or noncentered boxes over the points. It produces linear log/log plots that are free from the oscillations that occur in GPA for negative q (Tel, Fulop, and Vicsek 1989). The sandbox estimate of D_q is defined in Equation 2.9.

Equation 2.9 Generalized dimension spectrum by Tel's sandbox algorithm. R is radius length and i indexes the N circles around the randomly selected points.

$$D_q = \lim_{R \to 0} \frac{1}{q-1} \frac{\log \left[\frac{1}{N} \sum_{i=1}^{N} p_i^{(q-1)} \right]}{\log R} \qquad q \neq 1$$

2.2 Methods

An important rule in fractal theory involves the choice of embedding dimension to represent the data. The embedding dimension must be at least as high as the highest possible fractal dimension rounded up to next whole number plus one (Greenside et al. 1982; Takens 1981). Methods of estimating the global fractal dimension are known to be biased. The bias in fractal dimension estimates increase with embedding dimension (Ramsey and Yuan 1989). Higher embedding dimensions require longer sequences. Therefore, it is important to reduce the effect of bias by minimizing the embedding dimension. A positive side effect of this goal is a corresponding decrease in computation time.

The choice of DNA sequences of adequate length is also important. It has been suggested that the number of points required to estimate the correlation dimension within 5% of its true value is at least 42^M where M is the largest integer less than the dimension of the fractal (Smith 1988). However for simple models, 5,000 is a rough lower bound for the number points needed to achieve approximate results (Ramsey and Yuan 1989). The estimate obtained for attractors is known to be biased high but this bias decreases with sequence length; estimates for random noise are actually biased low (Ramsey and Yuan 1989). The effect of finite length may be reduced by applying the widest possible range of scaling. In a random walk, the resolution is limited by the distance travelled since this distance is a function of the number of steps in the walk. (See Chapters 3 and 4 for discussion and results of embedding dimension and length issues for DNA sequences.)

Human sequences were selected for analysis from GenBank version 55, based entirely on length of the sequence. There were 164 human nucleic acid sequences of length 4,500 to 15,000 and all were included in the study. The sequences analyzed are not necessarily representative of the human genome. Several sequences are from gene families, some chromosomes are underrepresented and the sample is severely deficient in noncoding DNA. The sequences came from GenBank so the sample includes mostly important or interesting genes rather than genomic sequences in general.

2.3 Results

The fractal dimension of 164 human DNA sequences, represented as four-dimensional pseudorandom walks (see section 3.2.2) was estimated using four methods: (1) Gate's Manhattan distance, (2) Maximum radial distance, (3) Grassberger-Procaccia algorithm (GPA), and (4) Tel's sandbox algorithm (Table 2.1).

These results reveal several things. All estimates were calculated using a four-dimensional embedding scheme. All algorithms produced global fractal dimensions greater than 2.0. This is evidence that a four-dimensional embedding with this representation is appropriate. It also indicates that any embedding dimension less than four is invalid according to the embedding dimension rule. Global fractal dimension estimates for

Table 2.1

Statistics on the estimated global fractal dimension, D, of 164 human DNA sequences by four different algorithms.

Method	Mean D	Standard Deviation	Minimum D	Maximum D
Gate's	1.536	0.185	1.172	2.348
Radius	1.653	0.185	1.265	2.194
GPA	1.660	0.209	1.165	2.254
Sandbox	1.631	0.137	1.300	2.253

the same 164 DNA sequences using a two-dimensional embedding scheme corroborates this fact. Virtually every DNA sequence produced a D that was lower for the two-dimensional embedding compared to four-dimensional embedding with an average decrease of 0.264.

Table 2.2 compares the calculated D by GPA method for two- and four-dimensional embeddings. A one-tailed paired t-test indicates that D for the 164 DNA sequences using the two-dimensional embedding is significantly lower ($p<0.0001$) than D using four-dimensional embedding. (A formal discussion and analysis of the effect of embedding dimension is in section 3.3.1.)

GPA and sandbox methods use basically the same underlying approach. It appears to provide more accurate estimates of D since the standard deviation of sandbox method is lower than GPA. The difference between mean D for GPA and sandbox method is due to nonoptimal box covering of the random walks using GPA. The

Table 2.2

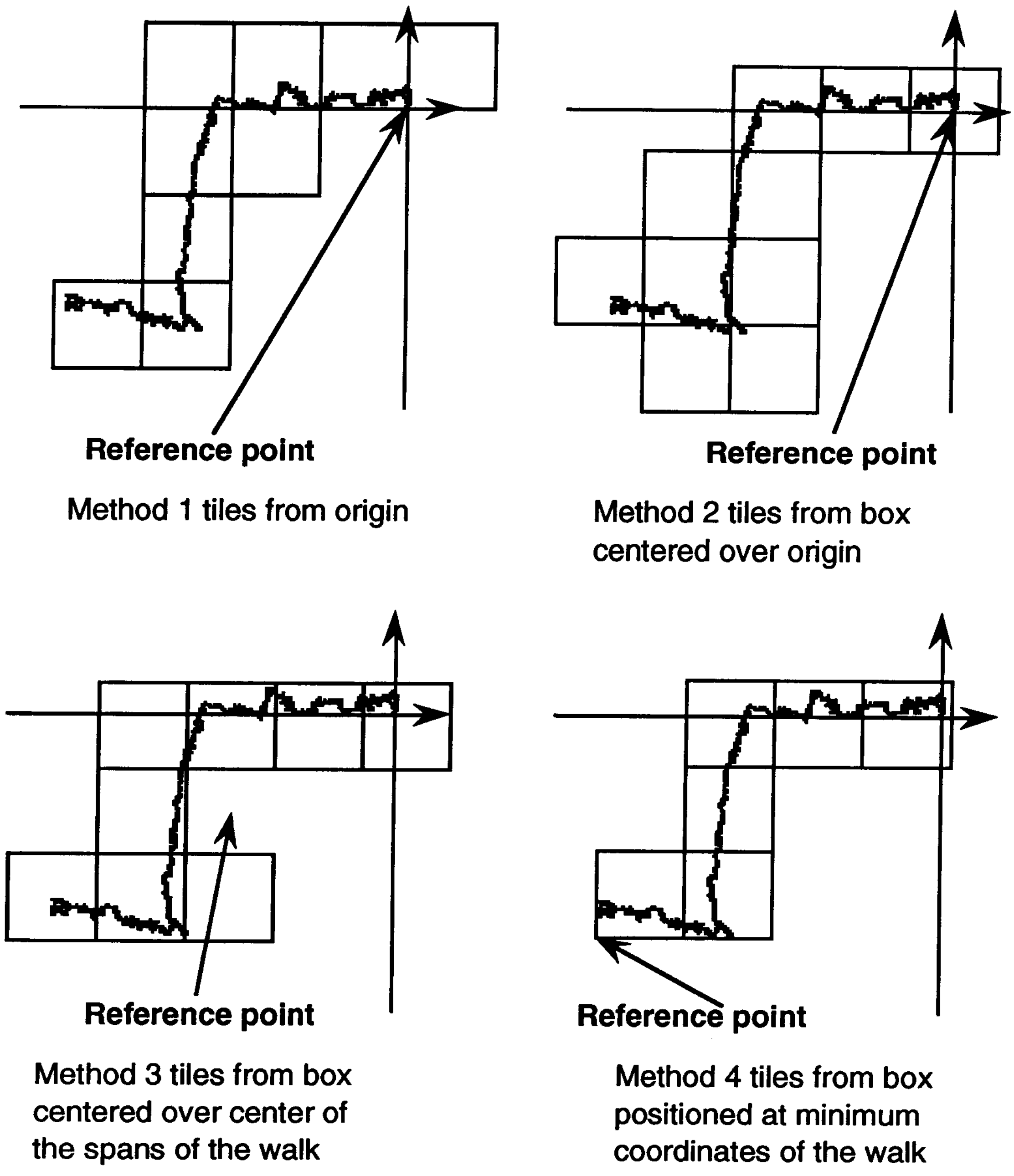
The two-dimensional embedding scheme produces significantly suppressed estimates of global fractal dimension compared to the four-dimensional embedding scheme.

	Mean D	Standard Deviation	Minimum D	Maximum D
2D GPA	1.395	0.146	1.084	1.802
4D GPA	1.660	0.209	1.165	2.254

box-covering method in the above calculations used a tiling that started at the origin, Method 1. Three other tiling approaches were tried to see if better results for GPA could be obtained. Tiling Method 2 tiled from a box centered over the origin. Tiling Method 3 tiled from a box centered over the center of the random walk. Tiling Method 4 tiled from a box positioned at the minimum value of each axis. Figure 2.1 demonstrates these four tiling schemes.

The results of estimating the global fractal dimension of 164 human DNA sequences by the four different tiling schemes are in Table 2.3. Paired t-test between the estimates for the 164 sequences by tiling method indicates that all tiling methods compute significantly ($p<0.0001$) different D with only one exception. Tiling methods 2 and 3 give approximately the same D for each sequence ($p=0.213$).

The fractal dimension estimate of 164 DNA sequences by the maximum radial distance method shows a high degree of.



Method 1 tiles from origin

Method 2 tiles from box centered over origin

Method 3 tiles from box centered over center of the spans of the walk

Method 4 tiles from box positioned at minimum coordinates of the walk

Figure 2.1 A two-dimensional representation of four different methods of tiling to box count using GPA algorithm. When applied to four-dimensional space, Methods 2 and 3 produce essentially the same global fractal dimension. Methods 1 and 4 produce significantly different D.

Table 2.3

Estimates of the global fractal dimension of 164 DNA sequences using four tiling approaches indicate significant sensitivity to box tiling for the GPA algorithm.

Tiling Approach	Mean D	Standard deviation	Minimum D	Maximum D
Method 1	1.377	0.079	1.182	1.589
Method 2	1.304	0.058	1.145	1.467
Method 3	1.308	0.060	1.163	1.500
Method 4	1.419	0.084	1.207	1.624

correlation with the estimates of D obtained by box counting (r^2=0.737). A test for difference between means of the two samples indicates that is no difference between the two groups (p=0.75). A paired t-test also indicates that there is no difference between the two estimates for individual sequences (p=0.44). Therefore, if only a generalized fractal dimension is needed to describe a system, the maximum radial distance method is a valid and very efficient approach.

Box counting provides the same estimate for D as maximum radial distance method, but only if issues of box-size range, sequence length, tiling methods, and logarithmic box-size increments are carefully addressed. If the range is too narrow, a poor line fit gives unpredictable estimates of D. If the maximum box size is set too low or too high, the estimate of D is decreased. Ds, which were equivalent to maximum radial distance estimates, were obtained by box-counting by using a minimum box size of five and a maximum box size equal to 30% of the maximum width of the fractal. A linear box size increment gives disproportionate weighting to small box sizes. Equation 2.6 measures the logarithmic change in number of boxes needed as the box size changes. Therefore, the box size increment in box counting should be logarithmic. Finding the box size range to obtain a good estimate of D using box counting can be difficult. Maximum radial distance method is a better algorithm than box counting because it is not plagued by this difficulty. Despite the fact that maximum radial distance is a better algorithm to estimate the global fractal dimension, it cannot be extended to compute the multifractal spectrum of generalized fractal dimensions.

The estimate of D for 164 DNA sequences using the Manhattan distance between the beginning point and end point of the random walk shows minimal correlation with box counting (r^2=0.699). However, there is a very significant difference between the means ($p<10^{-8}$) with the average fractal dimension by Manhattan distance significantly lower than box counting. Furthermore, a paired t-test indicates that there are significant differences between the estimates for individual sequences by method (p=.0001). The estimate of D by Manhattan distance has an even stronger correlation with maximum radial distance estimates (r^2=0.882). However, the difference between means and the paired t-test results are the same as the comparison with box counting--that is, the Manhattan distance estimator is significantly lower and differences exist for individual sequences.

Entropy, S, from thermodynamics, defined by Boltzmann

(Equation 2.10), is a measure of disorder, randomness, or lack of information in a system.

Equation 2.10 Boltzmann entropy. k_B is the Boltzmann constant $(\log n)^{-1}$ and p_i is the probability of the ith occurrence.

$$S = - k_B \sum_{i=1}^{n} p_i \ln p_i$$

For a perfectly ordered system S=0 and for a totally random system S=1. In the case of dimer frequencies, S measures the amount of divergence from the uniform distribution of dimers. There are 16 possible dinucleotides in a DNA sequence. The Boltmann entropy of dinucleotide frequencies is defined in Equation 2.11.

Equation 2.11 Boltzmann Entropy of dinucleotide frequencies of a DNA sequence. p_i is the probability of the ith dimer (of the 16 possible dimers) occurring in the sequence. Boltzmann constant is 1/log(16) in this case.

$$S = \frac{\sum_{i=1}^{16} p_i \ln p_i}{- \ln 16}$$

Dimer-frequency entropy measurements were obtained for DNA sequences ranging from 0.934 to 0.993, which indicate only minor divergence from uniform dimer frequency distributions. Furthermore, no correlation (r^2=0.024) was found between the estimate of D for a sequence and its dimer-frequency entropy.

Attempts to calculate multifractal spectra using GPA were unsuccessful. Nonoptimal box-covering plagued the process significantly, and it was difficult to obtain D_q curves that were monotonically nonincreasing. Log/log plots of the scaling range versus density changes revealed considerable sensitivity to edge effects and open areas in the random walk. The log/log plot in Figure 2.2 was puzzling at first. It did not seem logical that more of a larger sized box would be required to cover the graph. This phenomenon is explained graphically in Figure 2.3. The smaller sized box fits exactly inside the hole of the gasket so requires one less boxes to cover the gasket than the larger sized box. Attempts to calculate the multifractal spectrum using GPA were abandoned when it was discovered that Tel's sandbox method was much less sensitive to edge effects and open areas in the graph. It was much easier to obtain a monotonically nonincreasing D_q curve using the sandbox method. Chapter 5 details the research done with multifractal spectra.

2.4 Discussion

The fractal dimension obtained by box counting and maximum radial distance are equivalent. Both can be used to estimate the well-defined Hausdorff or similarity dimension, but maximum radial distance method is clearly more efficient. The Manhattan distance estimator calculates fractal dimension, but it is not equivalent to box counting or maximum radial distance. No attempt was made to prove or disprove its validity, but this

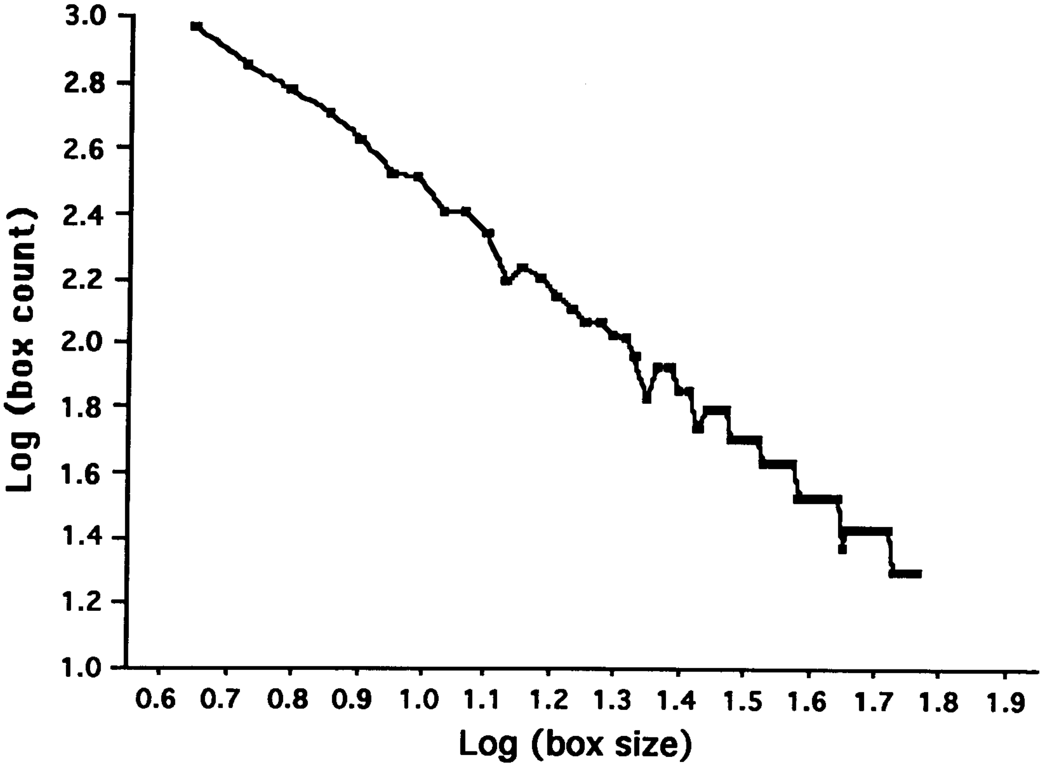
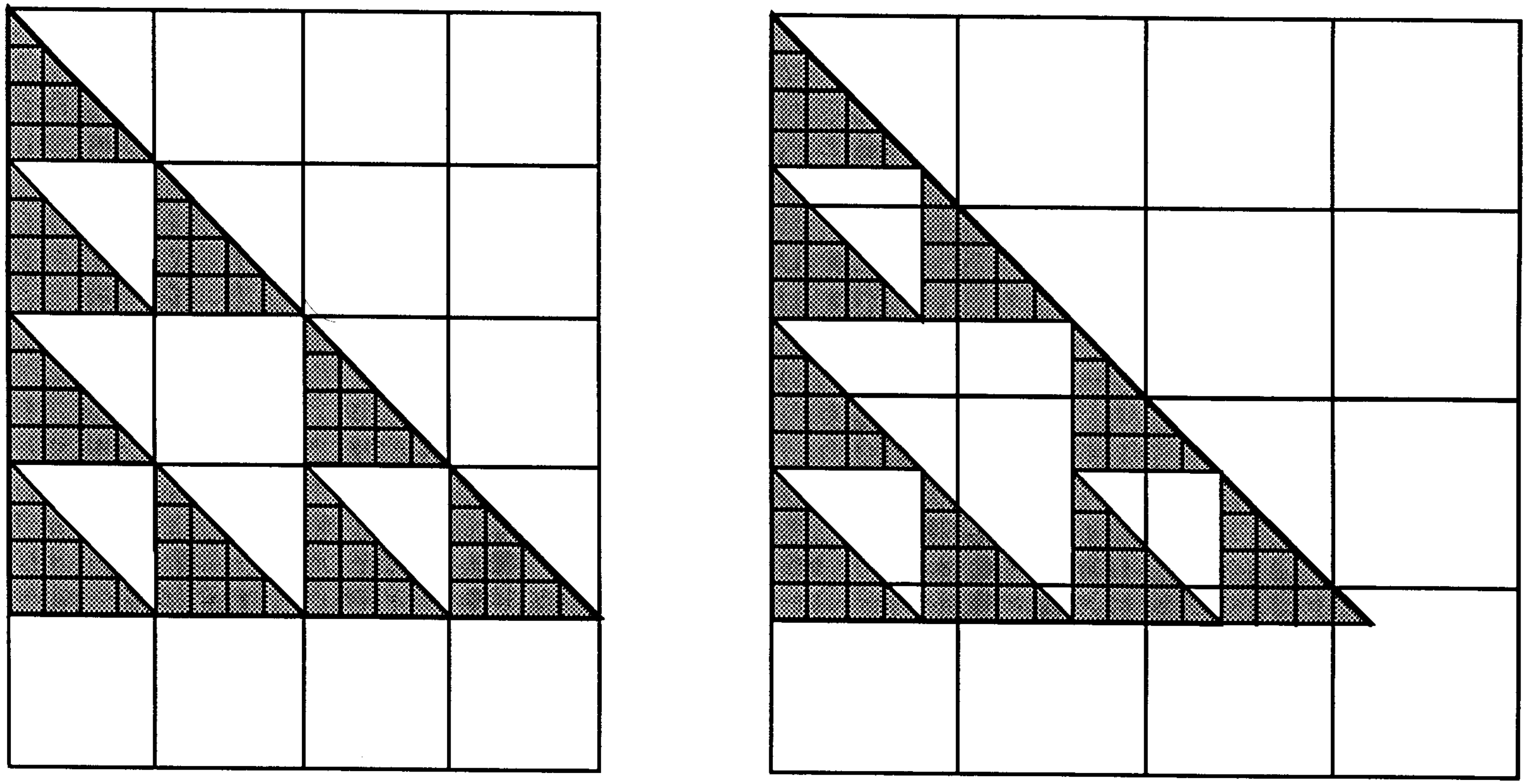


Figure 2.2 A puzzling log/log plot where the number of boxes required to cover a graph occasionally increases for a larger box rather than monotonically decreasing.



Requires NINE 4 x 4 Boxes to Cover

Requires TEN 5 x 5 Boxes to Cover

Figure 2.3 Example of a larger box size requiring more boxes to cover than a smaller box size in Sierpinski triangle gasket

research has shown that using a two-dimensional embedding produces invalid results. If the technique is valid when used with proper embedding dimension, it is not known where it belongs in the D_q spectrum of fractal dimensions.

Further research is needed to find a more systematic way of determining the proper scale to apply to individual fractals. The range to be used may be influenced by length of the sequence, the distance travelled along each axis of the random walk, the presence of long linear regions, and the size of the clusters. A systematic approach will allow a more definitive study of the fractal dimensions of DNA sequences.

3. GLOBAL FRACTAL DIMENSION OF HUMAN DNA SEQUENCES TREATED AS PSEUDORANDOM WALKS

3.1 Introduction

To a first approximation, 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 (Gates 1986). The pseudorandom walks derived from DNA sequences in Figures 3.1-3.6 show obvious qualitative deviations from the paired random walks shown in Figures 3.7-3.12. 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 (Dewey and Datta 1989; Helman, Coniglio, and Tsallis 1984; Isogai and Itoh 1984; Stapleton et al.

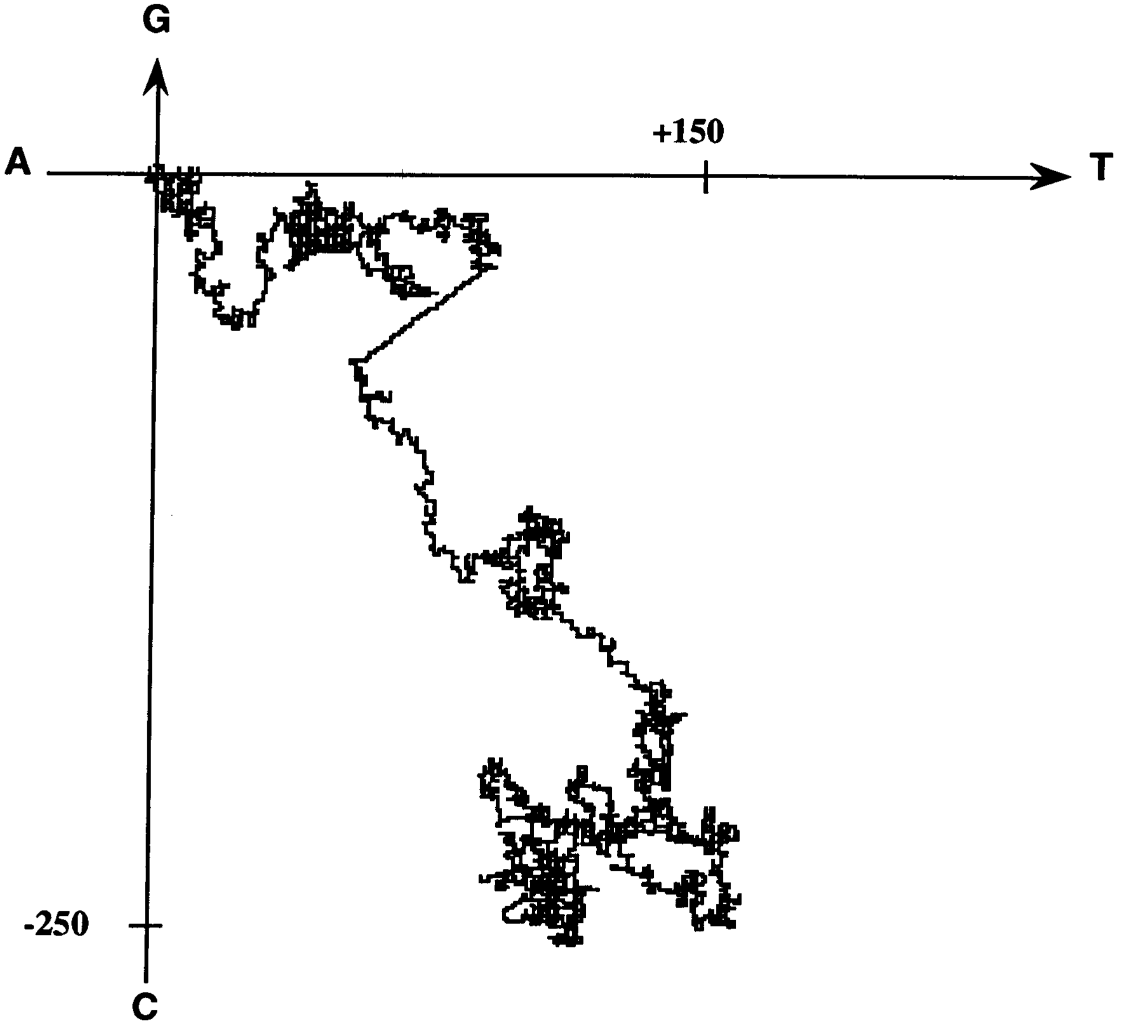


Figure 3.1 Pseudorandom walk of the human opsin gene sequence, Accession# K02281, 6,953 bps, D=1.650.

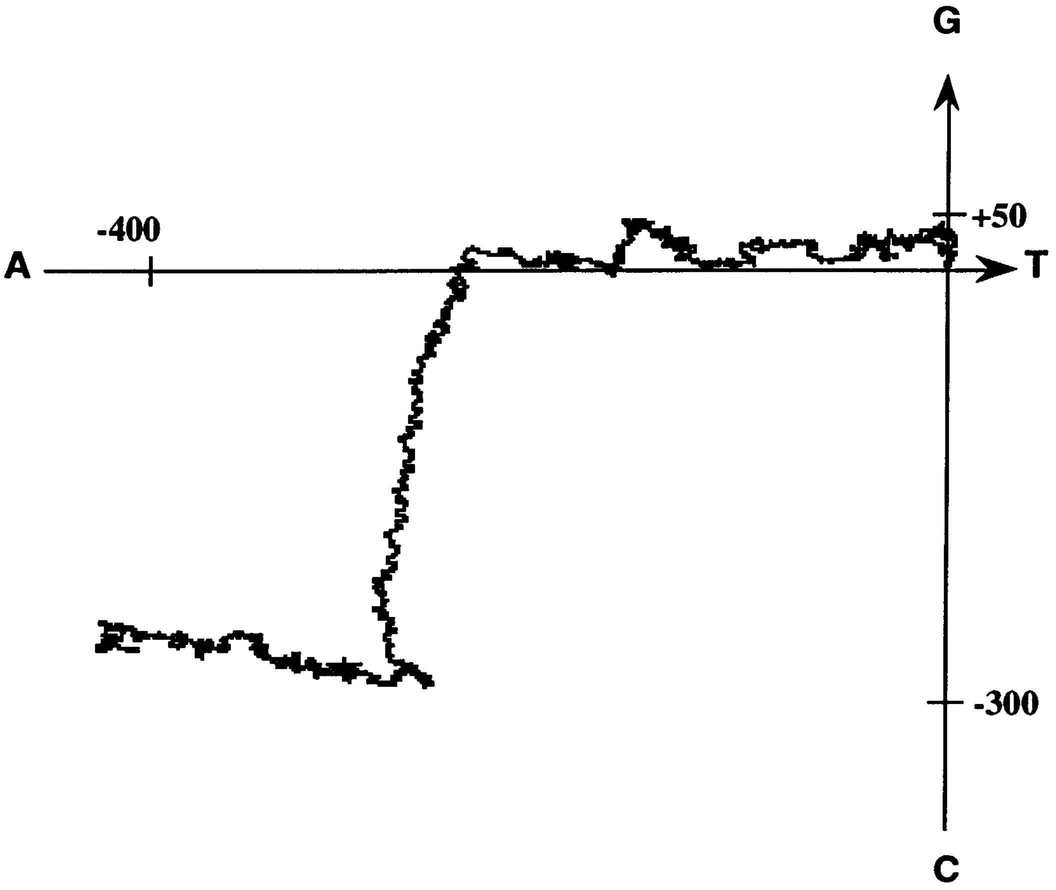


Figure 3.2 Pseudorandom walk of human factor V mRNA sequence, Accession# M16967, 6,909 bps, D=1.490.

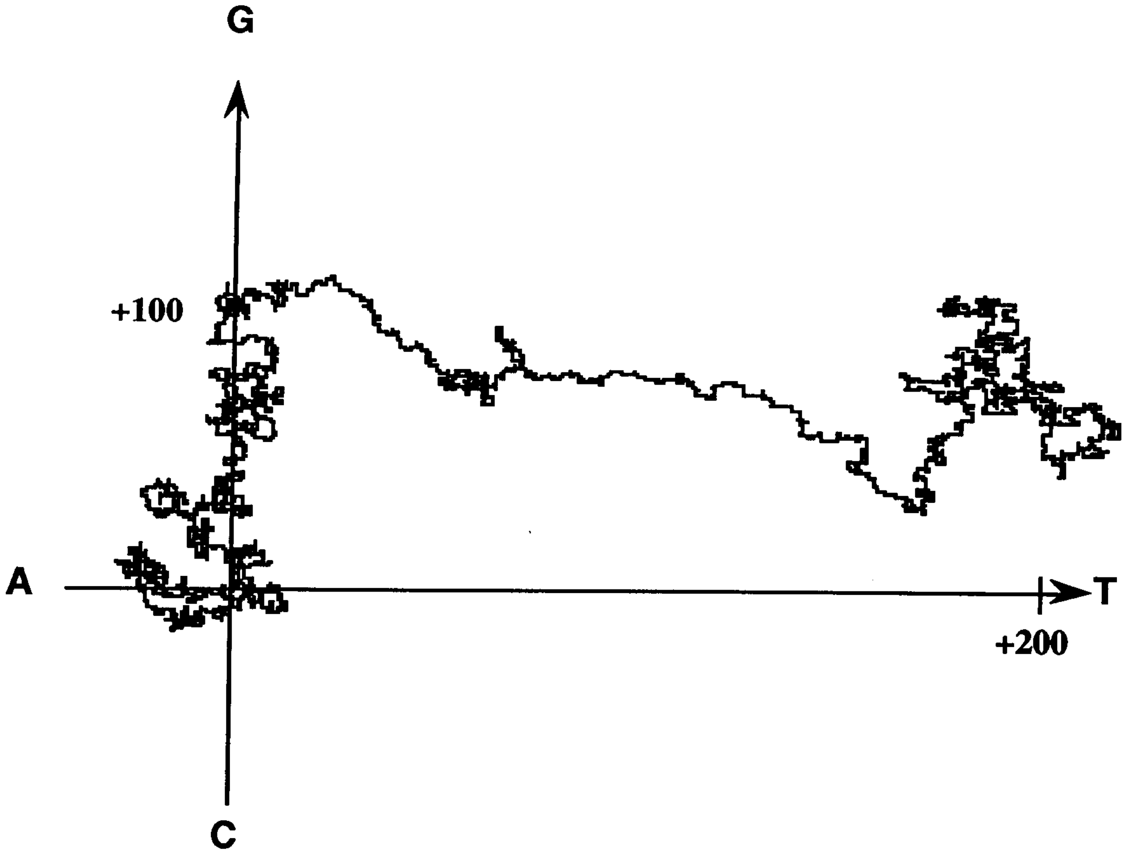


Figure 3.3 Pseudorandom walk of the human T-cell receptor germline beta-chain sequence, Accession# M14158, 4,913 bps, D=1.541.

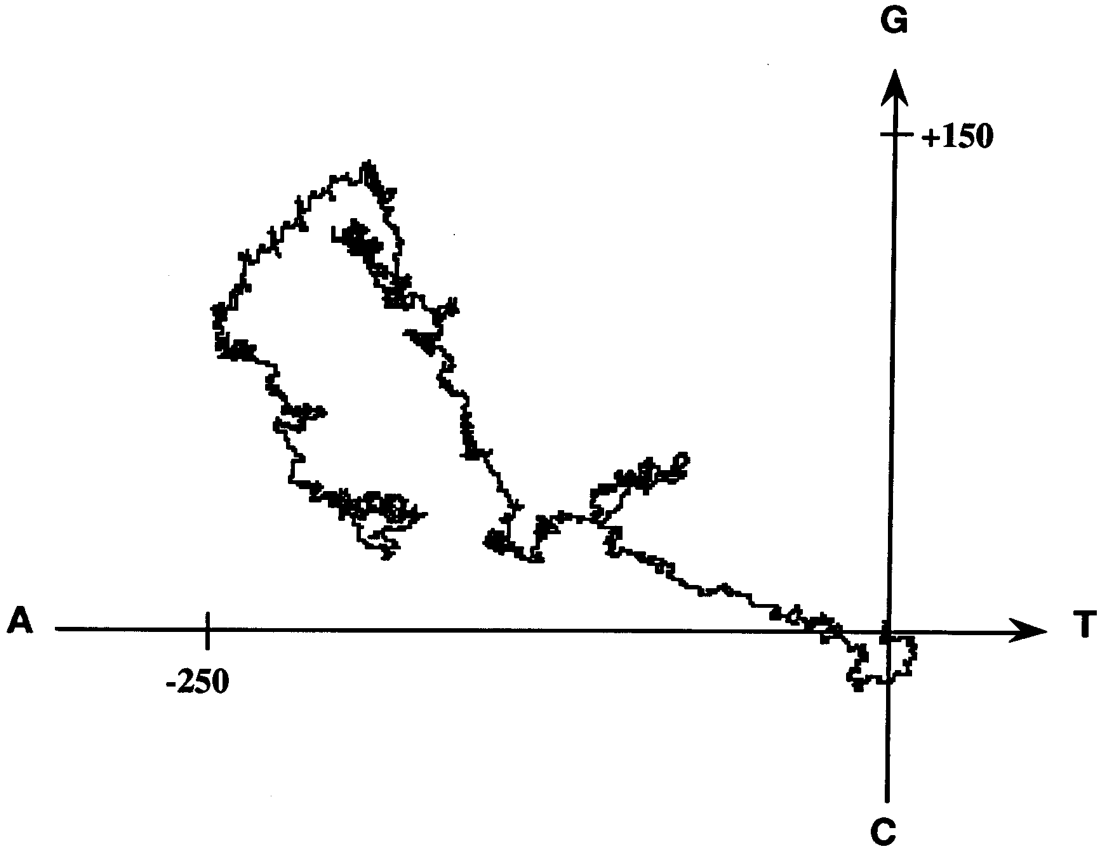


Figure 3.4 Pseudorandom walk of the human PRH1 gene sequence (Hae II-type subfamily), Accession# M13057, 4,946 bps, D=1.532.

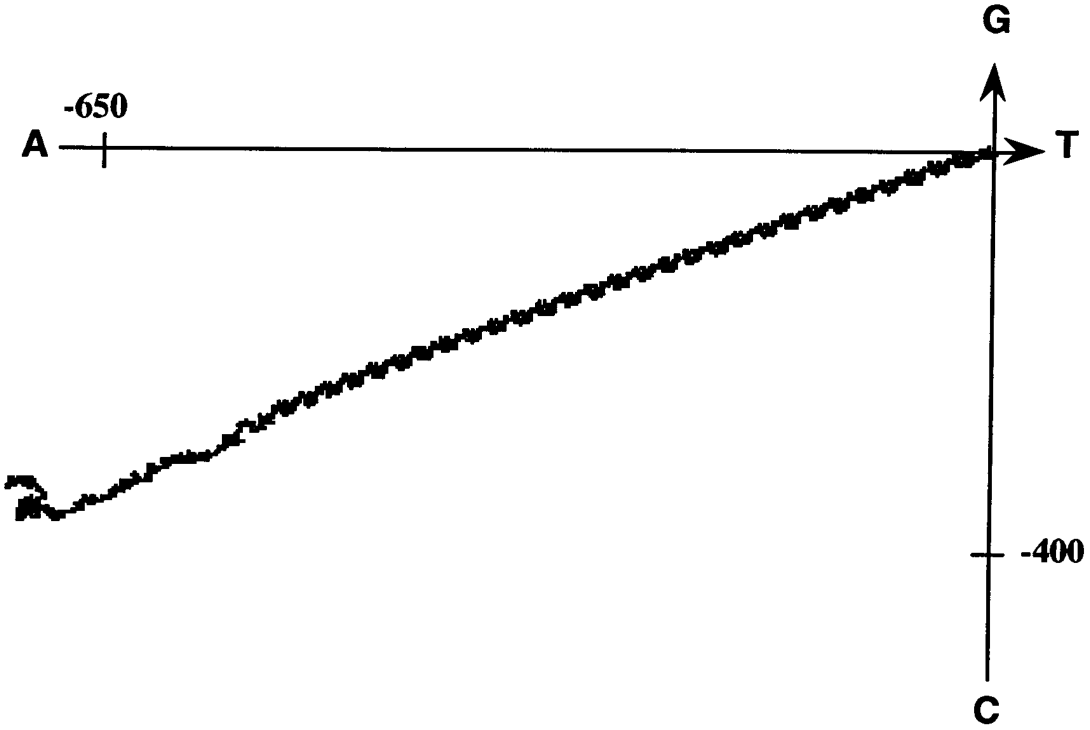


Figure 3.5 Pseudorandom walk of human apolipoprotein(a) mRNA sequence, Accession# X06696 M17399, 13,938 bps, D=1.547.

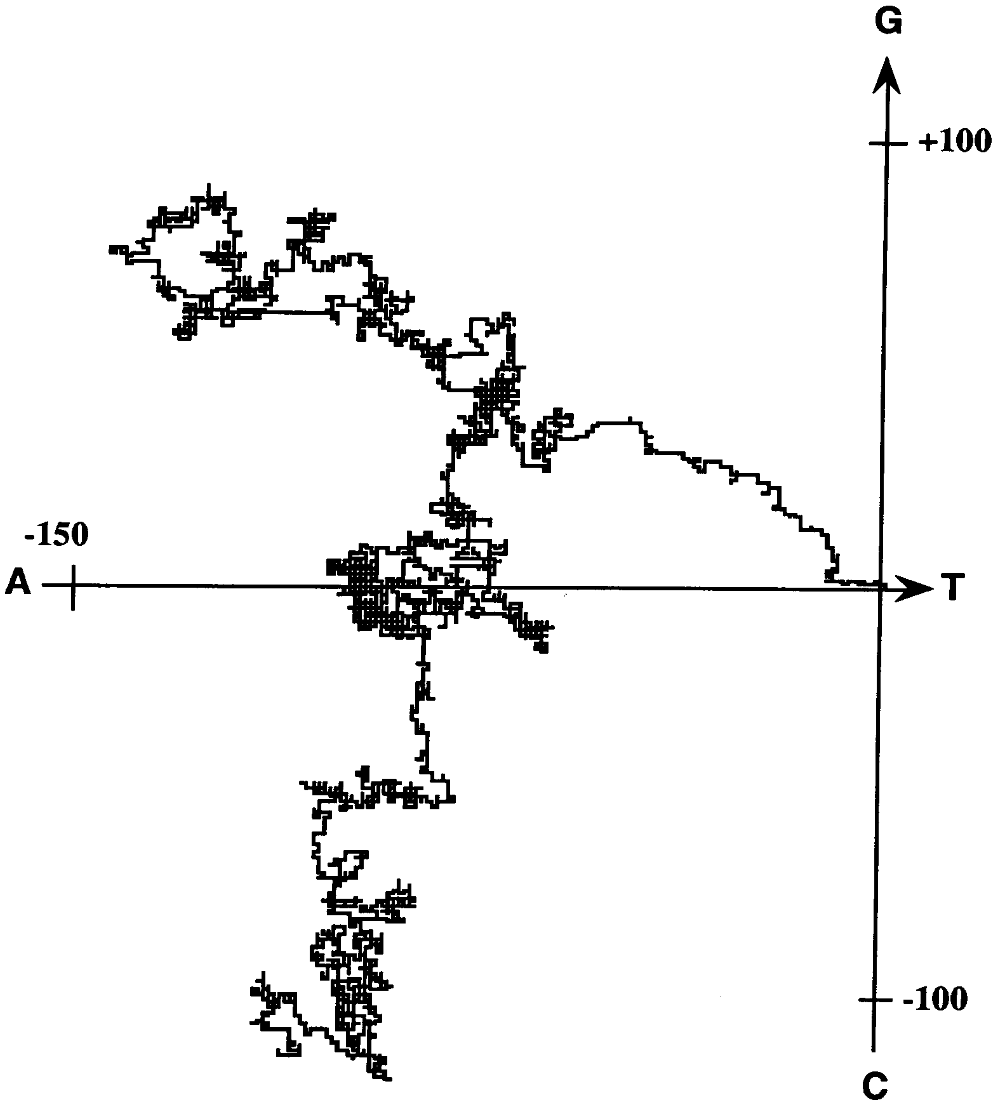


Figure 3.6 Pseudorandom walk of human alpha-1-acid glycoprotein-2 gene sequence, Accession# M21540, 4,944 bps, D=1.671.

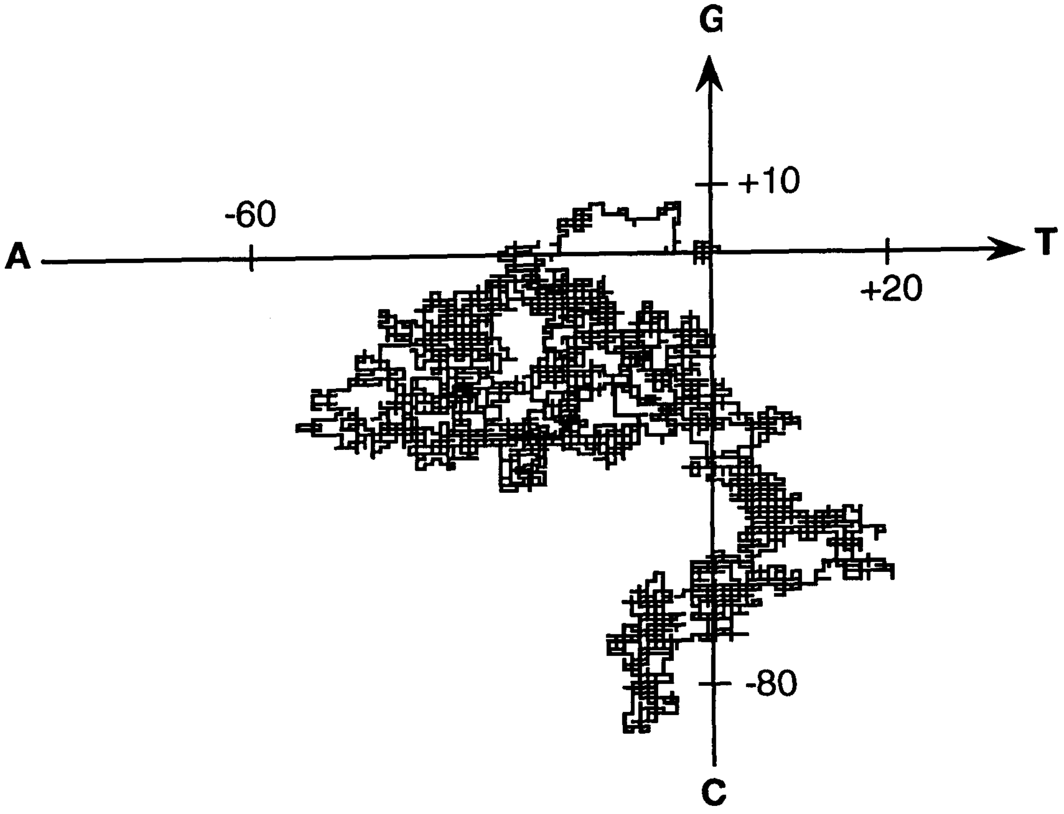


Figure 3.7 Pseudorandom walk of a random control sequence for human opsin gene, D=1.891.

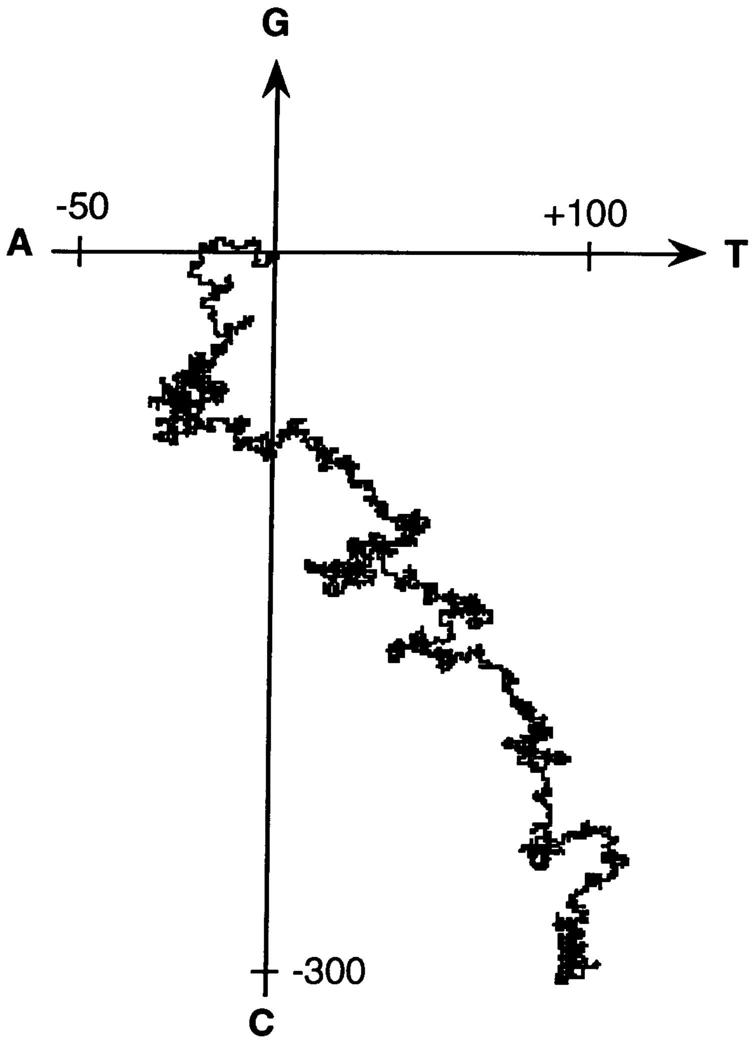


Figure 3.8 Pseudorandom walk of a base-matched control sequence for human opsin gene, D=1.701.

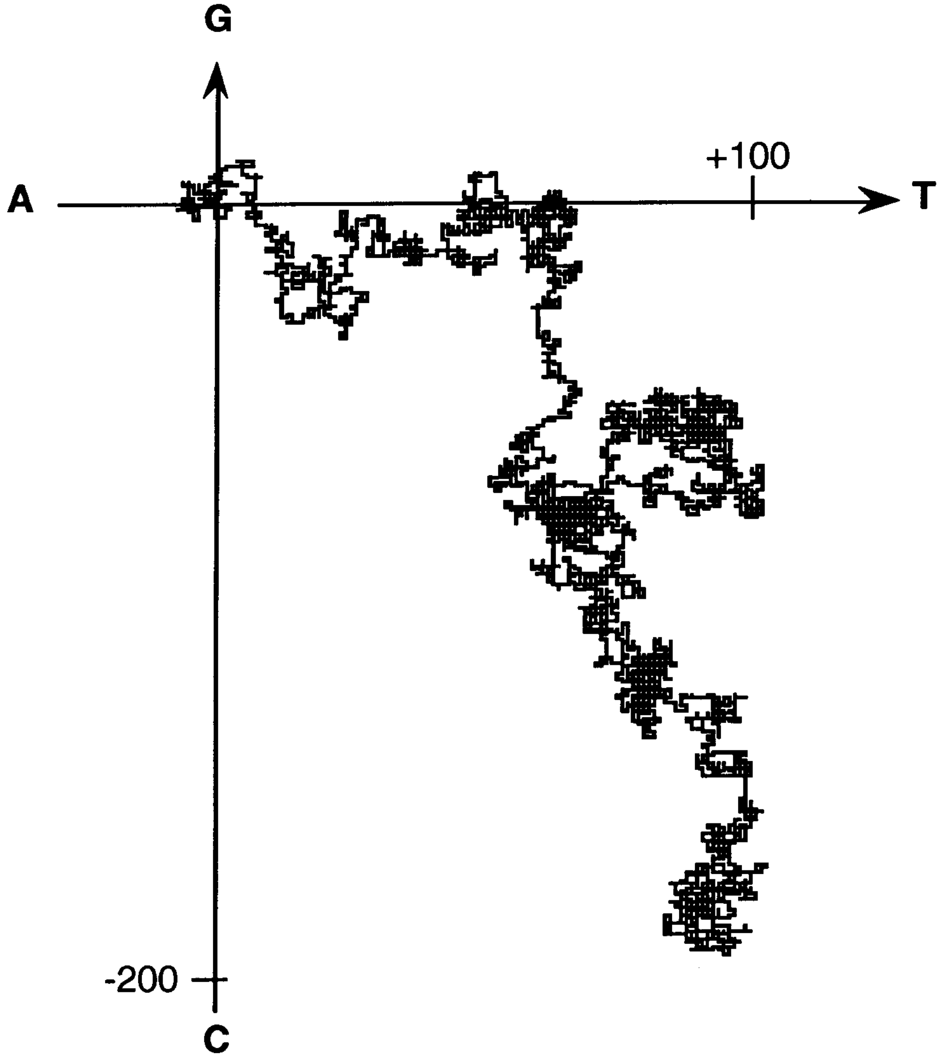


Figure 3.9 Pseudorandom walk of a dimer-matched control sequence for human opsin gene, D=1.744.

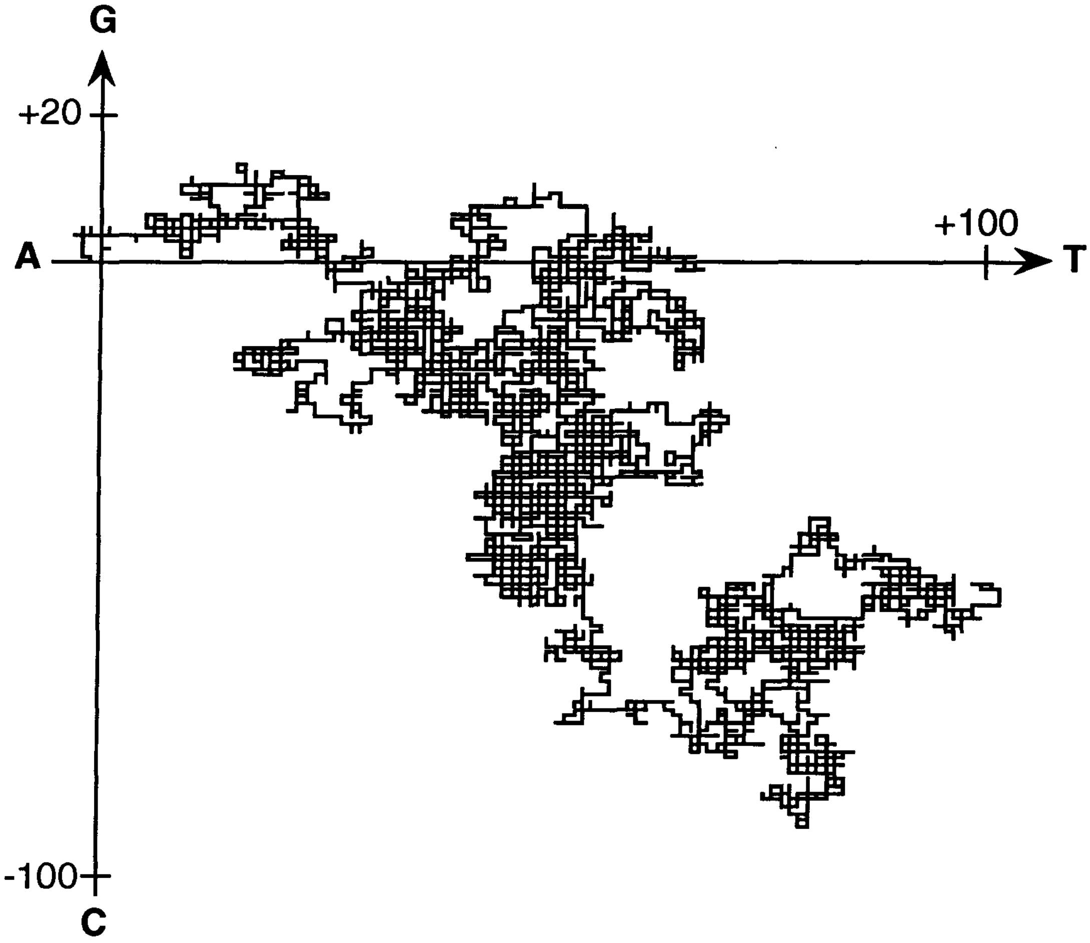


Figure 3.10 Pseudorandom walk of a random control sequence for human factor V mRNA, D=1.779.

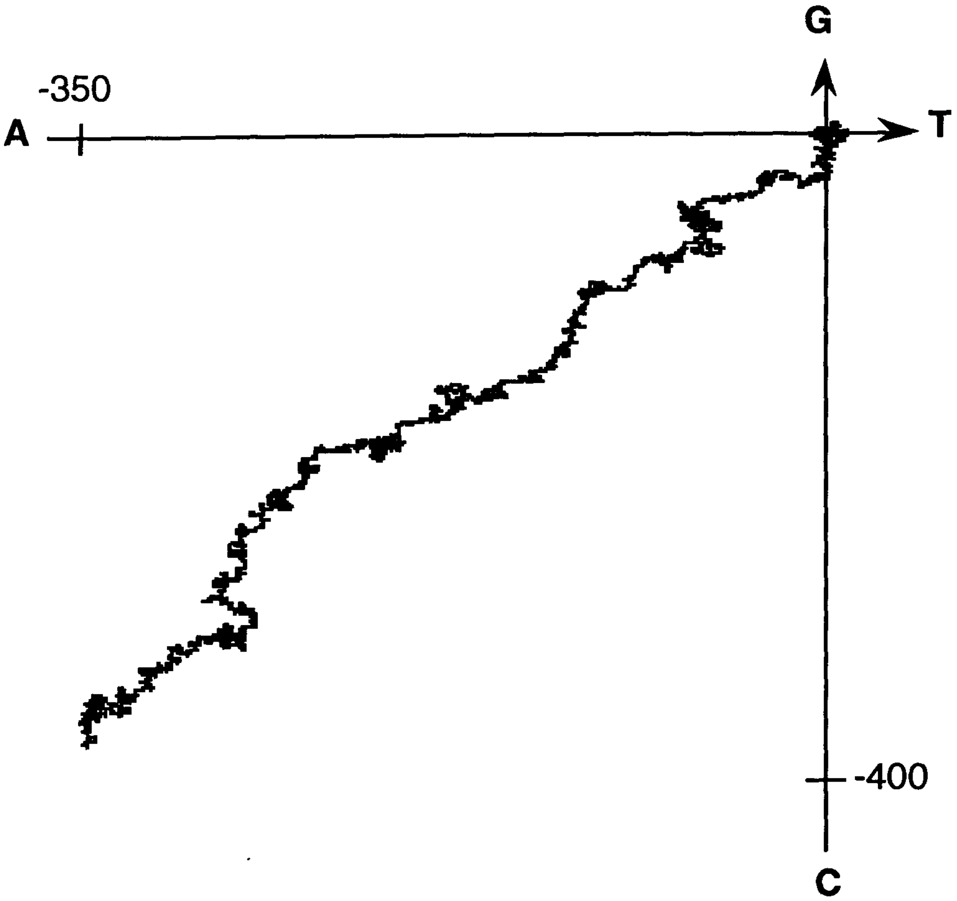


Figure 3.11 Pseudorandom walk of a base-matched control sequence for Human factor V mRNA, D=1.536.

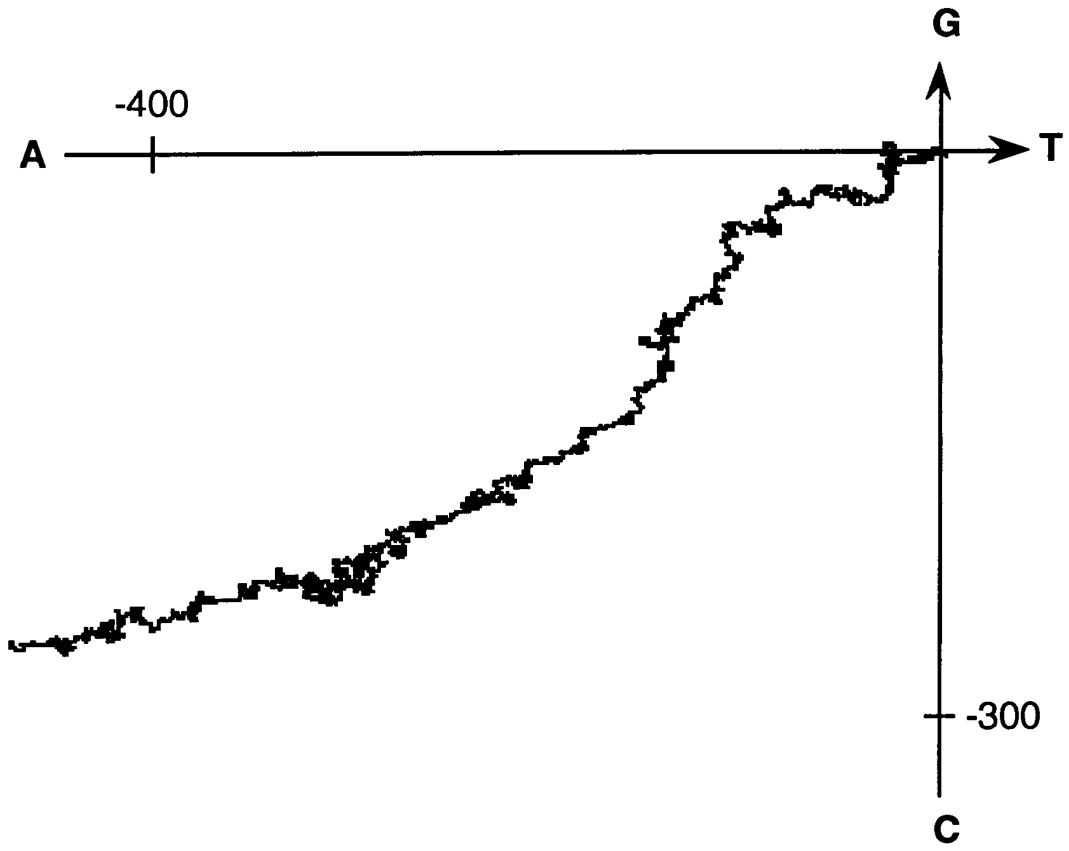


Figure 3.12 Pseudorandom walk of a dimer-matched control sequence for human factor V mRNA, D=1.600.

1990; Wang and Shi 1990) to three-dimensional structures of DNA (Takahashi 1989) and RNA (Purugganan 1989). There also has been some limited fractal analysis of DNA sequences.

Gates (1986) suggested that nucleic acid sequences could be represented as random walks in two-dimensional space with the bases C/G on opposite ends of one axis and A/T on the other. He suggested two methods to calculate D, which were reviewed in Chapter 2 (see Equations 2.2, 2.3, and 2.4).

Luo (Luo and Tsai 1988) calculated D for 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 D using the mean square separation between endpoints 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 (Shannon 1948). Their study assumed that the D of a single, relatively short DNA sequence is representative of all DNA of an organism.

Jeffrey (1990) investigated a graphic representation of nucleic acid sequences using iterated function systems (Barnsley 1988). 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 (1991) 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.

Just as I was completing this dissertation, an article was published about long-range correlations found in nucleotide sequences (Peng et al. 1992). They represented DNA sequences as walk in a one-dimensional embedding with steps based on whether the base was a purine or pyrimidine. They concluded that long-range correlations were present in intron-containing genes and in nontranscribed regulatory DNA sequences. Long-range correlations were absent in DNA seuqences that were purely coding DNA. One the problems in the research design that may have led to this conclusion is the fact that they used a one-dimensional embedding. This research has shown conclusively that at least a three-dimensional embedding is required to obtain valid results.

These five 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 or lower.

A detailed exploration of fractal analysis including the estimation of D is presented in this chapter. The technique of pseudorandom control sequences is used to evaluate the Ds of 164 relatively long human sequences (4,500-15,000 bases). The issues of adequate sequence length, proper embedding dimension, and scaling ranges have been addressed in the design of the calculations and analyses.

3.2 Methods

3.2.1 Methodological issues

As discussed in section 2.2, sequence length is an important issue because of finite length errors. To evaluate the effect of finite sequence length on the estimate of D, random sequences were generated composed of equal base frequencies over a range of lengths, 25 of each length, and D was estimated for each. Figure 3.13 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 5,000 is 1.847 with a standard deviation of 0.101. The standard deviation of D is 5.5% of the mean at length 5,000. Although D did not converge to D=2 at these lengths, finite-length errors can be controlled by always

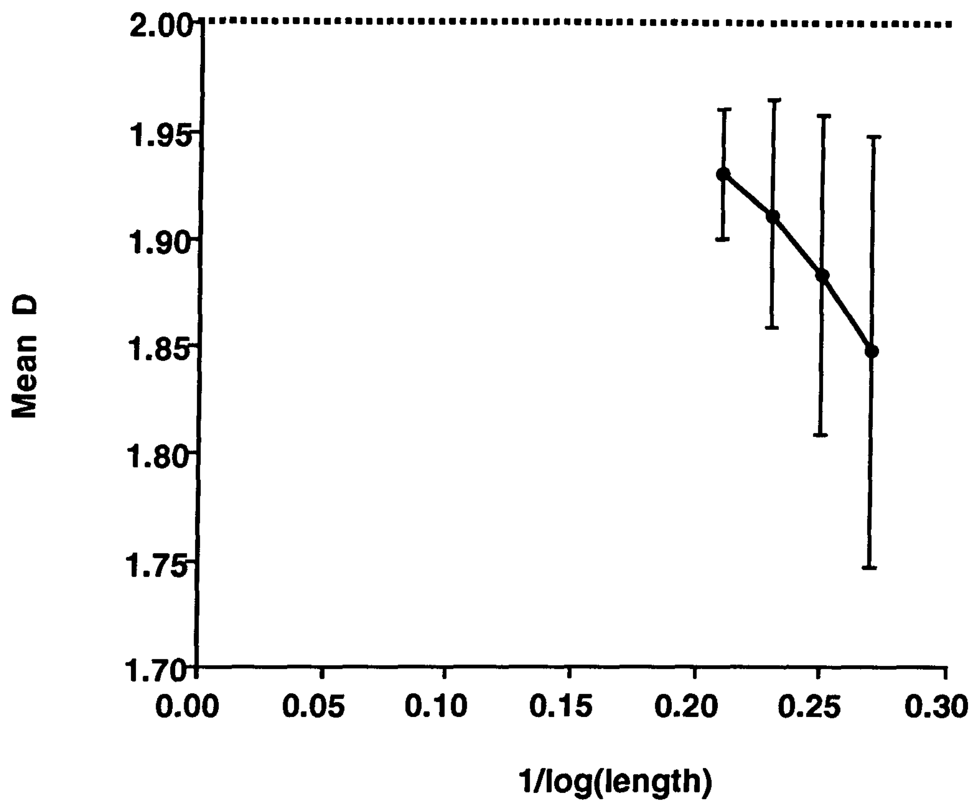


Figure 3.13 D versus length for random sequences. D increases with the length of random sequences and the standard deviation decreases.

comparing length-matched sequences. DNA and control sequences of the same length will be affected by finite length errors in precisely the same way, which does not affect the statistical analyses. Therefore, reasonable estimates of D are possible for random walks at least 4,000 bases in length with a standard deviation of D less than 6% of the mean. To assure convergence to a D of 1.995, a sequence longer than 500,000 is needed, which rules out analysis of available DNA sequences, typically 5,000-50,000 bases in length.

Human sequences were selected from GenBank version 55 for analysis based entirely on length of the sequence. There were 164 human nucleic acid sequences of length 4,500 to 15,000. (Average length was 7,178 bases.) All 164 were included in the study. Although 57 mRNA sequences are included, the group is loosely referred to 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 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. The sequences came from GenBank so the sample includes mostly important or interesting genes rather than representative sequences.

3.2.2 Random walk representation

The following natural method was used to convert a DNA sequence into a pseudorandom walk in N dimensions. A DNA sequence is represented as a series of vectors $\vec{x_j}$ 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. The axis assignments were specifically chosen 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 pseudo-random 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 strict biological constraint on the representation. Other symmetries suggested by biololgical factors are discussed below and in section 3.3.2. For a true random walk, the assignment of the symbol types is arbitrary and the direction in which the walk is read should be unimportant. Therefore, to preserve the symmetries of the random walk in the embedding structure, the following were required:

1. *Complementarity symmetry*: 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'.

2. *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'.

3. *Compatibility symmetry*: 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.

4. *Substitution symmetry*: D remains unchanged under the single exchange of either A<-->T or G<-->C.

Complementarity symmetry (1) is a special case of (2) and (4). (4) is also suggested by the natural biological grouping of A and T as weak-bonding and G and C as strong bonding bases.

In two dimensions complementarity (1), reflection symmetry (2), and compatibility (3) are satisfied for any single base representation. Substitution symmetry (4), however, requires that {A}=-{T} and {G}=-{C}. Otherwise, the sequence AG is a zero-step trajectory, whereas AC is not, which violates substitution symmetry (4). Therefore, the axis assignments employed were:

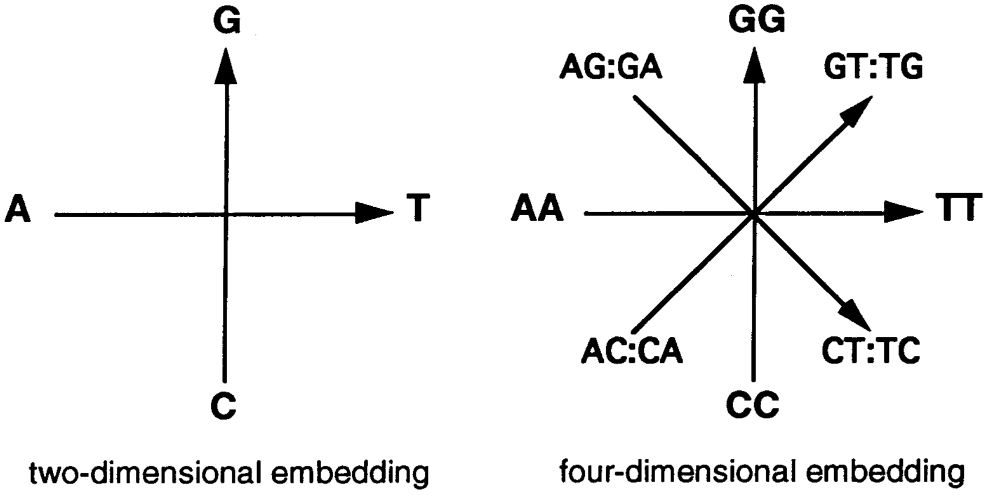
Axis 1: {A}=(-1,0) and {T}=(1,0)
Axis 2: {C}=(0,-1) and {G}=(0,1).

In four and higher dimensions a new $\vec{y}_i$ for each base in the sequence is begun. Thus, the representation is independent of reading frame and each base is used in two successive vectors. The four-dimensional embedding was determined as follows: Complementarity (1) 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 (2) requires that {AC}={CA}, {AG}={GA}, {AT}={TA}, {CG}={GC}, {CT}={TC}, and {GT}={TG}. Thus, {AC}, {CA}, {GT},.and {TG} must be grouped on one axis and {AG}, {GA}, {CT}, and {TC} on another axis. If {AC}={TG} are paired rather than {AC}={CA}, compatibility with the two-dimensional scheme is violated. Similarly, compatibility symmetry means that {AC}={CA}={GT}={TG}=0 cannot be true or {AG}={GA}={CT}={TC}=0 by substitution symmetry. Therefore, the only possible grouping is to set {AC}={CA}=-{GT}=-{TG}<>0 and {AG}={GA}=-{CT}=-{TC}<>0. In four dimensions the only scheme that obeys these conditions is:

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 degree angle lines in two dimensions. Figure 3.14 gives a graphic representation of the two-dimensional and four-dimensional embedding schemes.

To expand to six dimensions, {AT}, {TA}, {CG}, {GC} can be split or the other two quartets ({AC}, {CA}, {GT}, {TG} and {AG}, {GA}, {CT}, {TC}) from axes 3 and 4 can be split. The latter was chosen to preserve the compatibility of the assignments of the



two-dimensional embedding

four-dimensional embedding

Figure 3.14 Embedding schemes in two and four dimensions for pseudorandom walk representations of DNA sequences.

zero trajectories in lower dimensional embeddings as much as possible. However, the choice here is arbitrary. (The six-dimensional embedding was not used for any of the major calculations.) The six-dimensional embedding is:

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 eight-dimensional embedding for dimer pairs. Two axes for the AT/TA and CG/GC pairs are added, which were zero-step trajectories in lower dimensions:

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 the fact that the estimate of D is not independent of representation can be exploited by relaxing the symmetry conditions to allow other representations. As discussed in section 1.1, DNA sequences that code for protein have a strong bias for content and arrangement of certain dimers. The axis assignments can be changed to emphasize these dimers in the estimated fractal dimension to 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 in Equation 2.1. A true random walk of infinite length will be space filling with D=2 for all embedding dimensions (Mandelbrot 1983; Rudnick and Gaspari 1987) However, the representation of a DNA sequence in dimensions greater than 2.0 is not a true random walk. In a true random walk every step is totally independent of any of the previous or subsequent steps. The sliding dimer scheme used to map DNA into a pseudorandom walk is a correlated walk because each step in the walk has partial dependence on the previous step. An important parameter of any random walk is the mean-square displacement of the walker after *n* steps. The mean-square displacement of a random walk is defined in Equation 3.1.

Equation 3.1 Mean-square displacement of a random walk after *n* steps is a function of the square root of *n* multiplied by a constant, *b*, for all *n*.

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

The mean-square displacement for a walk with finite correlation is defined in Equation 3.2.

Equation 3.2 Mean-square displacement of a random walk with finite correlation after *n* steps is a function of the square root of *n* multiplied by a constant, *A*, where *A* depends on the exact nature of the correlation.

$$\langle R_n^2 \rangle = An \qquad n \rightarrow \infty$$

The fact that the exponent in Equation 3.2 is just a scaling exponent means that this is basically a variant of the maximum radial distance method of estimating the global fractal dimension in Equation 2.5. The four-dimensional embedding scheme used is a single partial correlation C_1, corresponding to a walk with each step dependent on its immediate predecessor. The mean-square displacement for this correlated walk is defined in Equation 3.3 (Barber and Ninham 1970).

Equation 3.3 Mean-square displacement of a correlated random walk after n steps where each step depends on its immediate predecessor. C_1, the correlation function is less than 1 for single partial correlation.

$$R_n^2 \approx \left(\frac{1 + C_1}{1 - C_1}\right) n$$

The numerator will be greater than 1 and the denominator will be less than 1 in this representation, so the multiplier will increase the mean-square displacement after n steps. For example, an uncorrelated random walk (C_1=0) will have a mean-square displacement of $\sqrt{25{,}000}$ or about 158 units. A random walk with a 0.25 correlation will have a mean-square displacement of $\sqrt{(1.25/0.75)\ 25{,}000}$ or about 204 units. The increased mean-square displacement will reduce the density of the walk yielding a lower fractal dimension. A totally random DNA sequence mapped into the four-dimensional space using a sliding dimer window is this type of correlated walk. Each step has the possibility of being followed by one of four dimers rather than one of sixteen. For example, a step in the CT direction can only be followed by either a step in the TA, TC, TG, or TT direction and TA is a zero-vector dimer. The graph in Figure 3.13 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 because the representation is a correlated walk rather than a true random walk. As stated before, this lack of convergence does not affect the statistical results because finite length effects and the correlation functions are the same for DNA and length-matched controls. The standard deviations are less than 7% of the mean for the shortest sequence evaluated, which allows reliable statistical analyses despite lack of true convergence.

3.2.3 Control sequences

Control sequences were generated to match the length of each DNA sequence using a random number generator. Pairing the DNA sequences with control sequences of the same length controls the effects of finite length. To avoid introducing sequential correlations in the control sequences, a linear congruential random number generator (Press et al. 1988) was used with a randomized shuffle and an effective period of at least 714,025, which is significantly longer than any of the sequences analyzed. Three types of control sequences were generated: (1) random controls: each base occurs with a probability of 0.25; (2) base-matched controls: the frequency of each base is determined from the DNA sequence and that frequency is used to generate the control sequences; and (3) dimer-matched controls: the frequency of each dinucleotide pair is determined and the control sequences are generated using the probability of the what the next base will be, given the base selected previously.

3.2.4 Estimating the global fractal dimension

The global fractal dimension, known as the Hausdorff dimension, was calculated using the sandbox algorithm (Tel, Fulop, and Vicsek 1989). 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 (Tel, Fulop, and Vicsek 1989). D is defined in Chapter 2, Equation 2.7.

To estimate D for a given sequence, 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 is calculated. The critical parameter for the sandbox method is the range of radii (Tel, Fulop, and Vicsek 1989). 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 $\sqrt{1 + 1 + 1 + 1}$ or two metric units. Within these constraints, the range of radii should be as large as possible but must try to minimize the amount of probable overlap of radius regions. After extensive investigation of the scaling properties of random and pseudorandom walks, the optimum scaling range was determined to be [2,26] for this range of sequence lengths. A radius range [2,26] was used with an increment of two. 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. (See section 2.1.5 for the discussion on the effects of sampling procuderes.) The random walks applied in this research may visit a site multiple times, so a frequently visited site may be sampled more than once in a random sampling of data points. It was decided to randomly sample 1% of the data points of each random walk. Studies on the sampling rate (data not shown) indicated that a 1% sampling produced the same fractal dimension as 10% sampling. 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 be an important descriptor of the behavior of individual random walks. Thus, random point sampling incorporates density into the estimate of D.

3.3 Results

3.3.1 The effects of sequence length and embedding dimension

The effect of sequence length on the estimate of D was evaluated for the 164 DNA sequences. Although length clearly affects the error in the estimate of D for sequences longer than 4,000 bases, no correlation (r^2 = .024) was found between the length of a DNA sequence and its estimated D. A minimum length cutoff of 4,500 base pairs was determined.

Ten of the 164 DNA sequences were randomly selected 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 two, four, six, and eight dimensions. D increased between two- and four-dimensional embeddings but remained relatively constant between four, six, and eight dimensions (Figure 3.15).

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 an embedding dimension of at least four must be used to satisfy the embedding dimension rule ($D \geq 1+[2.348]=4$).

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

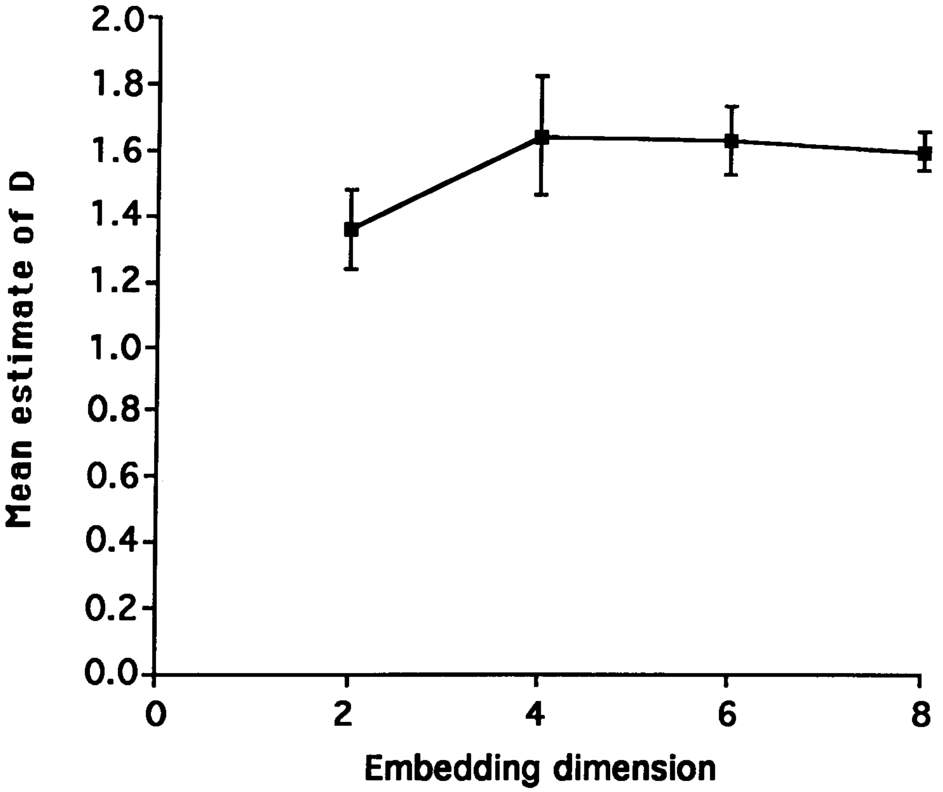


Figure 3.15 Fractal dimension versus embedding dimension averaged for 10 DNA sequences. The increase in D from two to four dimensions indicates that a two-dimensional embedding is insufficient. The gradual decrease in D for larger embedding dimensions demonstrates the effect of finite sequence length for higher embedding dimensions.

that the two-dimensional embedding was insufficient and a higher embedding was required. Embeddings greater than four 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 four was the result of bias due to finite sequence length.

3.3.2 Estimate of D for DNA sequences and controls

The estimated Ds for the 164 human sequences using the four-dimensional embedding scheme were compared to three control types: (1) random, (2) base-matched, and (3) dimer-matched. All controls match the paired sequence in length. To estimate the statistical distribution of D, 30 of each control type were generated for each DNA sequence. A z-score (Equation 3.4) was then calculated for each DNA sequence.

Equation 3.4 Standardized score or z-score where D_S is the estimated D of a sequence, $\overline{D}_C$ is the mean D for its matched controls, and $sd(D_c)$ is the standard deviation of the controls.

$$z = \frac{D_S - \overline{D}_C}{sd\ (D_C)}$$

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 the 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 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.

The aggregate results for all 164 DNA sequences evaluated are presented in Table 3.1. 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 Figure 3.16 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 Figure 3.17 indicates that over 50% of the 164 sequences have a rank below eight when compared to 30 base-matched or 30 dimer-matched controls. The number of sequences with a D rank of one was significant in both cases (p=0.0001 by chi-square test) with 23% (38/164) ranking lowest among their base-matched controls and 25% of the sequences (41/164) ranking lowest among their dimer-matched controls. Approximately 75% of the sequences had a rank lower than 16 for both base and dimer-matching.

It was expected that dimer-matched controls would match the DNA sequences better than base-matched controls. The fact that dimer-matching increased the magnitude of the difference rather than decreasing it requires further explanation. The base-matched and dimer-matched controls had bases and dimers

Table 3.1

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
No. of Sequences	164	164	57
DNA Sequences			
mean D	1.631	1.641	1.613
sd (D)	0.137	0.140	0.130
Random Controls			
mean D	1.863	1.865	1.859
sd (D)	0.027	0.027	0.027
mean z	-2.624	-2.603	-2.665
sd (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
sd (D)	0.127	0.126	0.129
mean z	-0.865	-0.833	-0.925
sd (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
sd (D)	0.109	0.105	0.111
mean z	-1.068	-1.193	-0.834
sd (z)	1.331	1.444	1.059
t value	-10.280	-8.543	-5.950
p value	$< 10^{-18}$	$<10^{-13}$	$<10^{-7}$

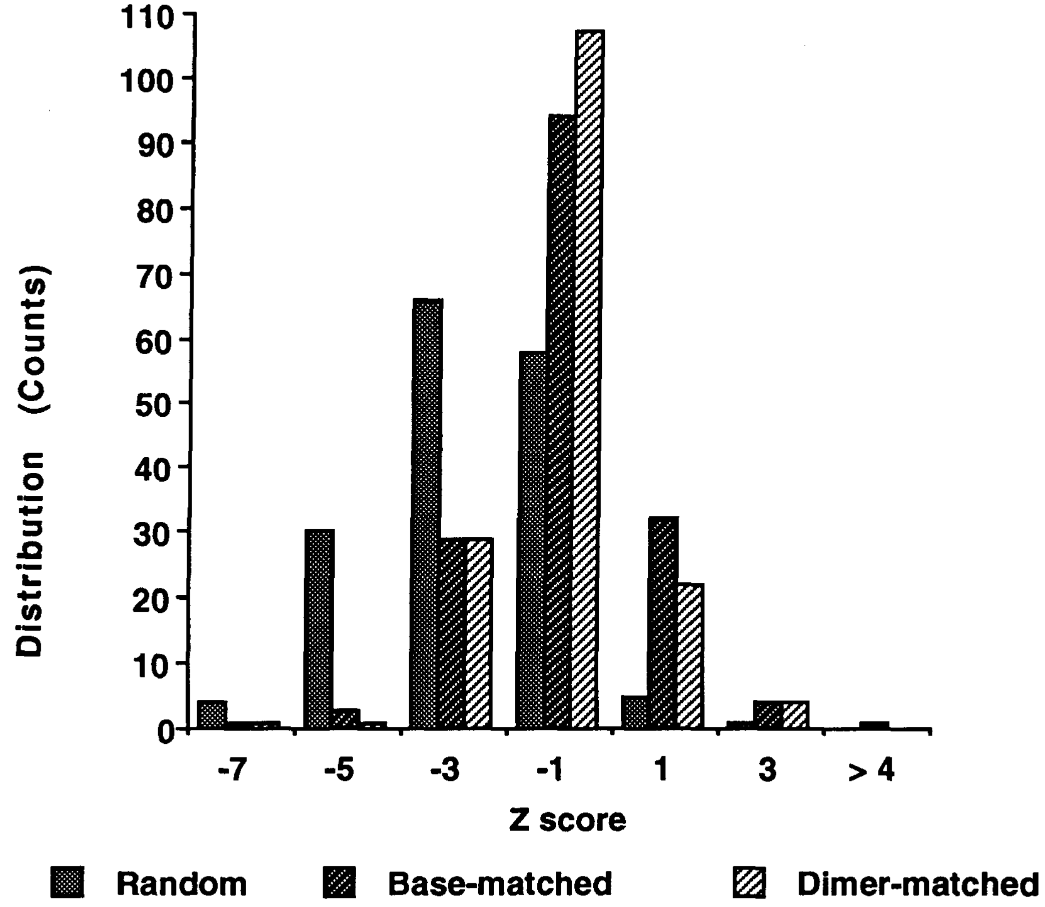


Figure 3.16 The distribution of z-scores for D of 164 DNA sequences compared to random, base-matched, and dimer-matched controls. The distribution is shifted in the negative direction indicating that D of human DNA is significantly lower than for all controls.

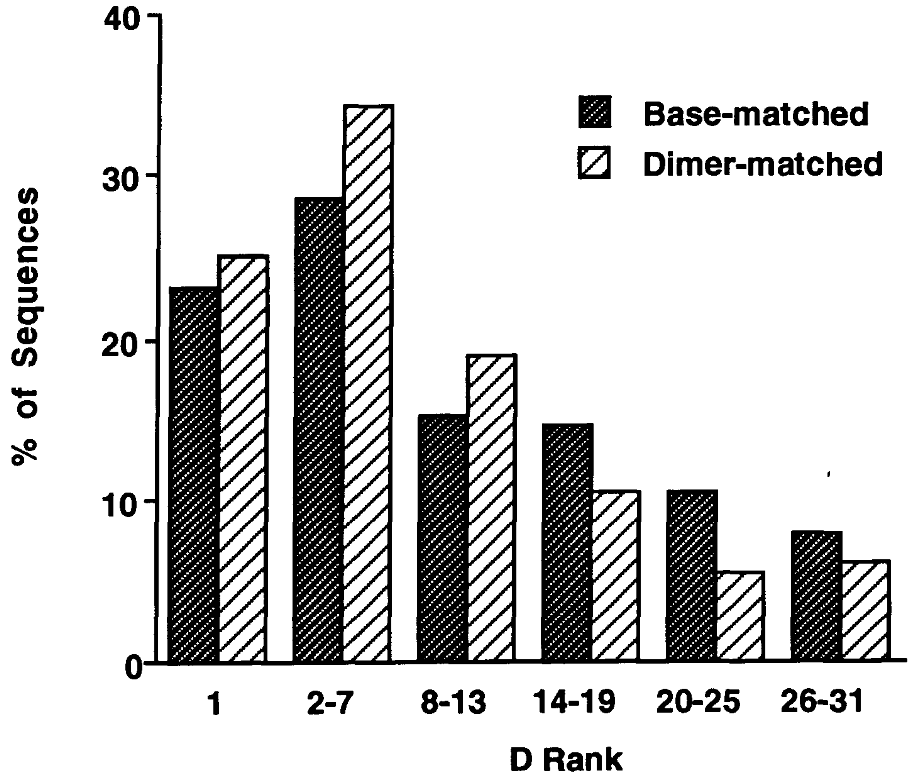


Figure 3.17 The distribution of D rankings for 164 human DNA sequences. The rank of D indicates how many matched-control sequences had a lower D.

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. Each DNA sequence was divided into 500-base subsequences, and the base and dimer content in these subsequences were compared to the overall base and dimer content. Wide fluctuations in base and dimer content were found within the sequences. Table 3.2 summarizes the results of base distribution analysis.

Only 8/164 (4.9%) of the sequences show uniform distribution of all four bases, whereas 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 Figure 3.18). Over 95% of the sequences

Table 3.2

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

Base	Number of sequences showing nonuniformity	%
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

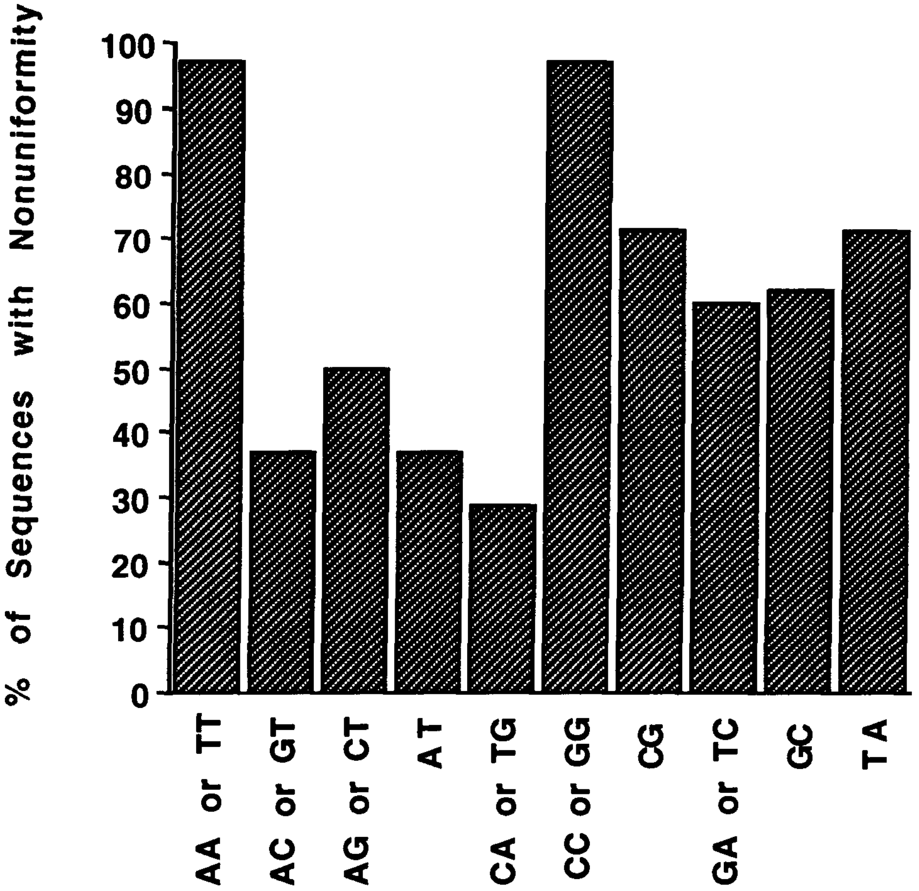


Figure 3.18 The distribution of dimers within sequences. The distribution shows marked non-uniformities with symmetries between mirror image dimers.

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 the requirement of reflection symmetry in the embedding scheme.

The 164 nucleic acid sequences in the sample consisted of two types: (1) mRNA sequences composed primarily of coding segments but with 5' and 3' untranslated segments; and (2) genomic DNA sequences, which are predominantly introns and other noncoding segments. The mean D for the genomic DNA group was 1.641 ±0.14 and for mRNA 1.613 ±0.13. However, this difference was not significant (p=0.20 by an unpaired t-test). Therefore, the z-scores of DNA and mRNA groups for random, base-matched, and dimer-matched controls were compared to determine if there was any difference in D between genomic DNA and mRNA. It was 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 ±1.444) was significantly (p=0.05) lower than the mean z-score for mRNA (-0.834 ±1.059).

3.4 Discussion

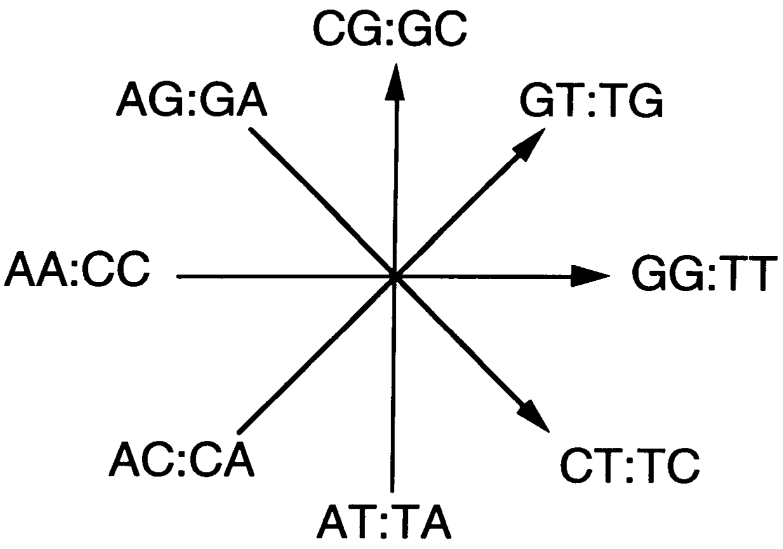
Matching for base frequencies and even dimer frequencies does not explain the nonrandomness of DNA sequences. The D estimates for the DNA sequences are significantly lower than the

Ds 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. Quasi-linear 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, which have been associated with nucleosome formation sites (Kimura, Takeya, and Takanami 1989; Pennings et al. 1989; Shrader and Crothers 1989; Uberbacher, Harp, and Bunick 1988). This finding correlates with the results of Markov chain analyses (Almagor 1983; Blaisdell 1985; Garden 1980; Kleffe and Langbecker 1990), which found strong nearest-neighbor effects in DNA sequences. There are also families of repetitive elements present in human DNA, which often contain internal short repeats.

The Ds 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 the power to discriminate between coding and noncoding populations.

This difference in D between genomic DNA and mRNA sequences agrees with the findings of Blaisdell (1983) 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. It was 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 (Elton 1975).

The results obtained in this study were based on one of many possible axis assignments. Do these results and conclusions change when an alternate four-dimensional scheme is used? To address this issue, a subset of 33 DNA sequences were studied, randomly selected from the original set, and D was calculated using a second alternative embedding scheme in which all dimers step (Figure 3.19).



Second four-dimensional embedding

Figure 3.19 Second four-dimensional embedding scheme in which all dimers step.

So the new four-dimensional embedding is as follows:

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 the two-dimensional representation (compatibility symmetry). The incompatibility is caused by the fact that this new representation converts the zero-vector dimers of AT/TA/CG/GC into nonzero vectors. A significant difference was found (p=0.005 by paired t-test) in the estimate of D for individual DNA sequences from that obtained using the first embedding scheme. There was no correlation (r^2=0.159) between the two estimates of D (Figure 3.20).

D obtained for DNA was compared to D obtained for base-matched controls using the second embedding scheme (Table 3.3). 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 p-value of $<10^{-8}$ for the 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 (Figure 3.21).

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. Although it is not proposed that all axis assignments will produce the identical result (some may be more or less discriminating than the two used in this research), this equivalence is strong evidence that the qualitative differences between random controls and DNA will persist regardless of embedding scheme.

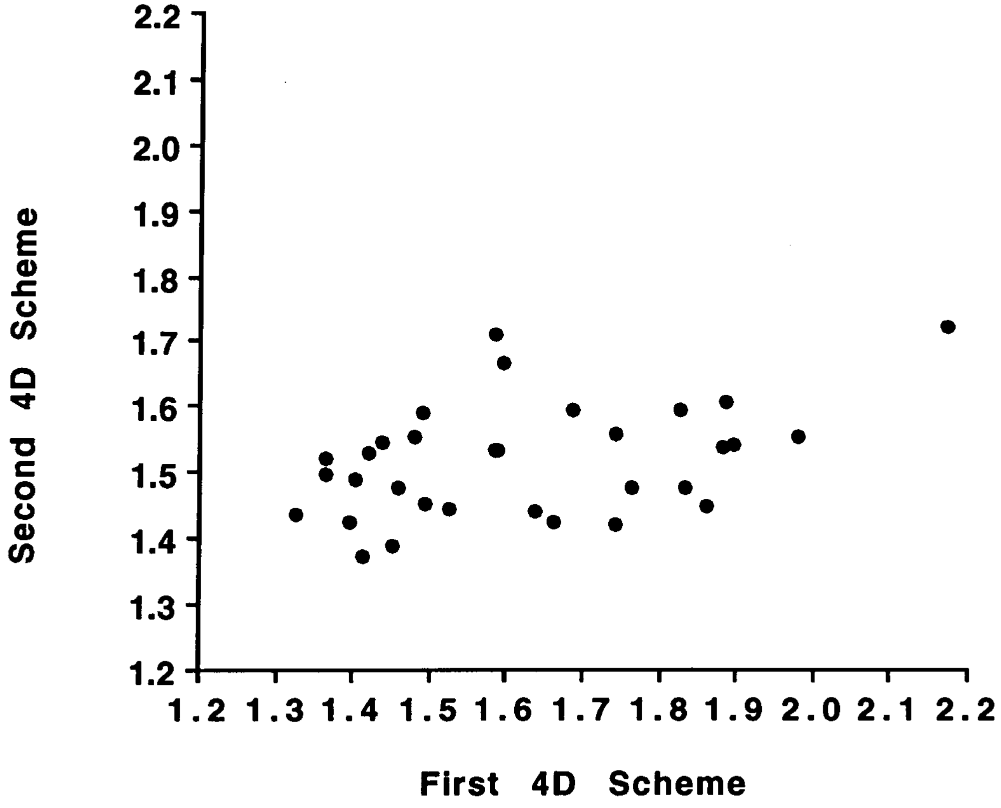


Figure 3.20 Scatterplot for a subset of 33 DNA sequences. There is no correlation between estimates of D using the first embedding scheme in Figure 3.14 and the second embedding scheme in Figure 3.19.

Table 3.3

Global fractal dimensions for DNA sequences and controls using a second four-dimensional embedding. The estimated global fractal dimension for a 33 sequence subset of original data using the embedding scheme in Figure 3.19 is significantly different from that obtained using the scheme in Figure 3.14. However, it is still significantly lower than D for base-matched controls.

	First 4D Embedding	Second 4D Embedding
No. of sequences	164	33
DNA sequences		
mean D	1.631	1.519
sd (D)	0.137	0.085
Base-matched Controls		
mean D	1.702	1.580
sd (D)	0.127	0.128
mean z	-0.865	-1.338
sd (z)	1.812	2.171
t value	-6.113	-3.540
p value	$< 10^{-8}$	0.0006

4. EFFECTIVE MULTIFRACTAL SPECTRA FOR RANDOM WALK AND SIERPINSKI CARPET

4.1 Introduction

The multifractal formalism is a useful way to characterize the spatial inhomogeneity of fractals (Grassberger and Procaccia 1983; Halsey et al. 1986; Hentschel and Procaccia 1983). It has been widely applied to analyze both theoretical and experimental fractal patterns. Of the common examples of a multifractal, the best known is the exactly soluble two-scale Cantor set, which is representative of a large class of point-like deterministic attractors. For many purposes, however, it is more convenient to have a model that is stochastic and higher-dimensional. The simplest example is the random walk on a lattice, which finds many applications in polymer physics, biology and economics. In this chapter numerical calculations of the D_q spectrum are presented for a random walk and Sierpinski carpet. These results should be particularly useful to compare to short data sets for which a true bulk multifractal spectrum cannot be obtained.

4.2 Methods

A true random walk of infinite length is space-filling with D=2 in two-dimensional embeddings. Higher embeddings also produce D=2 (Rudnick and Gaspari 1987). Local fluctuations can result in a local fractal dimension above two, so a four-dimensional embedding is used to work in a space at least one higher in dimension than the set examined (Takens 1981). The random walk is defined in Equation 2.1. The direction of each step is chosen at random with equal probability using a random number generator. To avoid sequential correlation, a linear congruential random number generator was applied that uses a randomized shuffle and produces a sequence with an effective period of at least 714,025 (Press et al. 1988). This is much larger than the longest DNA sequence in that data set, so no sequential correlations in the computations are introduced.

The D_q spectrum was calculated using the sandbox method of Tel that has been shown to converge substantially faster than box counting for the two scale Cantor set (Gould and Tobochnik 1990; Tel, Fulop, and Vicsek 1989; Viscek, Family, and Meakin 1990). The D_q spectrum by Tel's sandbox algorithm is defined in Chapter 2 (see Equation 2.9).

For an ideal fractal, D_q is defined according to Equation 4.1.

Equation 4.1 The multifractal spectrum is defined by D_q where R, the radius of spherical balls and D_q is the fractal dimension estimate at different values of q. Large negative values of q emphasize the sparsest part of the walk and large positive values of q emphasize the densest part of the walk.

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

In practice a linear fit is performed on $D_q(R)$ over a range of R $[R_{min}, R_{max}]$ with $R_{min}=2$ and R_{max} adjusted to the length representing the most linear region of the log/log plot of radius versus data point density. A two-unit increment is used over the range R_{min} to R_{max} since the smallest particle size is the equal to the square root of 4.

To compare results from pseudorandom walks with a known fractal, extensive computations were done with the square Sierpinski gasket, also called Sierpinski carpet (Figure 4.1). In theory, this carpet is a monofractal (D=log(8)/log(3)=1.893) since it should have the same fractal dimension at all scaling ranges (Vicsek, Family, and Meakin 1990). However, this is based on an infinitely large square that is subdivided an infinite number of times. A finite-sized carpet will have an effective multifractal spectrum due to inclusion of the edges of the carpet.

Four different sampling techniques using the sandbox method were applied to the carpet to evaluate what effect the edge had on the estimate of the multifractal spectrum. Method 1 randomly sampled points anywhere on the carpet. Method 2 randomly samples points that are at least R_{max} or more inside the outside edge of the carpet (Figure 4.2). This insures that range being studied never runs off the outside edge as radius is increased but it does over sample the center edge. Method 3 samples points within a strip that is equidistant from the outside and inside edges (Figure 4.3). This method includes both outside and center edges equally. Method 4 randomly samples points that were within R_{max} of the outside edge of the carpet (Figure 4.4). This technique totally avoids the center region but runs off the edge of the carpet with about half of the points sampled.

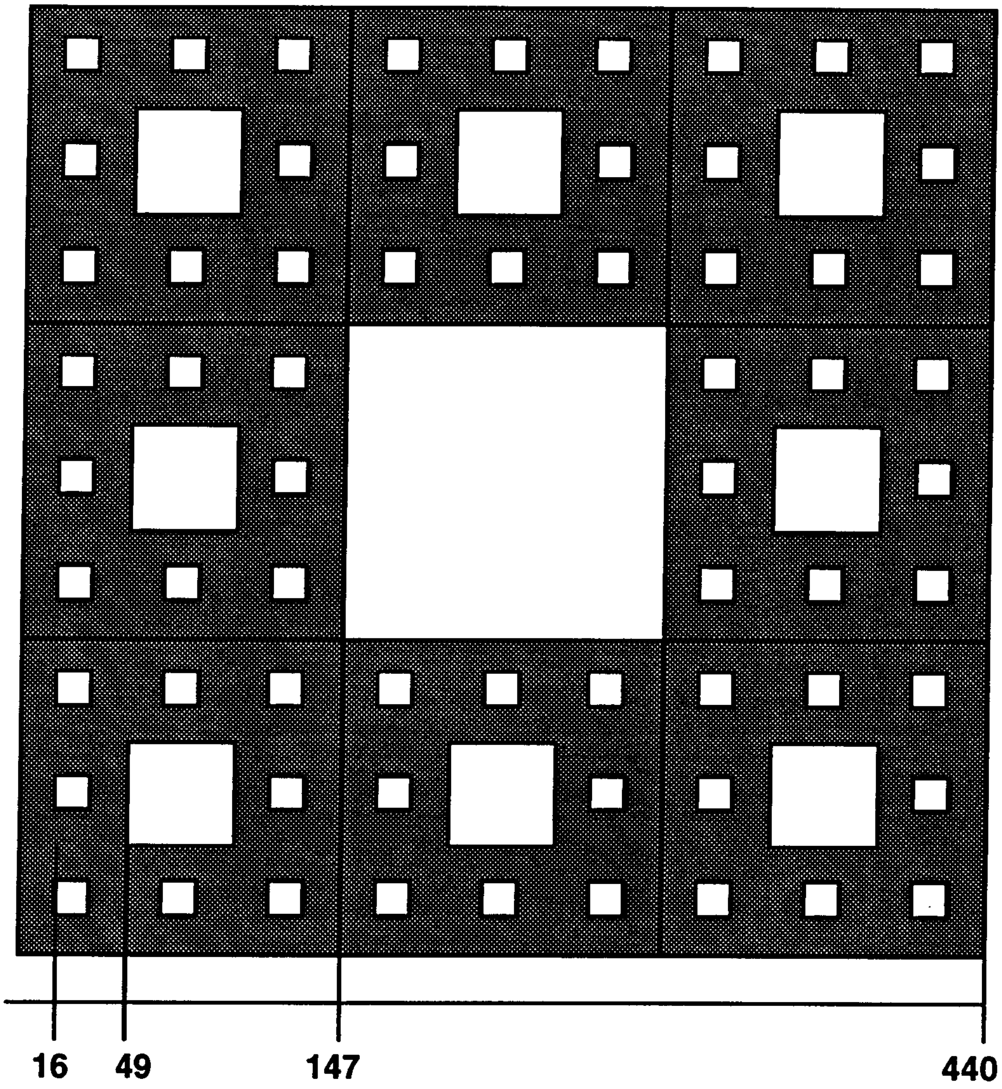


Figure 4.1 A 440x440-unit, level 3 Sierpinski carpet with the dimensions of intermediate components. Method 1 samples points randomly anywhere on the carpet.

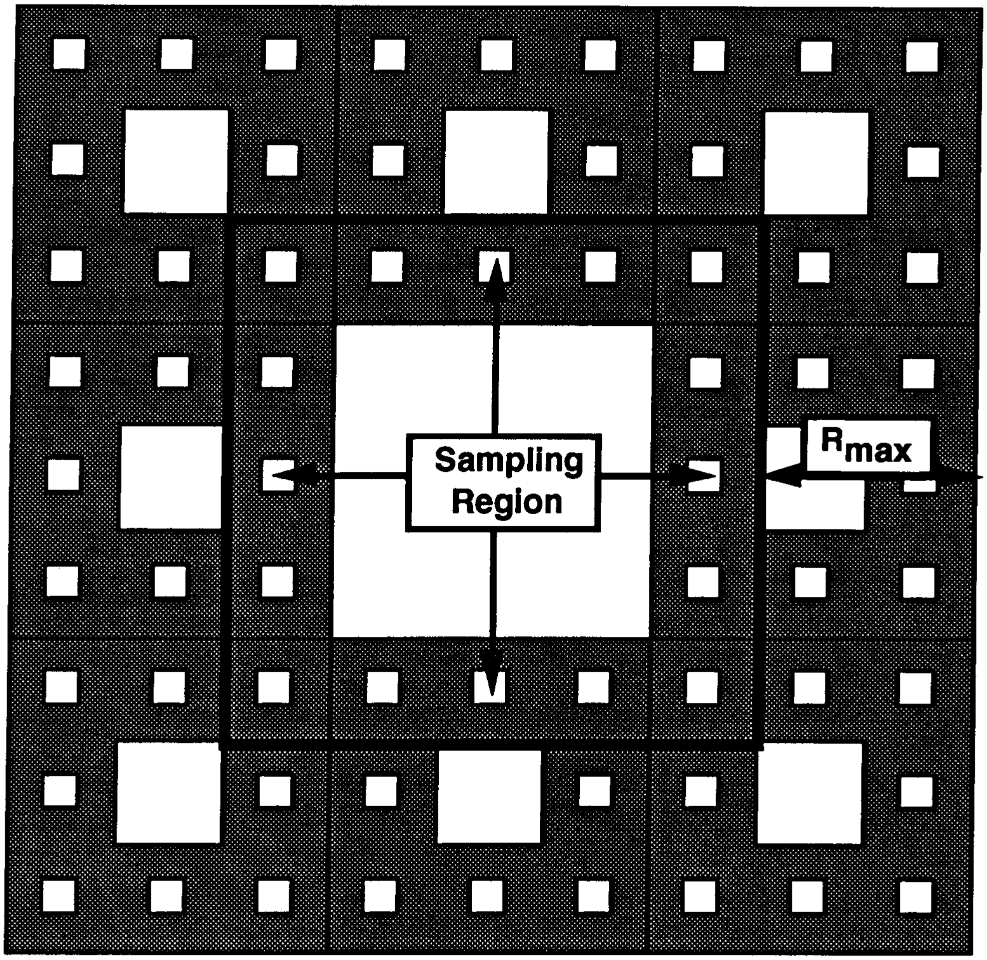


Figure 4.2 Method 2 samples in the middle of the carpet and avoids the outside edge completely but over samples the center edge.

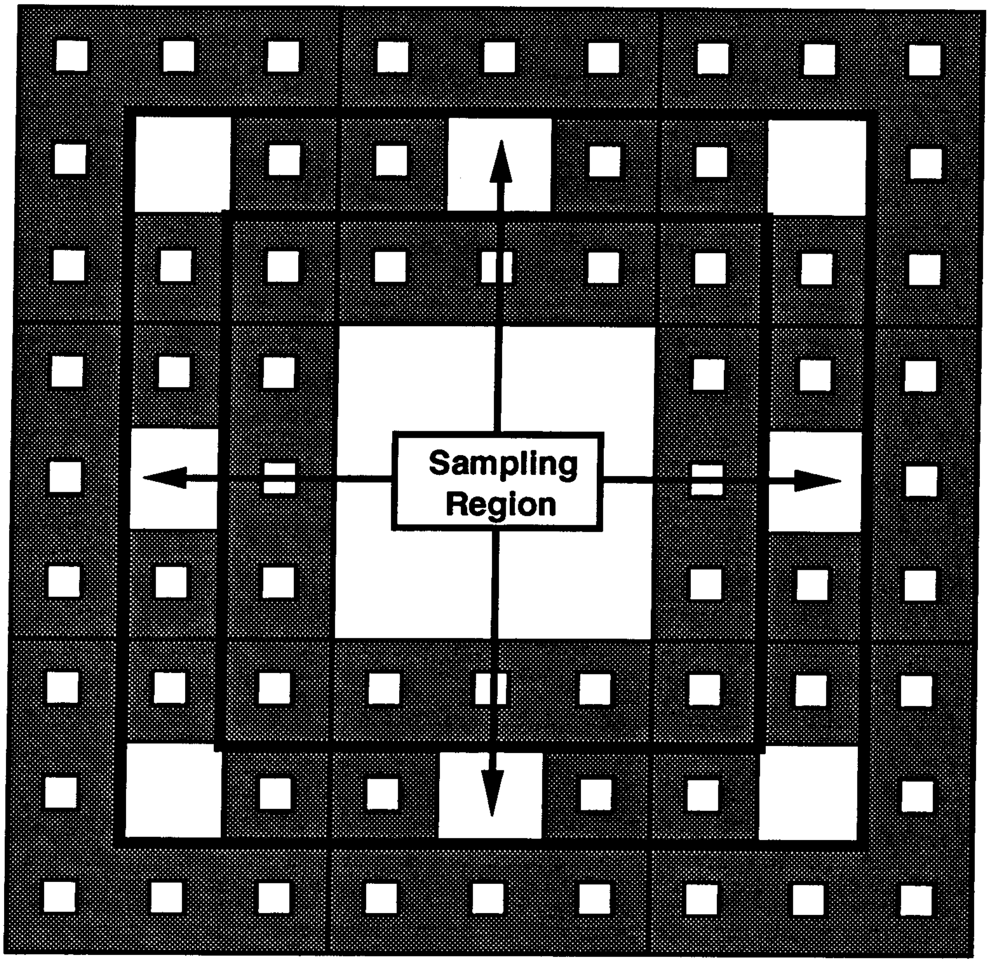


Figure 4.3 Method 3 samples points within a strip that is equidistant from outside and inside edge and avoids both the outside and center edge.

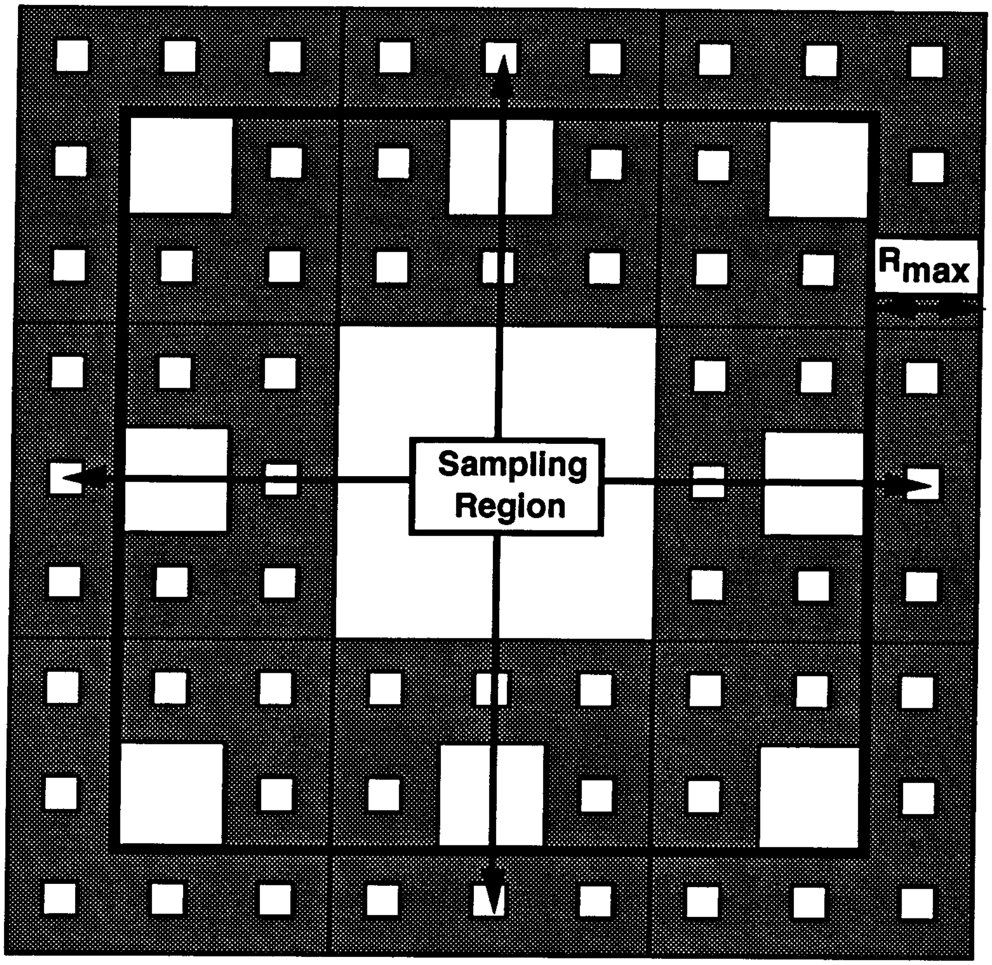


Figure 4.4 Method 4 samples points within R_{max} of the outside edge of the carpet and avoids the center edge completely.

The D_q spectrum was calculated for 10 random walks of lengths 50,000, 100,000, and 250,000 steps. As discussed in Chapter 3, the mean-square displacement of the path of a walker after n steps is a linear function of the square root of n (Equation 3.1). To apply a uniform scaling range to the random walks, the range used was $[2,R_{max}]$, where R_{max} is the average span of the walk along the four axes. Average span, like mean-square displacement, is a linear function of the square root of the number of steps in a walk and is defined in Equation 4.2.

Equation 4.2 Average span of a walk 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.

$$P = \frac{1}{d} \sum_{i=1}^{d} \left(i_{max} - i_{min}\right)$$

Random walks of various lengths were generated, 30 of each length, and then the mean-square displacement and the average span were calculated for each walk. Figure 4.5 demonstrates that average span and mean-square displacement both scale by the square root of the length of the walk. They differ only by a constant with average span slightly smaller than mean-square displacement. Therefore,, to compute D_q for random walks of various lengths, R_{max} is set to the average span of each walk.

Several different sampling rates were applied in the sandbox technique. Sampling rates above 2% did not significantly change D_q (data not shown). Therefore, in each case, 2% of the points in each random walk were sampled.

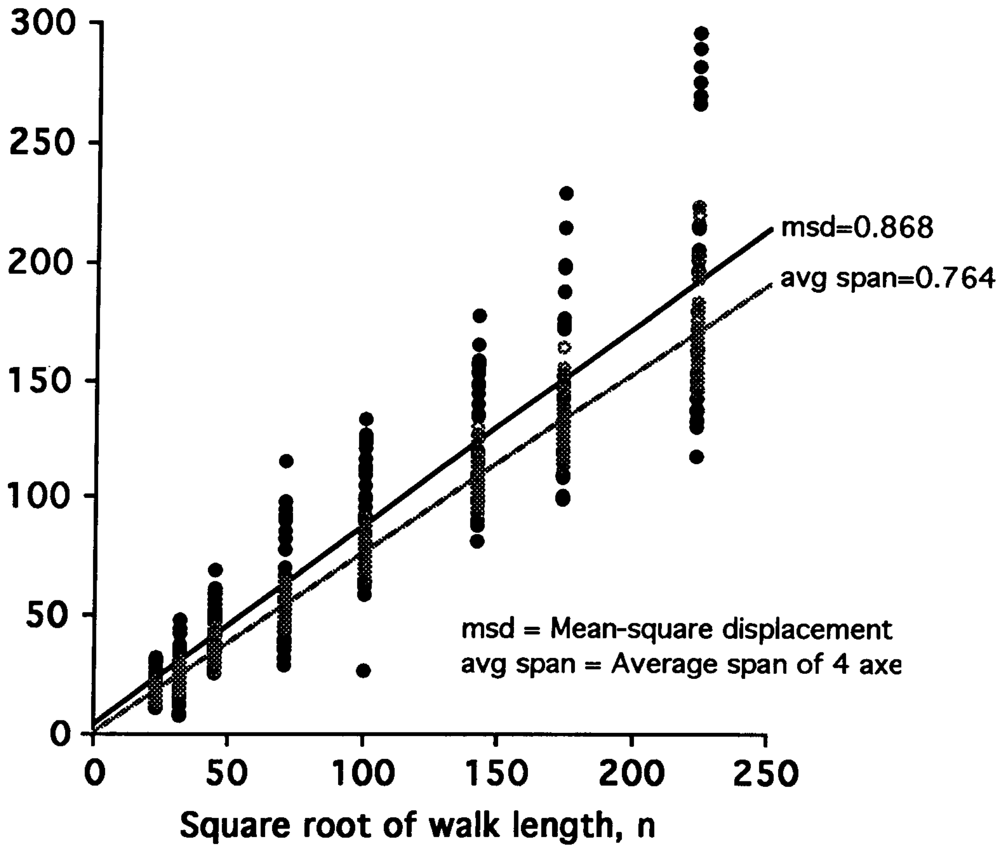


Figure 4.5 Average span and mean-square displacement of a random walk are both a linear function of the square root of n, the length of the walk.

4.3 Results

Results of a level 3 Sierpinski carpet using different sampling techniques were revealing. Sampling Method 1, which used randomly selected points anywhere on the carpet, produced an effective multifractal spectrum at R_{max} equal to the actual size of the fractal (see Figure 4.6). This correlates with the application of the average span for R_{max} for pseudorandom walks. If R_{max} is less than the size of fractal, the D_q curve is not monotonically nonincreasing. If R_{max} is greater than the size of the carpet, the D_q curve is monotonically nonincreasing but the D_q

values are decreased. The amount of decrease appears to be a linear function of R_{max}. The multifractal spectrum has significantly lower values than the theoretic fractal dimension because blank area beyond the edge of the carpet is included (i.e., running off the edge of the fractal). Also, the blank area at the center of the carpet is encountered.

Results of sampling Method 2, which randomly samples points that are R_{max} or more from the outside edge of carpet, are in Figure 4.7. This technique never runs off the outside edge of the fractal but over samples the center region.

Results of sampling Method 3, which sampled within a strip that was equidistant from the outside edge and center region, are in Figure 4.8. This technique runs off the edge of the carpet and samples the center region about equally. At R_{max}=50, which represents the distance of the edge of the sampling region from the outside edge and center, a multifractal spectrum is produced that is much closer to the theoretic value of D=1.893 than any of the other techniques. Smaller R_{max} produces D_q curves that are not monotonically nonincreasing and larger R_{max} increase the D_q values.

Results of sampling Method 4, which selects points that are within R_{max} of the edge, are in Figure 4.9. This method clearly fails to produce any D_q curve that is monotonically nonincreasing for any R_{max}. This method runs off the edge very inconsistently and never includes the center region. The D_q values do not approach the theoretic D=1.893 for any q either.

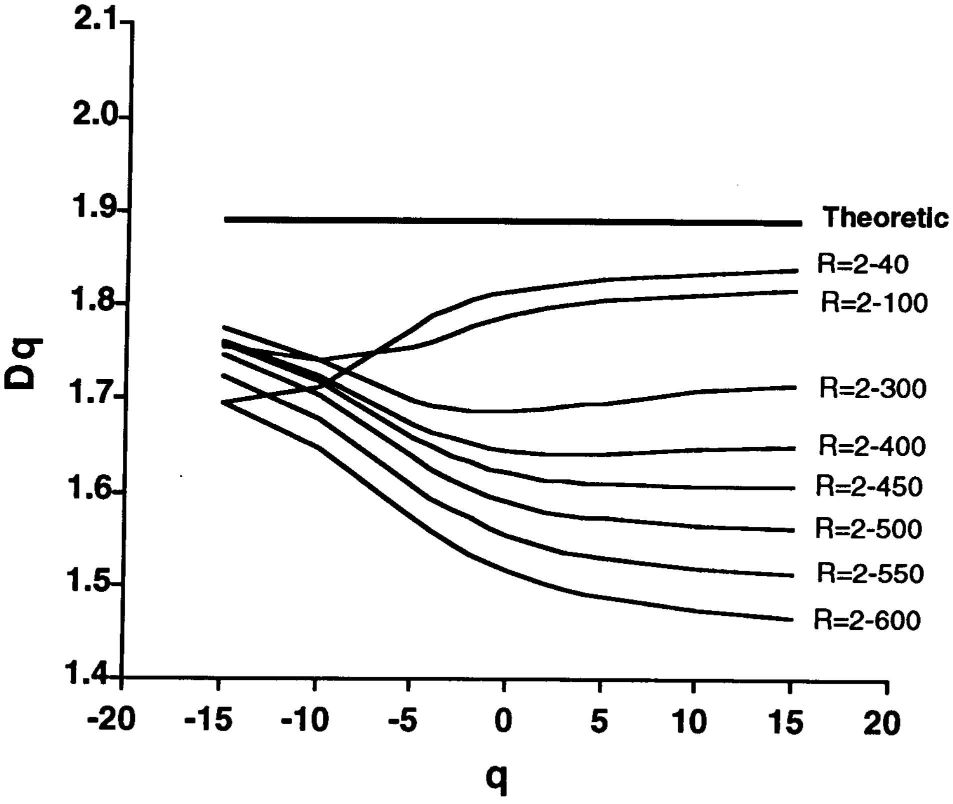


Figure 4.6 Multifractal spectra for a level 3 carpet of size 440x440 using Method 1, which randomly sampled points for different R_{max} values, reveals that monotonically nonincreasing spectra are obtained when R_{max} is equal to size of fractal (440 units) or larger. All spectra values are less than the theoretic value of D=1.893.

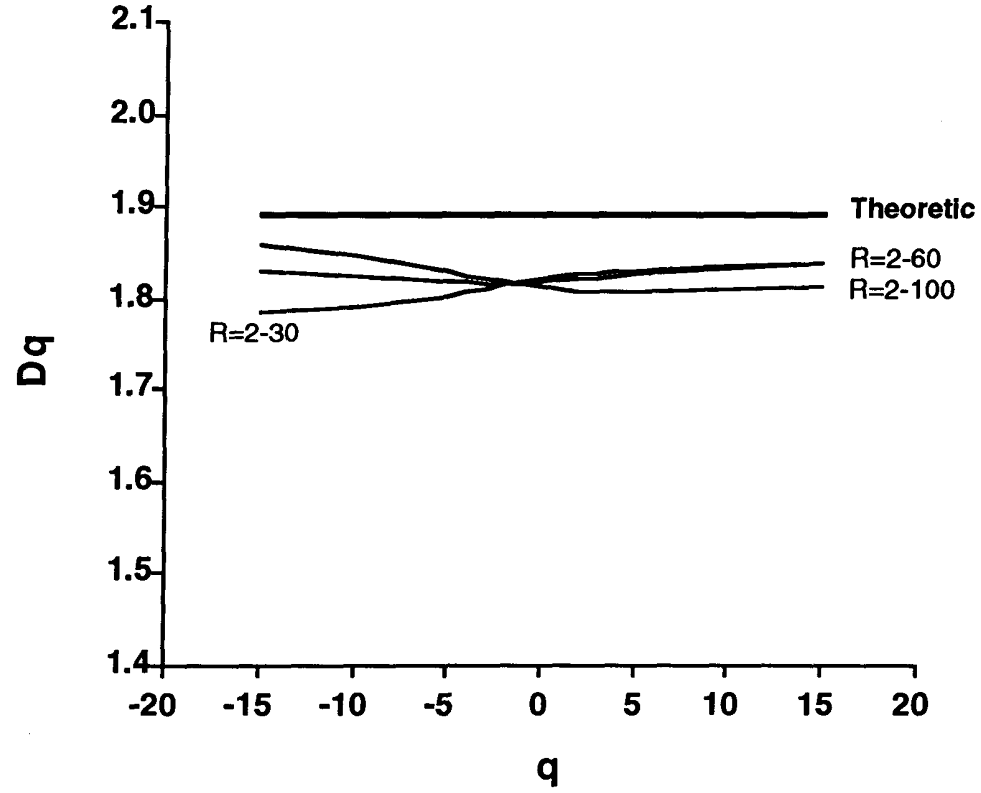


Figure 4.7 Multifractal spectra of a level 3 Sierpinski carpet using Method 2 sampling, which totally avoids the outside edge but over samples the center portion. R_{max}=100 is almost monotonically nonincreasing, and all points sampled were 100 units or more from the outside edge.

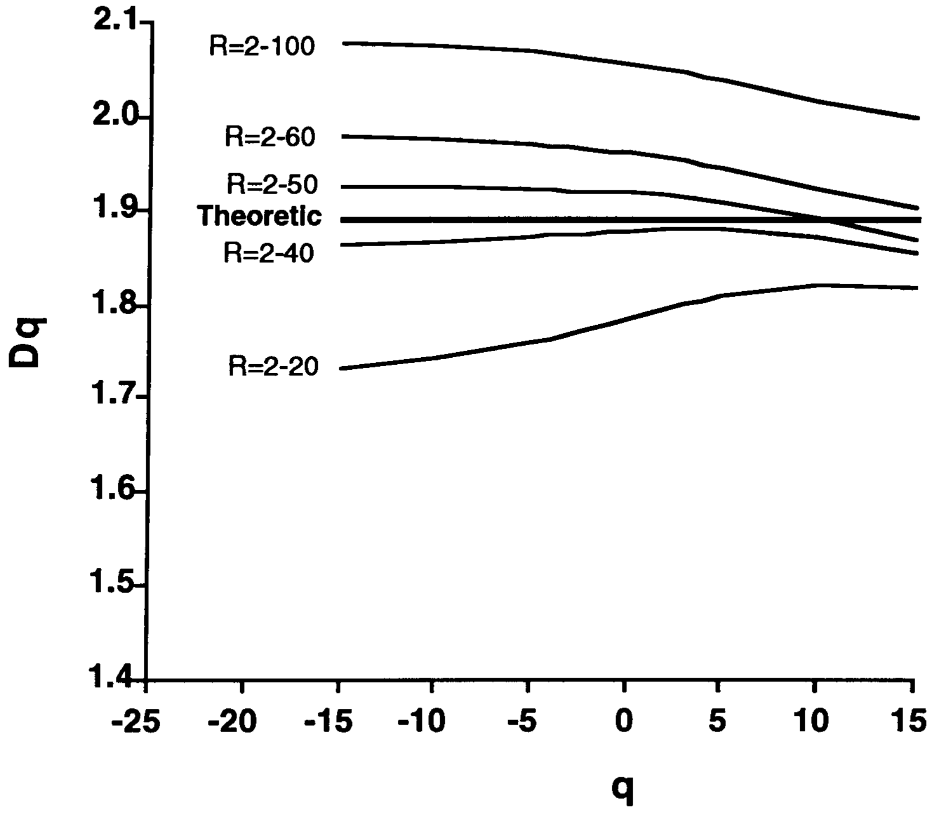


Figure 4.8 Multifractal spectra of a level 3 Sierpinski carpet using Method 3, which samples within a strip that is equidistant (50 units) from outside edge and center region, produces an effective spectrum that is close to the theoretic D=1.893 for R_{max}=50.

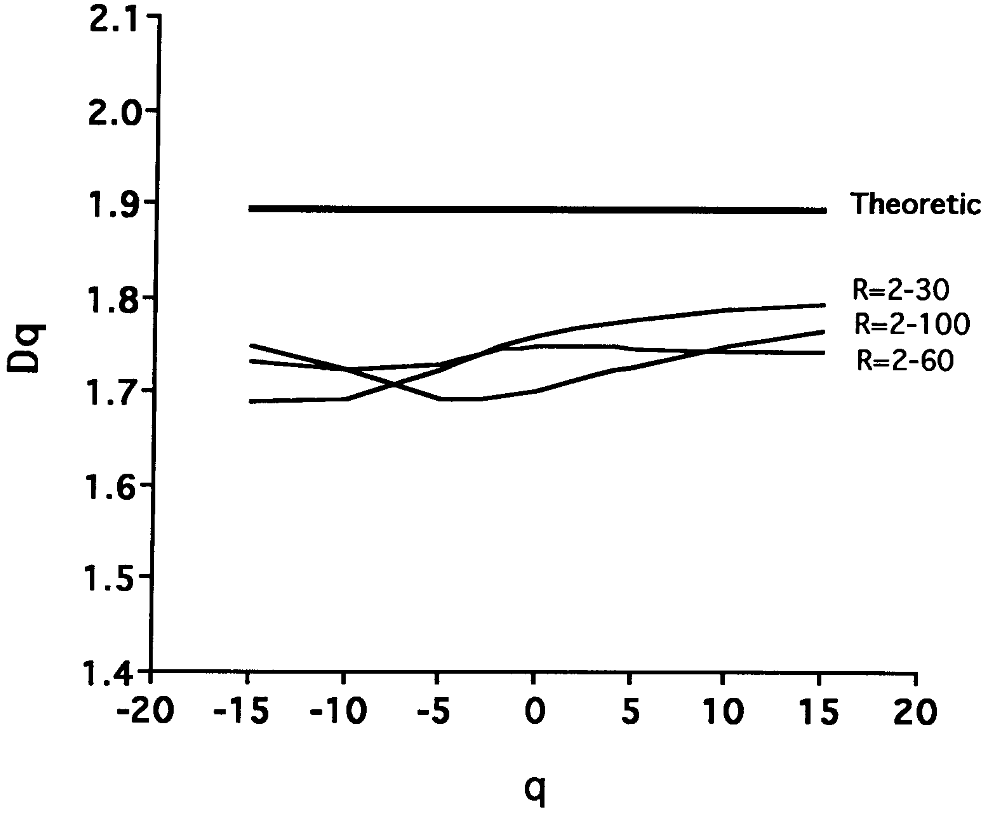


Figure 4.9 Multifractal spectra of a level 3 Sierpinski carpet using Method 4, which samples points within R_{max} of the edge, fails to produce a monotonically nonincreasing multifractal spectrum for any scaling range.

Figure 4.10 compares the multifractal spectra obtained by the four sandbox sampling techniques for $R_{max}=100$, which is about 23% of the size of this fractal. This demonstrates that different multifractal spectra are possible depending on how the sampling of points is done. Method 3, which avoids both the outside edge and center region, produces a monotonically nonincreasing D_q curve, which is elevated over the theoretical fractal dimension of the carpet at this R_{max}. Method 4, which samples along the outside edge, produces a poor D_q curve with very depressed values and is not monotonic. Method 1, which randomly samples points anywhere on the carpet, does not produce a monotonic D_q curve either. Method 2, which totally avoids outside edge but over samples the center region, produces a D_q curve that increases slightly at positive q at this R_{max}. However, the D_q values are all within 5% of theoretical fractal dimension of 1.893.

These simulations with the Sierpinski carpet reveal two important aspects of estimating the multifractal spectrum. First, the theoretical D_q curve is best approximated by sampling points that totally avoid both the outside and center edges using a narrow scaling range [2,40]. An effective multifractal spectrum,, which is monotonically nonincreasing, is achieved by randomly sampling points over the fractal and using a scaling range with R_{max} equal to size of the fractal. Scaling ranges smaller than this do not produce a monotonic D_q curves. These facts can now be applied to estimating the multifractal spectra for random walks.

The D_q curves were estimated for random walks of length 50,000, 100,000 and 250,000, 10 of each length. Over a very

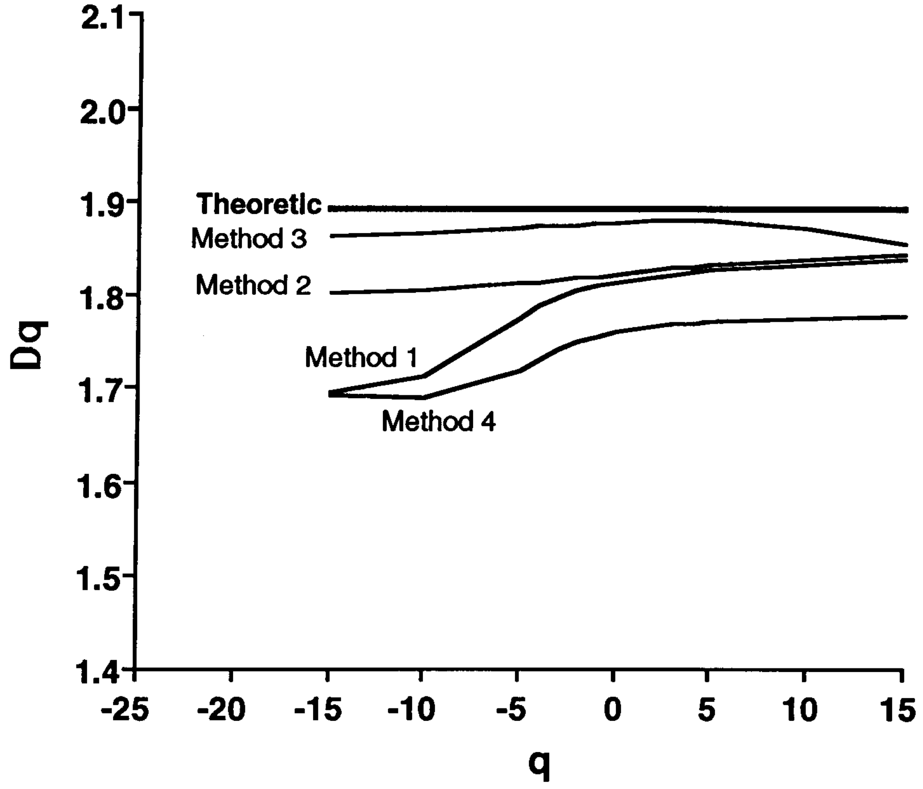


Figure 4.10 Multifractal spectra of a level 3 Sierpinski carpet, obtained by four sampling methods using scaling range of R=2-40, demonstrates the effect of outside and center edges on the calculated D_q. Method 3, which samples within a strip and totally avoids both center and outside edges at this scaling range, produces an almost monofractal D_q curve that is very close to the theoretic fractal dimension of this Sierpinski carpet.

limited scaling range, [2,20], the random walks of all three lengths are monofractal with the average D_q close to 2.0 for all q (Figure 4.11). The standard deviations are smallest at q=0 for all lengths and largest at the extremes, q=-15 and q=+15, with all standard deviations less than 5% (Gould and Tobochnik 1990; Viscek, Family, and Meakin 1990). Therefore, the repeatability of the result from walk to walk is good. In general, the longer the random walk, the smaller the standard deviations. Empirically, it was found that the shortest random walk to yield a valid monofractal spectrum was 4,000 steps with standard deviations of about 10%. However, this is over a scaling range [2,4] with all line fits based on just the changes in mass between radius=2 and radius=4.

A larger R_{max} examines the surface of a random walk as well as the bulk. The log/log plot in Figure 4.12 demonstrates the linearity obtained over a long scaling range of a 50,000 step random walk at q=-15.

When R_{max} is set equal to the average span of the walk along the four axes, the three D_q curves are identical in shape, and the values themselves depend on the length of the walk (Figure 4.13). The longer the walk, the lower the standard deviations obtained. It is interesting to note that the entire spectrum for 250,000 steps is above 2.0, the theoretical value for an infinite random walk.

Figure 4.14 reveals that there is very little difference in the average multifractal spectrum over a fairly wide scaling range. Here the 250,000-step random walks were scaled with R_{max} equal

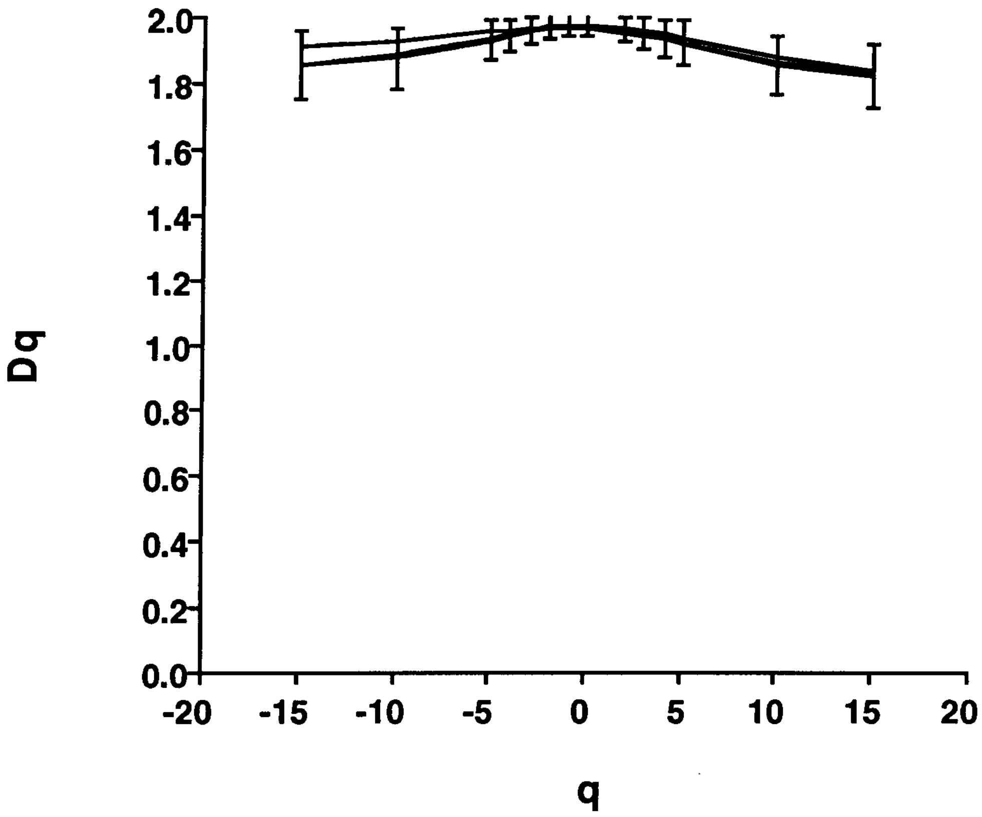


Figure 4.11 The average D_q curves for 10 random walks of 50,000, 100,000, and 250,000 steps are monofractal over a very limited scaling range [2,20] with fractal dimension close to 2.0. Error bars indicate the standard deviations obtained for 50,000-step walks.

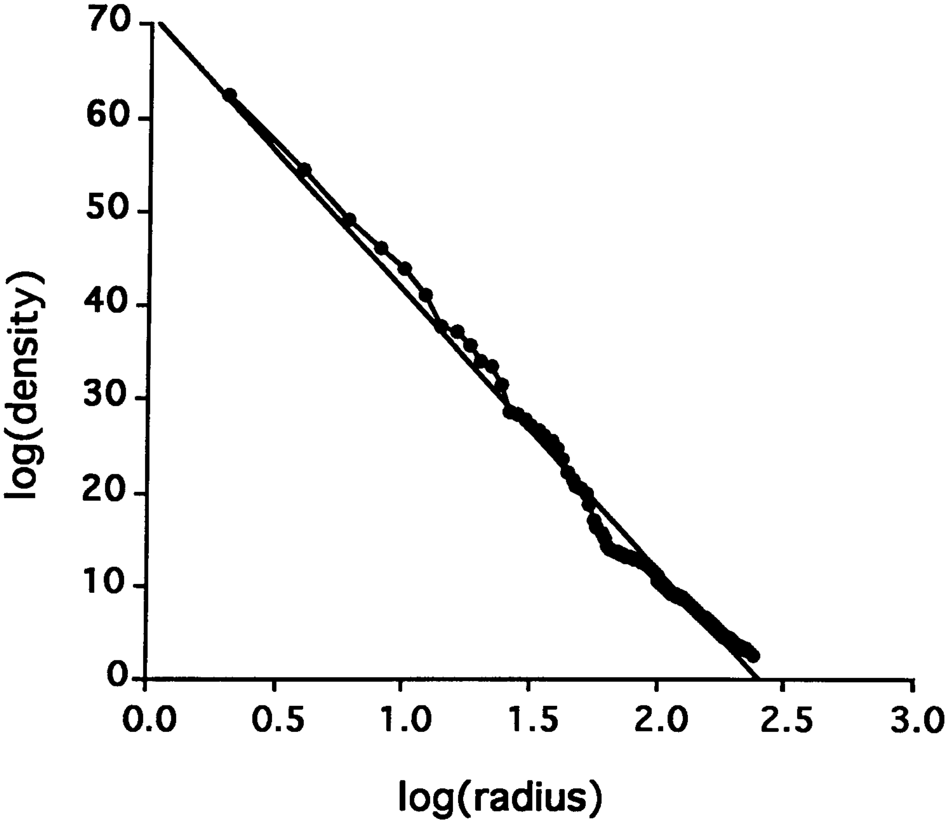


Figure 4.12 Log/log plot for a 50,000-step random walk using the sandbox algorithm for q=-15 shows a long linear scaling range. A line fit for this plot produces r^2=0.989 indicating a very good fit.

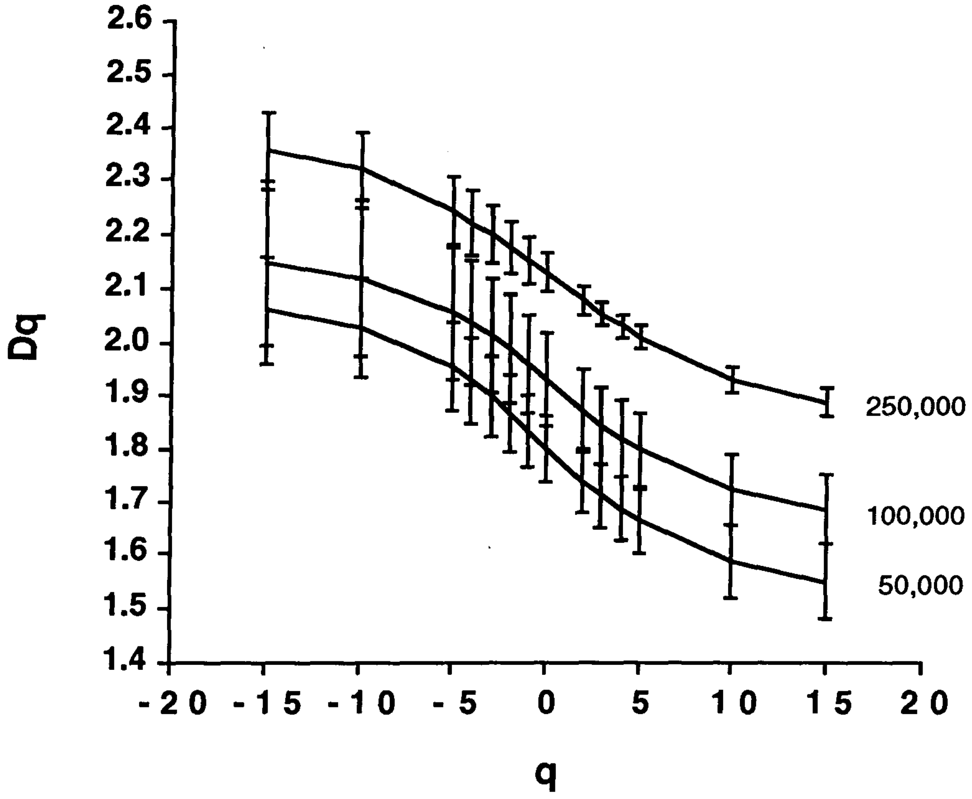


Figure 4.13 The average D_q curves for 10 random walks of length 50,000, 100,000, and 250,000 over a long scaling range are multifractal, and the spectra are unique for each length. The scaling range is $[2,R_{max}]$ where R_{max} equals the average span along the four axes for each walk length. R_{max} equals 160, 200, and 240 for the 50,000-, 100,000-, and 250,000-step walks respectively.

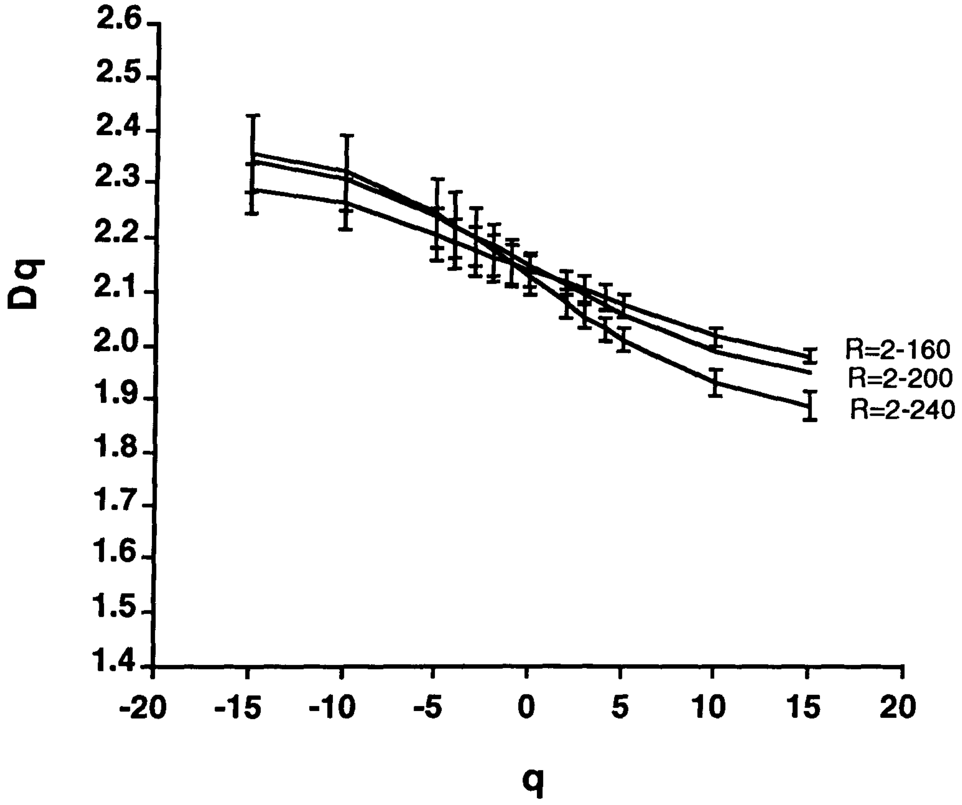


Figure 4.14 The average D_q curve for 10 random walks of 250,000 steps changes very little over a rather wide scaling range.

to 160, 200, and 240 respectively with minimal change in the D_q curve.

Figure 4.15 shows the multifractal spectrum obtained when R_{max} is set to incorporate half of the data points of the walk on average rather than average span. The curves are distinctly different with this scaling range and the spectrum for 250,000-step walk still exceeds the theoretical value of 2.0.

For standard box counting, the D_q curve does not converge for these parameters (Ramsey and Yuan 1989; Smith 1988). Also, the sandbox method appears to be computationally faster than box counting. Because the mean radius of a random walk scales with the square root of its length, D_q curves were calculated with R_{max}=100 for the 50,000 length walk, R_{max}=120 for the 100,000 length walk and R_{max}=224 for the 250,000 length walk. It was expected that these curves would be indistinguishable. These results show that the curves for 50,000 and 100,000 were within one standard deviation of each other for all q, with better agreement at negative q. The 250,000 curve was significantly different for all q values. This result appears to be caused by the inclusion of a significant sampling of the walks' surface. Both the surface and the bulk of a random walk scales as the square root of the length of the walk. However, the exponents are different (Figure 4.5).

A variety of sampling methods were used to calculate the multifractal spectrum of the Sierpinski carpet that include or exclude the center and/or outside edge of the carpet. This cannot be done for a random walk. Finding the outside edge of the walk is

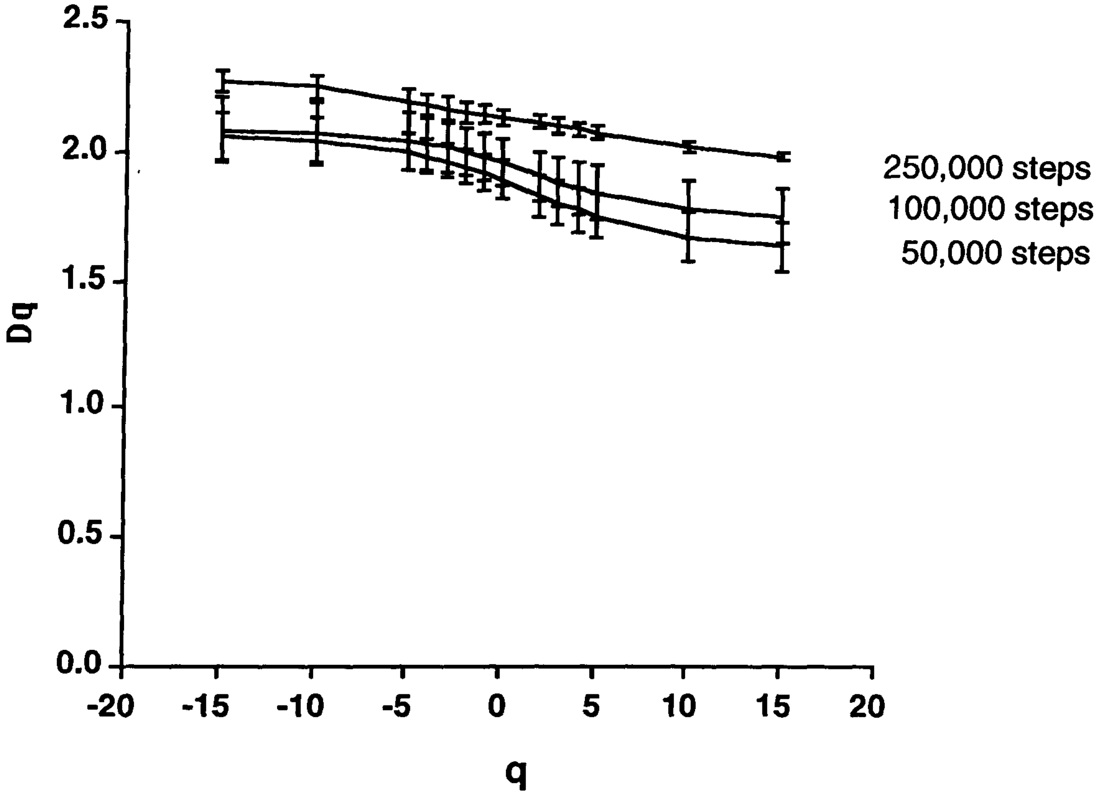


Figure 4.15 The average D_q curves for 10 random walks of length 50,000, 100,000, and 250,000 with scaling range $[2, R_{max}]$ where R_{max} is set so that the ball around each sampled point, on average, contains 50% of the data points of the walk. These D_q curves are also multifractal.

a convex hull problem that is computationally intractable. An infinite length random walk would essentially have no edge. Finite length random walks will invariably have an edge, and there is no way to totally avoid the edge in computing D_q. There has been some work on adjusting measures of fractal dimension for edge effect (Taylor and Taylor 1991). However, these adjustments apply to graphs of continuous functions when using box counting and do not apply for finite unit-step random walks. The random walk is also different from the carpet in that there will not be regions within the fractal without data points that have distinct edges. However, random walks will typically have holes within the bulk of the walk of varying sizes and shapes. This blank space within a walk is difficult to predict, and estimating the effect it has on multifractal spectrum is even more difficult. 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. It was found that the average span of the walk among the axes provides an adequate scaling range. Average span produces reproducible results among walks of the same length. It is important to note that these values are arbitrary but do provide a usable spectrum for comparison purposes.

Fractal analysis of DNA sequence data is the motivation for developing techniques to calculate multifractal spectrum. DNA sequences are mapped into four-dimensional pseudorandom walks according to mapping procedure and the axis scheme in Chapter 3 (Figure 3.14). The multifractal spectra of these walks are then computed using the sandbox algorithm with R_{max} equal to the average span. Three types of control sequences are then used for comparison: (1) random bases where each base has equal probability of occurring; (2) base-matched where the control sequences are generated using the probability of each base occurring according to proportion of bases in the DNA; (3) dimer-matched where the control sequences are generated using the probability of each dimer-pair occurring according to proportion of dimers in the DNA. The multifractal spectra of these control types are then calculated. Figure 4.16 shows a clear difference between the human beta globin gene and all three control types.

4.4 Conclusion

A unit-step random walk on a lattice is homogeneous over a scaling range that is very limited (R_{max} about 10% of average axis span). Scaling ranges larger than this yield a multifractal spectrum that may be used to examine the perimeter of the walk in addition to the internal structure. The sandbox algorithm gives converged spectra for much shorter random walks than does box counting. The problem of edge effect in computing the multifractal spectrum with the square Sierpinski carpet has been demonstrated. Because the edge of a finite length random walk cannot be avoided, methods must be used to obtain a converged multifractal spectrum that can be applied consistently. The ability to calculate a converged effective multifractal spectrum allows the use of the random walk as a typical model fractal for

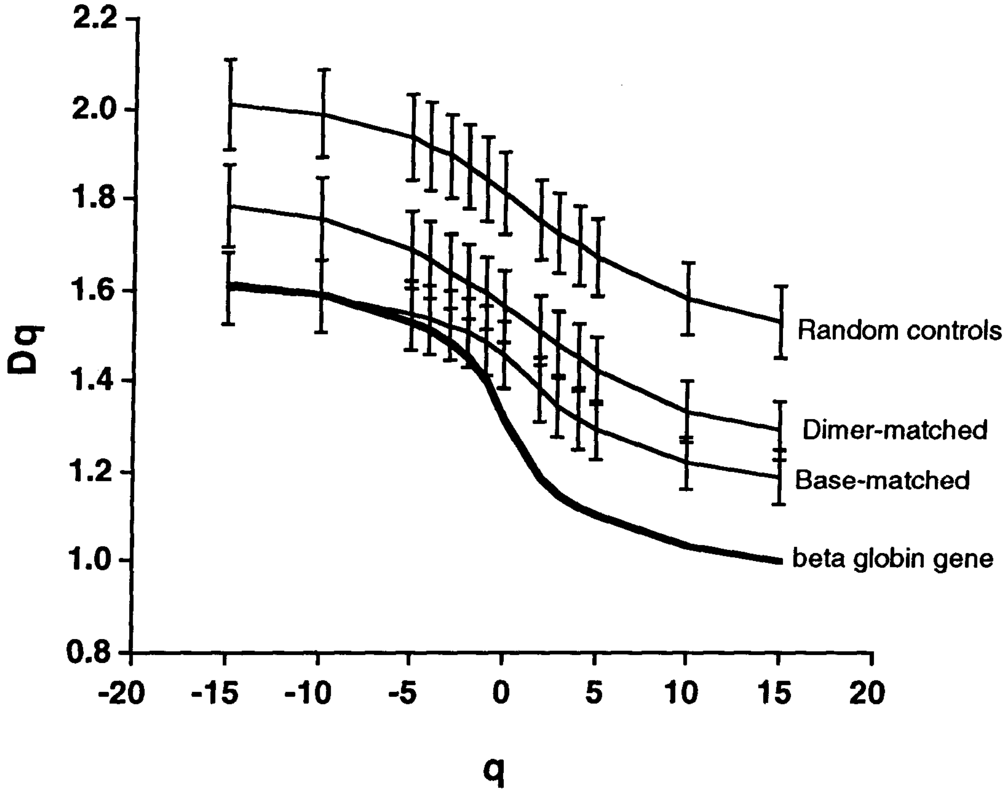


Figure 4.16 The multifractal spectrum of the human beta globin gene (length=73,326 base pairs) converges at R_{max}=700. Significant differences exist between its D_q curve and the average curves of random, base-matched, and dimer-matched control sequences.

comparison with short experimental data sets such as DNA sequences mapped into pseudorandom walks.

5. MULTIFRACTAL SPECTRA DISTINGUISH VERTEBRATE mtDNA SEQUENCES FROM INVERTEBRATE mtDNA SEQUENCES

5.1 Introduction

In Chapter 3, I have shown that the global fractal dimension is useful in the study information content. However, the global fractal dimension produces a measure averaged over the entire time series. Important local patterns may be lost or masked with this averaging. The multifractal spectrum reveals more about localized patterns and may provide more information about internal organization of DNA sequences than just the global fractal dimension.

In Chapter 4, I have shown, using pseudorandom walks, that relatively long DNA sequences (minimum of 15,000 base pairs for nonrandom sequences) are needed to calculate a multifractal spectrum. A number of animal mitochondrial genomes have been totally sequenced. A total of 12 complete genomes, 4 invertebrates and 8 vertebrates, were found in GenBank. Animal mitochondria are typically 15,000-20,000 base pairs in length and have a high rate of mutation. Therefore, mitochondrial DNA (mtDNA) sequences are good material for the application of fractal analysis and the exploration of fractal dimension and information content. With mtDNA, the entire genome may be efficiently analyzed rather than minute segments of larger genomes. Furthermore, they are all about the same length so the estimated fractal dimensions can be compared directly without the need for control sequences.

Information content and entropy in DNA sequences have been studied before. Subba Rao (Subba Rao, Geevan, and Subba Rao 1982) observed that mutations in human hemoglobin genes tend to occur such that the frequency of a codon that mutates is greater than the frequency of a codon to which it mutates. He concluded that the codon frequency distribution should be more equiprobable after the mutation than before it. Thus, the entropy of a coding region of DNA should be a nondecreasing function of the number DNA generations.

Konopka (1984) disagreed with Subba Rao arguing that the entropy measure, H, assumes that all codons are equally probable regardless of length or origin. Therefore, it depends on relative codon usage frequencies and is independent of the length of a DNA coding region and is only useful for comparing genes coding the same polypeptide in the same genome or across species. He proposed a function, D, which adjusts for genetic code degeneracy, and found that the D value for human mitochondrial genes was almost the same as for human nuclear genes even though mitochondria mutate at a much faster rate. The value of D for the average mitochondrial gene was greater than the value for the corresponding human nuclear genes. He found the same results on comparing histones, which are slow-evolvers, and globins, which are fast-evolvers. He concluded that entropy is not a good indicator of evolutionary differences and that information content does not tend to increase with evolution.

Rowe (1983) studied the information content of viral DNA. He found that viral DNA is a Markov chain with memory of two, and that most of the structure is on the level of pairs and triplets with little or no structure on levels 4 and 5. Noncoding regions of viruses have a nonrandom structure, often containing a higher level structure than the surrounding genes. He did find some correlation between levels of information storage and virus families. A strong codon bias exists in viral genes, and genes that code for structural proteins often showed stronger triplet correlations than other genes. He concluded that new ways of detecting and measuring information storage are needed, particularly for long range correlations.

At least one study has been published that used fractal dimension to study information content in DNA sequences. Luo and Tsai (1988) used fractal dimension to study its relationship to evolutionary level. He found that the average fractal dimension (AFD) grows gradually with increasing evolutionary level and suggested that this represents randomization of vocabulary composition of genetic language perhaps due to random drift. He also noted an increased correlation of neighboring bases and suggested that this represented the clarification of grammatical construction of genetic language, which may be a result of natural selection. He estimated the fractal dimension of mammalian mtDNA at D=1.154, and found this to be significantly lower than nonmammalian mtDNA and eukaryote viruses.

5.1.1 Characteristics of mitochondrial genomes

Mitochondria are small organelles found in the cytoplasm of eukaryotic cells that produce energy for the cell. They exist as complete, compact genomes and carry out their own DNA replication, DNA transcription, and protein synthesis. Animal mitochondrial genomes form a circular double helix composed of 15,000-19,000 base pairs. Each organelle may have 5-10 of these DNA molecules.

The genome is almost entirely coding sequence with just 5% of the genome making up the displacement loop (D-loop), which does not seem to code for protein but may perform regulatory functions for structural genes. The genome codes for 2 ribosomal RNAs (12S and 16S rRNA subunits), 22 tRNAs for protein synthesis, and 13 proteins (3 cytochrome oxidase subunits, 7 NADH dehydrogenase, ATPase6, ATPase8, and cytochrome b). No introns have been found in mitochondria and there are very few intragenic bases, if any.

Mitochondrial genomes of animals are highly conserved, showing 50-90% homology for coding regions. The noncoding D-Loop is very divergent with almost no homology. However, there seems to be secondary structure homology in the D-Loop despite the lack of sequence homology. The D-Loop typically contains the origin of replication. Gene organization and arrangement are identical for the mitochondria of human, cow, rat, mouse, and frog. The chicken mitochondria genome has a simple translocation affecting only four genes. Invertebrate mitochondria do not show this conservation of genome organization. Yet, the coding DNA shows the same high level homology with other animals.

The mitochondrial genetic code is very much like the universal genetic code but does show some differences. Interestingly, the codon differences seem to be organism specific. For example, the codon UGA, which is a STOP codon in the universal code, specifies tryptophan in mtDNA. AGA and AGG, which specify arginine in the universal code, specify a STOP codon in mammalian mtDNA and serine in drosophila mtDNA. With only 22 tRNAs, mitochondrial genomes show a high level wobble in the third base position of codons. Codon preferences exist but are species specific and usually reflect the base composition of the genome. For example, drosophila mtDNA is 77% A+T and 94% of all codons end in A or T. These variations in genetic code suggest that random drift has occurred in the genetic code of mitochondria.

In animals, mtDNA is transcribed at the same rate from a single promoter region on each strand, producing two different giant RNA molecules, each containing a full-length copy of one DNA strand. Transcription is completely symmetrical. The transcripts made on one strand, called the heavy strand (H strand) are extensively processed by nuclease cleavage to yield the 2 rRNAs, 14 of the 22 tRNAs, and about 10 poly-A-containing RNAs.

The light strand (L strand) transcript is processed to produce only eight tRNAs and one small poly-A containing RNA. The remaining 90% of this transcript does not appear to contain any useful information (except being complementary to the coding sequences on the H strand) and is degraded.

Mitochondrial genomes show a high rate of mutation with a rate of nucleotide substitution 10 times that of nuclear genomes. This high mutation rate makes mtDNA a good subject for evolutionary studies.

5.2 Methods

Twelve complete mitochondrial genome sequences were found in GenBank. Information on base content of the eight vertebrate and four invertebrate mitochondrial genomes is in Figure 5.1. The invertebrate group consisted of mtDNA sequences from two species of sea urchin (Paracentrotus lividus and Strongylocentrotus purpuratus), Drosophila yakuba, and Leishmania tarentolae. The vertebrate group consisted of mtDNA sequences from human, chicken, frog, mouse, rat, cow, fin whale, and carp.

Using the method described in Figure 3.14, the 12 mtDNA sequences were mapped to pseudorandom walks in four-dimensional embedding scheme. The multifractal spectrum was calculated using the sandbox method (Equation 2.7). Results in Chapter 4 revealed that converged and reproducible D_q curves were possible when the scaling range was adjusted to the behavior of the individual walk. The scaling range applied to each walk was

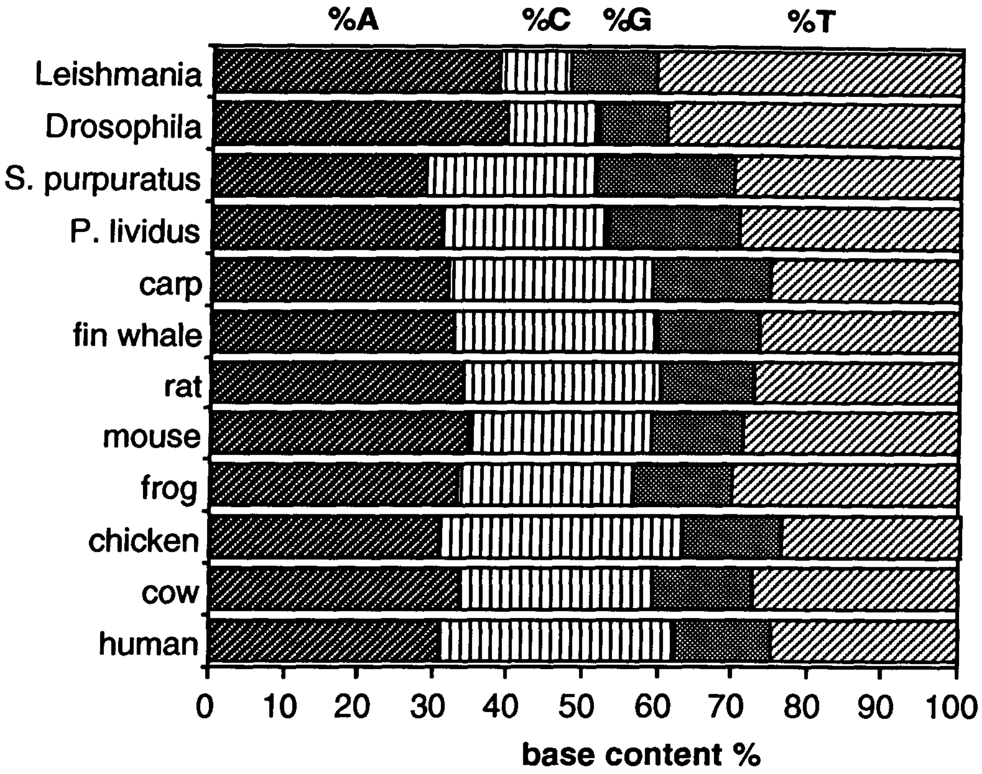


Figure 5.1 Base composition of the heavy strand of eight vertebrate and four invertebrate mitochondrial genomes.

its average span over the four axes, which was defined in Equation 4.2.

Previous researchers have used entropy to evaluate evolutionary levels, so entropy measures were calculated for the base, dimer, trimer, 4-mer, 5-mer, and 6-mer composition of each genome. Entropy, S, is a measure of disorder in a system. For a perfectly ordered system S=0 and for a totally random system S=1. In the case of dimer frequencies, S measures the amount of divergence from the uniform distribution of dimers. This is defined in Equation 2.11.

5.3 Results

The two-dimensional projections of the pseudorandom walks of the 12 mtDNA were very revealing. Virtually all of the vertebrate genomes show the same basic walk (Figures 5.2 and 5.3) even though they differ somewhat in base content. The high level of sequence homology is quite obvious with this graphic representation. The invertebrate genome walks (Figure 5.4) are dramatically different from the vertebrates. It is not surprising that the two sea urchin species have walks that look a lot alike. Drosophila and Leishmania genome walks are grossly very different even though their base contents show the same trend--excess A+T and deficient in C+G.

Predictably, the multifractal spectra for the vertebrates are all very close, reflecting the linear character of these walks. The multifractal spectra of the invertebrates also clustered together. In Figure 5.5, the multifractal spectra of the vertebrates are distinctly different from invertebrates.

Entropy measurements for the 12 genomes were calculated from base content up to a word size of 6. Figure 5.6 shows that all the vertebrate except the fin whale have basically the same entropy for sequence composition. The invertebrates, however, show two distinct patterns. The two sea urchins have entropy measures much closer to 1.0 than the vertebrates demonstrating

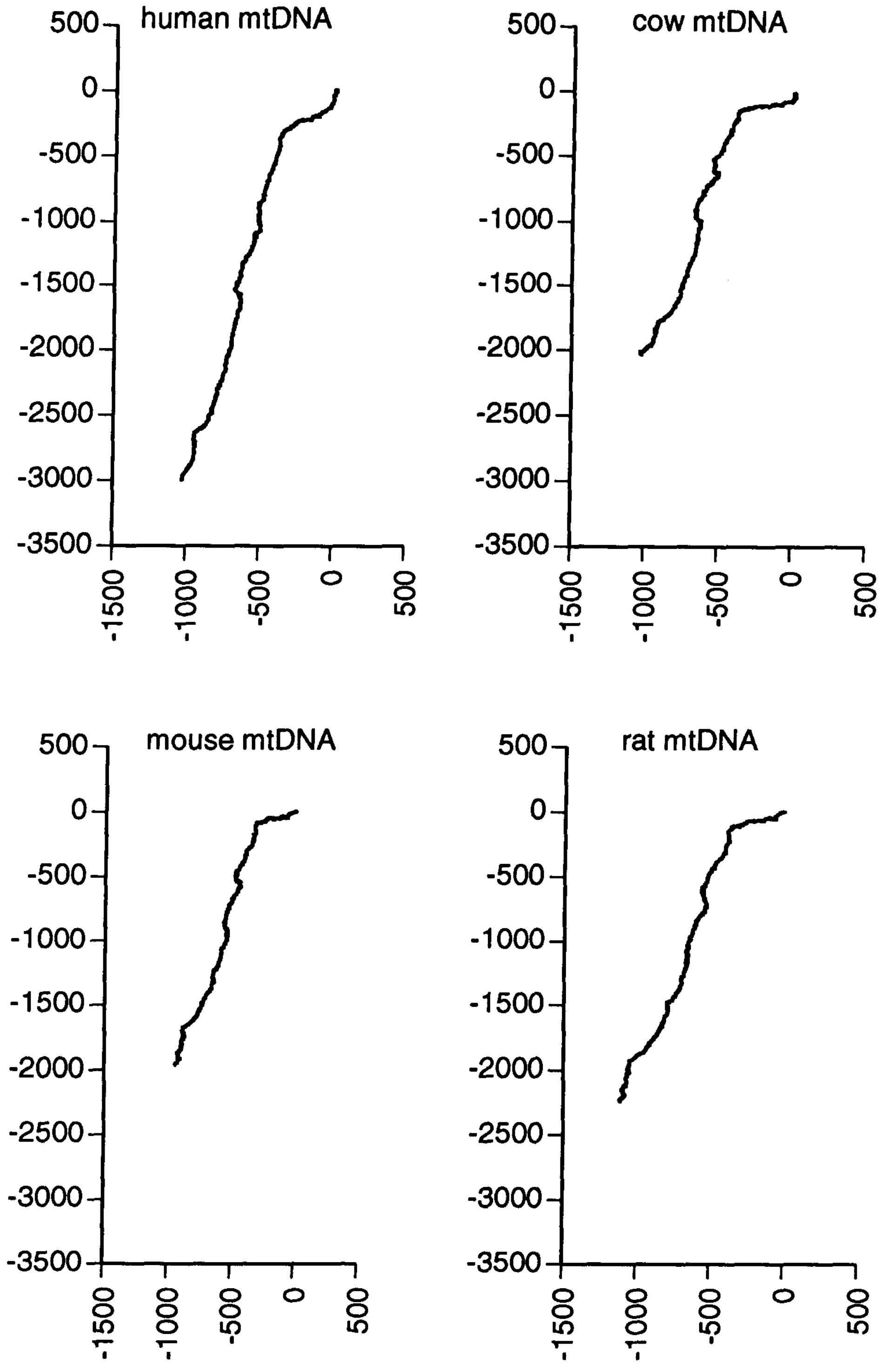


Figure 5.2 Two-dimensional pseudorandom walks of four vertebrate mtDNA genome sequences.

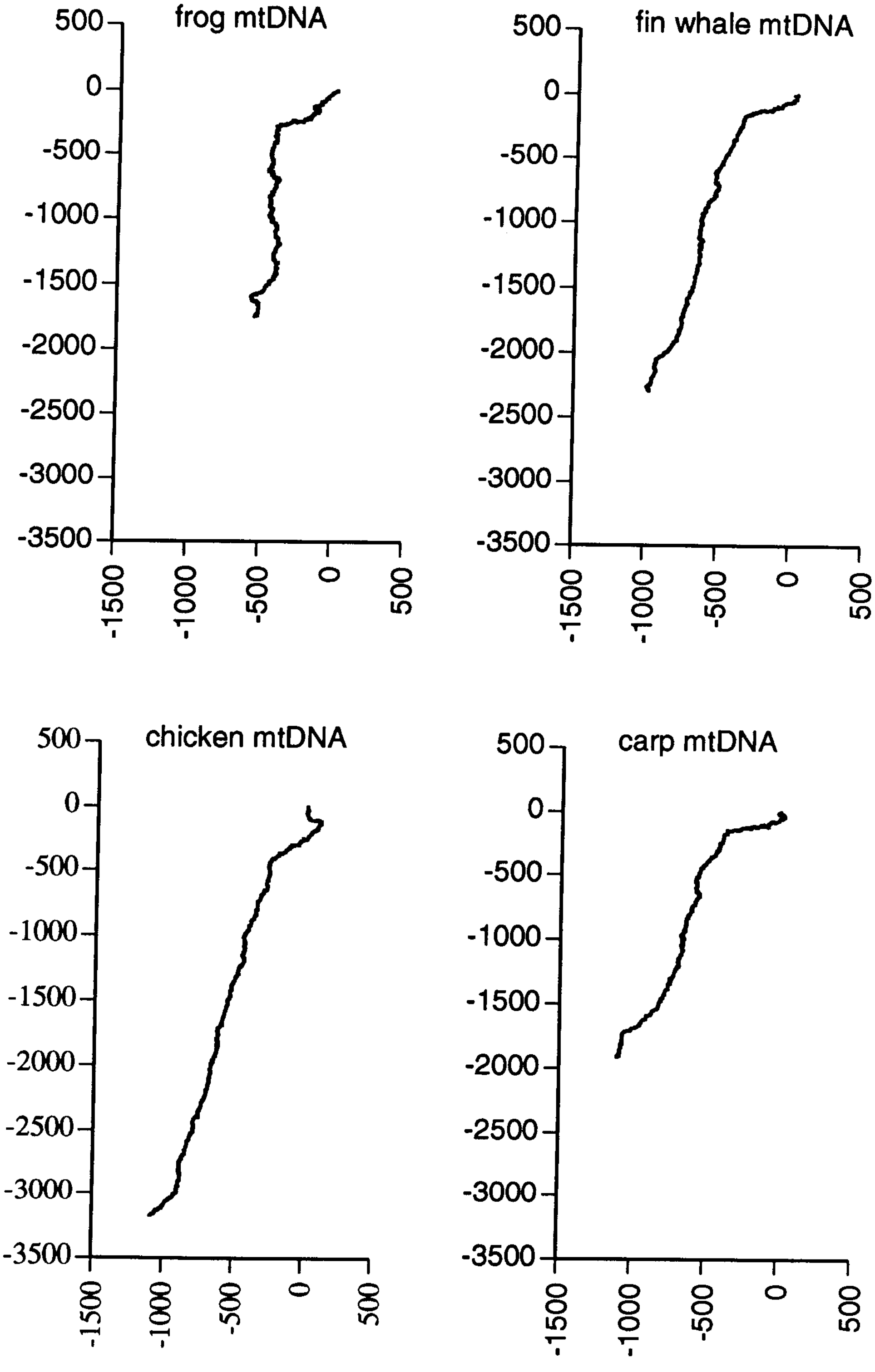


Figure 5.3 Two-dimensional pseudorandom walks of four additional vertebrate mtDNA genome sequences.

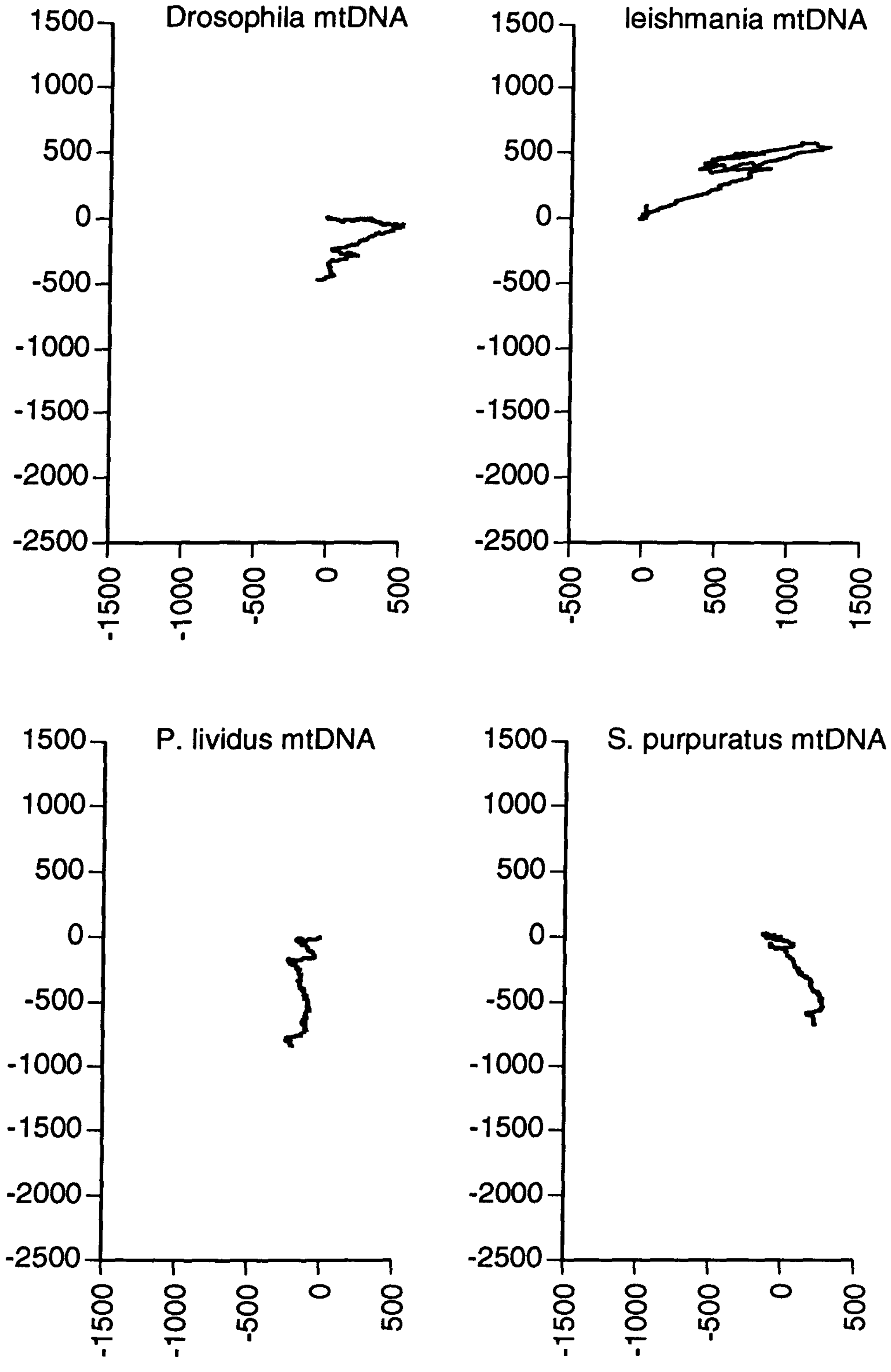


Figure 5.4 Two-dimensional pseudorandom walks of four invertebrate mtDNA genome sequences.

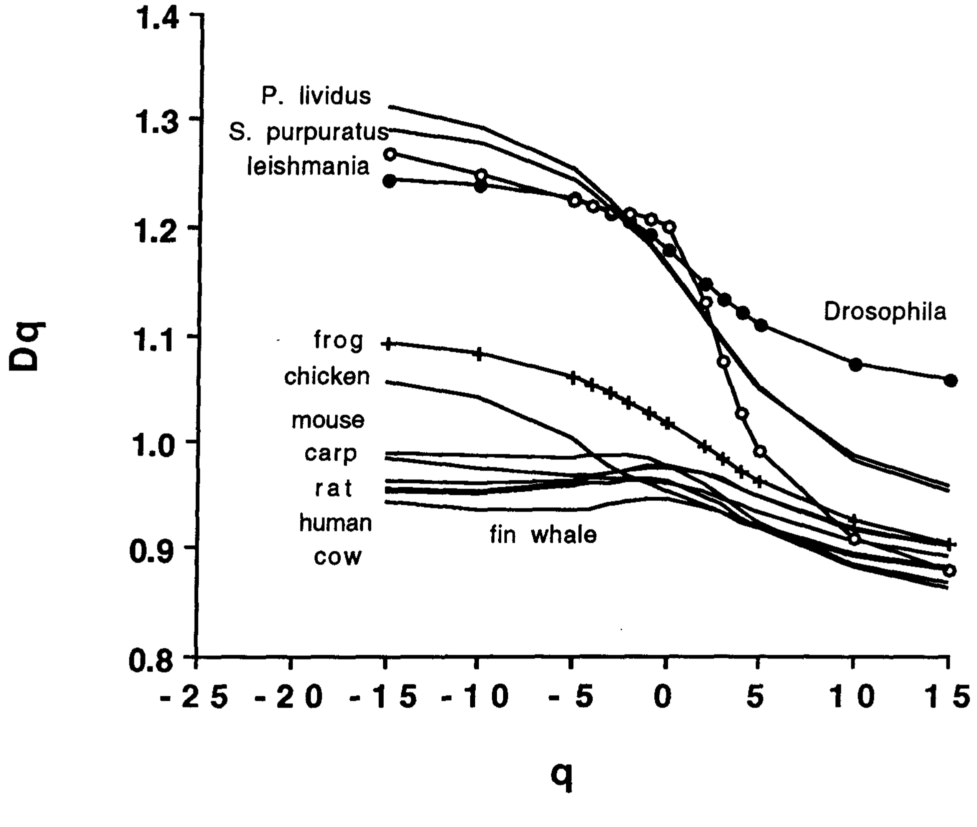


Figure 5.5 Multifractal spectra distinguish vertebrate mtDNA sequences from invertebrate mtDNA sequences.

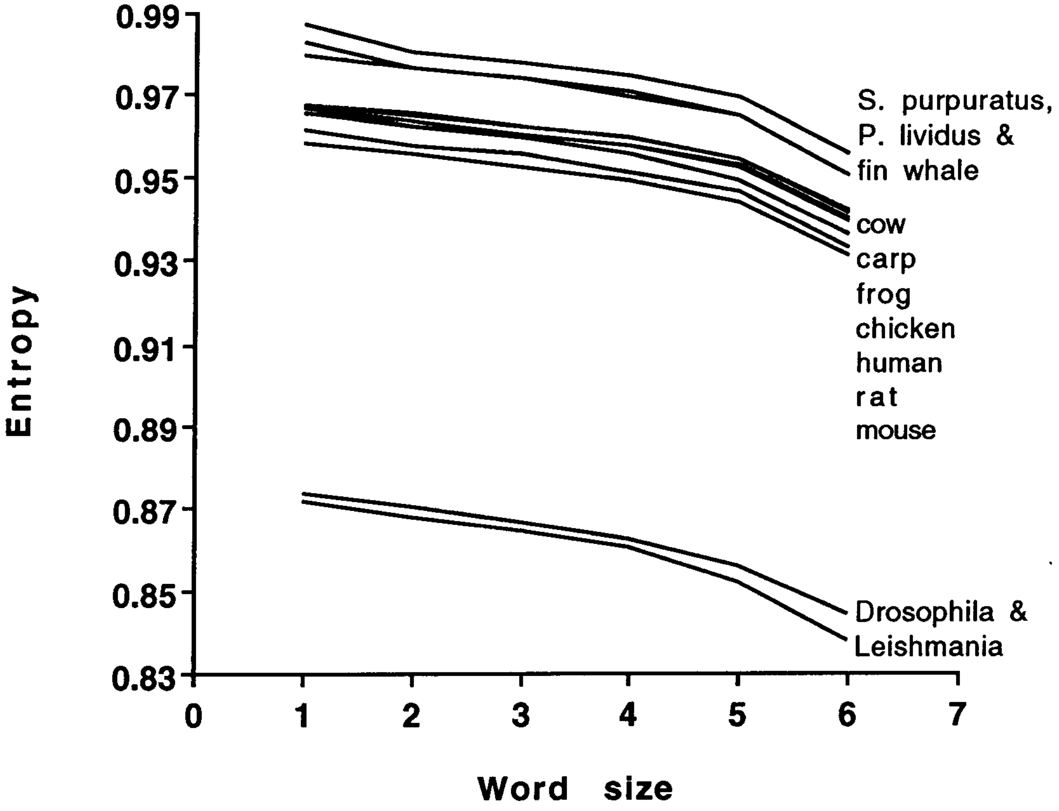


Figure 5.6 Entropy measurements of composition of mtDNA sequences by word size indicate that all but the fin whale have essentially the same entropy. The invertebrates show two distinct patterns with the fin whale most like the sea urchins.

more inherent disorder. Entropy for the fin whale also shows the same level of randomness as the sea urchins. Drosophila and Leishmania, on the other hand, have much lower entropy measures than the vertebrates, demonstrating more inherent orderliness in content. Despite the difference in entropy, however, the multifractal spectra for the invertebrates appear the same.

Vertebrate mitochondria have virtually the same gene order and organization, except chicken has a simple translocation. Invertebrate genomes have very different gene orders from the vertebrates as well as the others in the group. To study what effect gene ordering had on multifractal spectra, the invertebrate mtDNA was rearranged to match the gene order of the vertebrates. The multifractal spectra for the rearranged invertebrate sequences are the same as that for their natural ordering. Only 37 different genes are present in mtDNA, so this rearrangement would move large chunks of about 500 base pairs, which does not disrupt the underlying long-range correlations. Thus, gene order and organization are not the explanation for the differences between the multifractal spectra of the two groups (Figure 5.7).

Chi-square tests were performed using the base and dimer content data for the two groups of organelles. The base and dimer content of vertebrate genomes is significantly different from the invertebrate group ($p<.0001$). The genomes of vertebrate group are significantly different from each other ($p<.0001$) as are the invertebrates ($p<.0001$). Statistically speaking, the unique

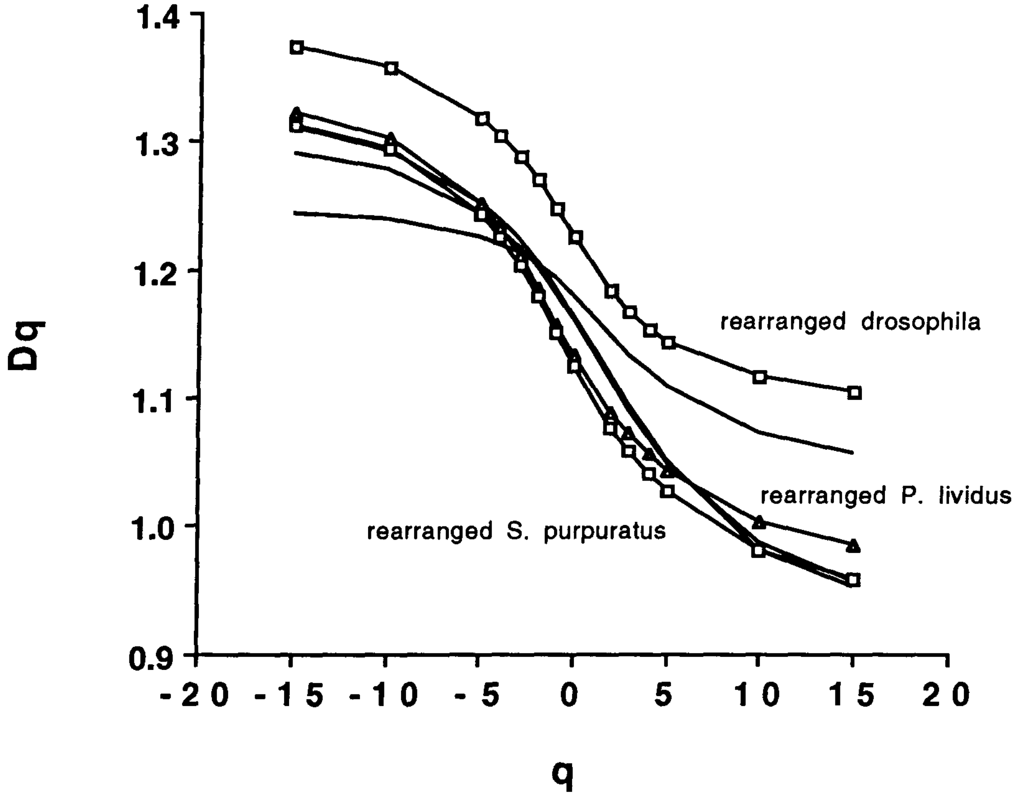


Figure 5.7 The multifractal spectra of invertebrates do not change when mtDNA is rearranged to the same gene order as vertebrate genomes.

groupings revealed by multifractal spectra cannot be explained by base or dimer content.

5.4 Discussion

The multifractal spectra of mtDNA sequences mapped into four-dimensional random walks reveal that invertebrate genomes are more randomly organized than vertebrate genomes. Long-range correlations in vertebrate mtDNA produced lower multifractal spectra for all vertebrates. The difference is not explained by base and dimer frequency differences since the groups show statistically significant differences among themselves. It is not explained by entropy measures of word content since the invertebrate group shows two divergent patterns of nonrandomness and the fin whale shows the same level of disorder as the sea urchins. Although the genomes of the invertebrate group have a very different gene order and organization, the difference is still present when the genomes are rearranged to match the mammalian mitochondrial genome order.

The multifractal spectra of mtDNA reveal the presence of long-range correlations that are significantly nonrandom. Vertebrate mtDNA sequences show more long-range correlations than invertebrate genomes. The lower multifractal spectrum for vertebrates indicates the presence of significant differences in information content that is independent of their base and dimer contents. These long-range correlations are visually obvious from the two-dimensional graphs of the random walks. The vertebrate genomes are significantly deficient in G with excess A and C and near-normal T content. Even though A+T content is about 57-60% of the genome, the random walks travel only a distance of 1,500 on the A<->T axis compared to 3,000 on the C<->G axis. Both vertebrate and invertebrate mtDNA is extremely deficient in both CG and GC and very rich in AA, AT, TA, and TT. This disparity in distance traveled on the two axes for vertebrate genomes must mean that A+T tends to occur in short runs such as AATT, ATAT, TATA, TTAA, or longer combinations like this. The two-dimensional projection places A and T on opposite poles of the same axis. Thus, the walks of these sequences oscillate in short spurts in either direction and fail to make any significant travel along that axis. In contrast, significant travel is accomplished along the C<->G axis with travel in the C direction. This behavior is uniform across the sequence, showing long-range correlations that are sustained throughout the genome. Invertebrates also show long-range correlations, but their paths have distinct changes in direction, indicating shorter long-range correlations that are nonuniform. The multifractal spectra successfully quantified these differences, which were not elucidated using entropy.

6. DISCUSSION AND CONCLUSION

6.1 What Has This Research Revealed?

Fractal analysis of DNA sequences has revealed several things. Human DNA sequences on average have a lower fractal dimension than three control types--random, base-matched, and dimer-matched sequences. The DNA sequences appear to have long-range correlations that are more than just near-neighbor effects. The nature of these correlations is not totally understood. One possible type of correlation is the presence of heterogeneous segments that differ in base and dimer content. This heterogeneity was most marked for the dimers AA, CC, GG, and TT with over 95% of the sequences analyzed showing nonuniform frequency distributions within a 500 base-pair window. This reflects the fact that base-runs of length 5-6 are common in DNA. Over 50% of the sequences showed heterogeneity for AG, CG, CT, GA, GC, GC, and TA. About a third showed heterogeneity for AC, AT, CA GT, and TG.

Sequences that code for protein have a lower fractal dimension than sequences that do not have a coding function. It is unlikely that the correlations in coding DNA, which produce the lower fractal dimension, can be totally explained by codon preference. DNA sequences were mapped into a four-dimensional space using a sliding-dimer scheme. This essentially ignores the fact that codons are found in triplets in a specific frame. A neural net algorithm for finding coding regions in anonymous DNA uses fractal dimension as one of eight inputs (Uberbacher and Mural 1991). My research concurs with their conclusions that coding DNA usually has a lower fractal dimension than noncoding DNA. The fact that coding sequences seem to have longer-range correlations than noncoding sequence producing lower fractal dimensions is consistent with research on the entropy of *Escherichia coli* sequences. Lauc (Lauc, Ilic, and Heffer-Lauc 1992) found that the entropy of coding sequences was lower than noncoding sequences. Rowe's research on viral DNA, however, yielded an opposite conclusion--that noncoding DNA was more correlated than coding DNA (Rowe and Trainor 1983).

Peng et al. (1992) concluded that intron-containing gene sequences showed long-range correlations whereas pure coding sequences did not. The findings of my research clearly indicate that both coding and noncoding DNA show some long-range correlations, and that coding DNA shows stronger correlations than noncoding. One explanation for this disagreement is that their research used a one-dimensional embedding. This research has shown that any embedding less than three is inadequate and will yield invalid results.

Mitochondrial DNA sequences from vertebrate genomes have a lower fractal dimension and multifractal spectrum than invertebrate genomes. Evolution by random mutation would produce more randomness and less correlation over time. Luo (1988) found that the fractal dimension in DNA sequences increased with increasing evolutionary level. However, his analysis was based on a single gene sequence for each organism, which may not be representative of the entire organism. This research analyzed the entire genome sequence of each organism.

The quantified fractal dimension of a DNA sequence alone does not discriminate coding from noncoding DNA sequences. It cannot be said that any sequence with dimension below a certain value codes for protein. The fractal dimension of a sequence must be compared with the fractal dimension of controls to be useful. However, if the difference between D for a sequence and its controls is examined, coding DNA can be distinguished from noncoding DNA.

6.2 Fractal Algorithms

Fractal analysis methods are computationally intensive. Even the most efficient algorithms would be difficult to apply to very long DNA sequences. The iterative nature of these algorithms make them ideal candidates for parallel processing. Without parallel processing, it is very important to use efficient programming to reduce computation time. New methods are needed to evaluate dimension estimates directly, without the need for control sequences.

A random walk representation was selected because it is a very simple random fractal. More complex representations may be more informative and provide more specific information about the sequences studied.

6.3 Unsolved Theoretical Problems

Much theory exists about random walks, but there are many theoretical issues about fractal dimension that have not been solved in this research and need to be addressed. What is the theoretical fractal dimension of a correlated random walk? Can finite size effects be estimated theoretically?

No theoretical research has been done on the multifractal spectrum of a correlated random walk. A true, uncorrelated walk is a monofractal with a fractal dimension of two for any embedding dimension. The multifractal spectrum of an object quantifies nonhomogeneity. Because the random walk representation used in this research is a correlated walk, it may be a valid multifractal. Research is need to quantify the theoretical values for the multifractal spectrum of a correlated random walk.

In this research, a multifractal spectrum was obtained for a theoretical monofractal, the Sierpinski carpet, by using a scaling range that went beyond the outer edge. Is the fractal dimension estimate obtained by "running off the edge" of the fractal meaningful? What does it really tell us? To be theoretically valid, a D_q curve must be monotonically nonincreasing. Smaller scaling ranges that did not run off the edge failed to produce valid D_q curves. The D_q values for negative q were consistently less than D_q at zero. Although not considered theoretically valid, does a D_q curve that is not monotonically nonincreasing contain useful information? To obtain D_q closer to the theoretical value, a different scaling range for each q might be needed rather than applying the same range for all q.

It would seem that since the mean-square displacement of a walk is a function of its length, scaling by a function that is a square root of the length should yield the same multifractal spectrum for random walks of all lengths. Yet the spectrum values are not the same and actually increased with the length of sequence. Why does the multifractal spectrum for a random walk increase with sequence length when scaling by a function that is a square root of the length?

There does not seem to be agreement on how entropy relates to information content. Does high entropy mean more random but more information content? Does low entropy mean less random? Is there more information in a system with low entropy or high entropy? How does fractal dimension relate to entropy?

6.4 Future Research

Ways to apply fractal analysis to smaller data sets need to be discovered. The smallest data set, in which a reasonable fractal dimension estimate could be obtained using the random walk representation, was about 4,000 base pairs. The shortest random walk to yield a valid multifractal spectrum was 15,000 base pairs. These are serious limitations. The average exon of a gene is about 120 base pairs. The coding portions of a gene sequence are rarely longer than 5,000 base pairs. To apply fractal analysis successfully to distinguish coding regions from noncoding regions, methods of estimating the fractal dimension of small data sets must be found. Perhaps other representations will reduce this length limitation.

Another issue is how to assign dimers to axes in a random walk representation. Will other axis representations indicate something different about these sequences or are the results too general? 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 oligonucleotides 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.

The random walk is a heuristic representation for DNA sequences. Is there a different representation that might be more informative? Lastly, what type of nonrandomness does fractal dimension actually measure? Is it just heterogeneity or does it describe other long-range correlations?

6.5 Conclusion

This research has shown that measurements of fractal dimension of DNA sequences may be quite useful in quantifying long-range correlations. The differences in long-range correlations between sequences may be useful in distinguishing functional DNA sequences from nonfunctional sequences. A significant amount of theoretical and computational work has been completed in this research to establish the basis of fractal analysis of DNA sequences. Ambiguous or contradictory findings are possible when fractal analysis is applied without adherence to basic rules and premises of the paradigm. Fractal analysis promises to uncover information about the internal organization of DNA sequences, which has not been possible by traditional methods.

APPENDIX

COMPUTER SOURCE CODE

```
/* Begin program sandbox.c:  does sandbox counting and saves
data to file to be processed by computedq.c  */

#include <stdio.h>
#include <math.h>
#include <malloc.h>

#define TRUE 1
#define FALSE 0
#define MaxDim 4

typedef struct {
    int coord [MaxDim];
} *Dptr, DPstruct;

static float *distblock;
static Dptr datablock;
static int minR, maxR, NUMRANDS;
static long rwlen;

#define  M  714025
#define  IA 1366
#define  IC 150889

float ran2 (idum)
    long *idum;
/*This is linear congruential random number generator that has an
effective cycle not less than 700,000 (Press et al., p. 212.)*/
{
    static long iy, ir[98];
    static int iff=0;
    int j;
    if (*idum < 0 || iff==0) {
        iff=1;
        if ((*idum=(IC-(*idum)) % M) < 0) *idum = -(*idum);
        for (j=1; j<=97; j++) {
            *idum=(IA*(*idum)+IC) % M;
            ir [j] = (*idum);
        }
        *idum = (IA*(*idum)+IC) % M;
        iy = (*idum);
    }
    j=1 + 97.0*iy/M;
    iy=ir[j];
    *idum=(IA*(*idum)+IC) % M;
    ir[j]=(*idum);
    return (float) iy / M;
}

long string_value (locusname)
    char *locusname;
/* Produce unique and reproducible seed value based on name */
{
    char *charptr;
    int stringlength, i;
    long tempseed = 0;

    charptr = locusname;
    stringlength = strlen (locusname);
    for (i=0; i<stringlength; ++i)
        tempseed = tempseed + *(charptr++);
    return tempseed;
}

long load_data_points (npoints, locus)
    long npoints;
    char *locus;
/* Read in data points from file and load into RAM datablock*/
{
    long np, bp;
    int x, y, z, w, q, d;
    char filename[50];
    Dptr this;
    FILE *pointFile;
    pointFile = fopen (locus, "r");
    if (pointFile == NULL) {
        printf ("%s file not found\n",locus);
        exit (1);
    }
    this = datablock;
    np = 0;
    while (!feof(pointFile)) {
        fscanf (pointFile, "%d%d%d%d%ld", &x, &y, &z, &w, &bp);
        this->coord[0] = x ;
        this->coord[1] = y ;
        this->coord[2] = z ;
        this->coord[3] = w;
        ++np; ++this;
    }
    fclose (pointFile);
    return np;
}

float euclid_distance (data, p1, p2)
    DPstruct data[];
    long p1, p2;
/* Compute euclidean distance between point 1 and point 2 */
{
    int dim, Diff[MaxDim];
    double sumsq=0.0;

    for (dim=0; dim<MaxDim; ++dim) {
        Diff[dim] = data[p1].coord[dim] - data[p2].coord[dim];
        sumsq += Diff[dim] * Diff[dim];
    }
    return sqrt(sumsq);
}

void find_distances (rn, data, dist, np)
    long rn, np;
    DPstruct data[];
    float dist[];

/*Compute the distance between point np and all other points and
store values in distance matrix */
```

```
{
    long dp, i;
    float  d;
     i=0; dp=0;

    while (dp<np) {
        if (dp != rn) {
            dist[i] = euclid_distance (datablock, dp, rn);
            ++i;
        }
        ++dp;
    }
}

void sort_distances (ra, n)
    float  ra[];
    long n;
/*Sort distances in distance matrix by increasing values.  This is
a heap sort (Press et al., p. 247). */
{
    long l, j, ir, i;
    float  rra;

    l = (n >> 1)+1;
    ir=n;

    for (;;) {
        if (l > 1)
            rra=ra[--l];
        else {
            rra=ra[ir];
            ra[ir]=ra[1];
            if (--ir == 1) {
                ra[1]=rra;
                return;
            }
        }
        i=l;
        j=l << 1;
        while (j <= ir) {
            if (j < ir && ra[j] < ra[j+1]) ++j;
                if (rra < ra[j]) {
                    ra[i] = ra[j];
                    j += (i=j);
                }
            else j=ir+1;
        }
        ra[i]=rra;
    }
}

void sandbox_count (npoints, dist, minR, maxR, locus)
    long npoints;
    int minR, maxR;
    float dist[];
    char *locus;
/* Count number of points within radius range minR to maxR using
the sorted distance matrix and output data to sandbox file */
{
    int r;
    long count, i;
    float radius;
    FILE *outfile;
    char outname[50];

    strcpy (&outname[0], locus);
    outfile = fopen (outname, "a");
    i=0; count = 1; r = minR;

    while (r <= maxR) {
        radius = r;
        while ( (dist[i] <= radius) && (i<npoints) ) {
            ++count; ++i;
        }
        fprintf (outfile, "%d\t%ld\n", r, count);
        r+=2;
    }
    fclose (outfile);
}

long process_data (npoints, locus, minR, maxR, NUMRANDS)
    long npoints;
    int NUMRANDS, minR, maxR;
    char *locus;
```

```
/*load data points, select NUMRANDS data points and do sandbox
for each point finding distances, sorting, and then counting*/
{
    long np, randnum;
    int nrands, nradii;
    float maxdistance, rn;
    long seed;
    char newlocus[50];

    np = load_data_points (npoints, locus, &avgspan);
    distblock = (float *) calloc (np, sizeof (float));
    if ( (float *) distblock == NULL) {
        printf ("     Unable to allocate Distance memory\n");
        exit(1);
    }

    seed = - (string_value(locus)); /*seed based on locus name*/

    nradii = (maxR - minR + 2) * 0.50;
    for (nrands=1; nrands<=NUMRANDS; ++nrands) {
        rn = ran2 (&seed);
        randnum = (rn * np) + 0.5;
        find_distances (randnum, datablock, distblock, np);
        sort_distances (distblock-1, np-1);
        sprintf (&newlocus[0], "%s.%d_%d.%d.sb", locus, minR,
            maxR, NUMRANDS);
        sandbox_count (np-1, distblock, minR, maxR,
            &newlocus[0]);
    }
    free ( (float *) distblock);
    return nradii;
}


sandbox (npoints, locus, fName, minR, maxR, NUMRANDS)
    long npoints;
    int minR, maxR, NUMRANDS;
    char *locus, *fName;
{
    char Lname[50];
    int nradii;
    FILE *monitor;
    datablock = (Dptr) calloc (npoints, sizeof (DPstruct));
    if ( (Dptr) datablock == NULL) {
        printf ("      Unable to allocate memory for datablock\n");
        exit (1);
    }
    else {
        nradii = process_data (npoints, locus, minR, maxR,
        NUMRANDS);
        monitor = fopen (fName, "a");
        fprintf (monitor, "%s.%d_%d.%d.sb\t%ld\t%d\t%d\n",
            &locus[0], minR, maxR, NUMRANDS, npoints, nradii,
                NUMRANDS);
        fclose (monitor);
        free ( (Dptr) datablock);
    }
}

main (argc, argv)
    int argc;
    char *argv[];
{
    char locus[50], fName[50];

    if (argc != 6) {
        printf ("Parameter  error\n");
        exit (1);
    }
    strcpy  (&locus[0],argv[1]);
    rwlen  =  atol(argv[2]);
    minR  =  atoi(argv[3]);
    maxR  =  atoi(argv[4]);
    NUMRANDS = atoi(argv[5]);
    strcpy  (&fName[0],"gene.batch");
    sandbox (rwlen, &locus[0], &fName[0], minR, maxR, NUMRANDS);
}

/*  End of program sandbox.c  */
/* Begin program computedq.c:  process data file created by
sandbox.c and compute Dq curve */

#include <stdio.h>
#include <math.h>
#include <malloc.h>

#define MaxQ 14
#define MaxSamples 1500

struct LOGARRAY {
    double boxsize, boxcount;
} *logblock;

int DQ[MaxQ];

struct CIRCLE {
    int radius;
    long sum, center [MaxSamples];
} *circleblock;

static float sqrarg;
#define SQR(a) (sqrarg=(a), sqrarg*sqrarg)


void fit (ndata, a, b, siga, sigb, chi2, q)
    double *a, *b, *siga, *sigb, *chi2, *q;
    int ndata;
/ * Linear regression (Press et al., p.527) */
{
    int i;
    double t, sxoss, sx=0.0, sy=0.0, st2=0.0, ss, sigdat;
    struct LOGARRAY *logdata;

    logdata = logblock;
    *b=0.0;
    for (i=0; i<ndata; ++i) {
        sx += logdata->boxsize;
        sy += logdata->boxcount;
        ++logdata;
    }
    ss=ndata;
    sxoss=sx/ss;
    logdata = logblock;
    for (i=0; i<ndata; ++i) {
        t=(logdata->boxsize)-sxoss;
        st2 += t*t;
        *b += t*(logdata->boxcount);
        ++logdata;
    }
    *b /= st2;
    *a=(sy-sx*(*b))/ss;
    *siga=sqrt((1.0+sx*sx/(ss*st2))/ss);
    *sigb=sqrt(1.0/st2);
    *chi2=0.0;
    logdata = logblock;
    for (i=0; i<ndata; ++i) {
        *chi2 += SQR((logdata->boxcount)-
            (*a)-(*b)*(logdata->boxsize));
        ++logdata;
    }
    *q=1.0;
    sigdat = sqrt((*chi2)/(ndata-2));
    *siga *= sigdat;
    *sigb *= sigdat;
}


void load_data (fname, nradii, ncenters, circ_array, maxradii)
    char *fname;
    int nradii, ncenters, maxradii;
    struct CIRCLE circ_array[];

/* load data from File created by sandbox.c.  Data structure is an
array of circles representing from size minRadius to maxRadius
sized balls.  Each of the randomly selected points has a count to be
filled into the appropriate circle element.  So there are nradii
circles with ncenters sandbox values.  Each circle has a radius
value and a sum value representing the total number of points
covered by all the balls that size.  Program is written so that
different scaling ranges can be applied using same sandbox file.
Although nradii may have been sandboxed, you may choose to only
go to maxradii over a shorter range */
```

```
{
    FILE *datafile;
    int c, r, radii;
    long npoints;

    if ( (datafile = fopen (fname, "r")) == NULL) {
        printf ("datafile not found\n");
        exit (1);
    }
    for (c=0; c<ncenters; ++c) {
        for (r=0; r<nradii; ++r) {
            fscanf (datafile, "%d%d", &radii, &npoints);
            if (r<maxradii) {
                circ_array[r].center[c] = npoints;
                if (c==0) {
                        circ_array[r].radius = radii;
                 if (r==0) circ_array[r].sum = 0;
            }
            circ_array[r].sum += npoints;
        }
    }
    }
    fclose (datafile);
}

void do_dqs (circ_array, nradii, ncenters, totalpoints, locus)
    struct CIRCLE circ_array[];
    int nradii, ncenters;
    long totalpoints;
    char *locus;
{
    FILE *outfile;
    double a, dq, siga, sigb, chi2, qs;
    struct LOGARRAY *LB;
    char outname[60];
    double qval, pTotal, pCover, dsum, radsize;
    int r, c, logpoints, q, i, lp;
    double dbltotal, dblcenter, dblnc;

    strcpy (&outname[0], &locus[0]);
    strcat (&outname[0], ".DQ");
    outfile = fopen (outname, "w");
```

```
     logblock = (struct LOGARRAY *) calloc (nradii,
         sizeof (struct LOGARRAY));
     if (logblock == NULL) {
         printf ("can't allocate logblock\n");
         exit (1);
     }
     dbltotal = totalpoints; dblnc = ncenters;
     for (q=0; q<14; ++q) {
         LB = logblock;
         qval = (double) (DQ[q]-1);
         logpoints = 0;
         for (r=0; r<nradii; ++r) {
             dsum = 0.0;
             radsize = (double) circ_array[r].radius;
             for (c=0; c<ncenters; ++c) {
                 dblcenter = circ_array[r].center[c];
                 pTotal = dblcenter / dbltotal;
                 pTotal = pow (pTotal, qval);
                 pTotal /= dblnc;
                 dsum += pTotal;
             }
             LB->boxsize = log10(radsize);
             LB->boxcount = log10(dsum);
             ++LB; ++logpoints;
         }

        fit (logpoints, &a, &dq, &siga, &sigb, &chi2, &qs);
        dq /= qval;
        fprintf (outfile, "%d\t%.3f\n", DQ[q], dq);

/* This section dumps log/log arrays if needed for visualization
of log/log plot
        LB = logblock;
        for (lp=0; lp<logpoints; ++lp) {
            printf ("%.6f\t%.6f\n", LB->boxsize, LB->boxcount);
            ++LB;
        }
* /

    }
    free ( (struct LOGARRAY *) logblock);
}
```

```
main (argc, argv)
    int argc;
    char *argv[];
{
    FILE *batchfile, *datafile;
    char bname[50];
    long npoints;
    int nradii, ncenters, maxradii;
    char locus [50];

    if (argc == 1) {
        printf ("No batch file specified\n");
        exit (1);
    }
    strcpy (&bname[0], *++argv);
    if ((batchfile = fopen (bname, "r")) == NULL) {
        printf ("%s NOT FOUND\n", &bname[0]);
        exit (1);
    }

    DQ[0] = -15;
    DQ[1] = -10;
    DQ[2] = -5;
    DQ[3] = -4;
    DQ[4] = -3;
    DQ[5] = -2;
    DQ[6] = -1;
    DQ[7] = 0;
    DQ[8] = 2;
    DQ[9] = 3;
    DQ[10] = 4;
    DQ[11] = 5;
    DQ[12] = 10;
    DQ[13] = 15;

    while (!feof(batchfile)) {
        fscanf (batchfile, "%s%ld%d%d%d",
            &locus[0], &npoints, &nradii, &ncenters, &maxradii);
        circleblock = (struct CIRCLE *) calloc (maxradii,
            sizeof (struct CIRCLE));
        if ( (struct CIRCLE *) circleblock == NULL)
            printf ("Unable to allocate memory for circles\n");
        else {
            load_data (&locus[0], nradii, ncenters, circleblock,
                maxradii);
            do_dqs (circleblock, maxradii, ncenters, npoints,
                &locus[0]);
            free ( (struct CIRCLE *) circleblock);
        }
    }
    fclose  (batchfile);
}

/* End of program computedq.c */
```

REFERENCES

Almagor, Hagai. 1983. A Markov analysis of DNA sequences. *Journal of Theoretical Biology* 104: 633-645.

Barber, Michael N., and B. W. Ninham. 1970. *Random and restricted walks: Theory and applications.* New York: Gordon and Breach, Science Publishers.

Barnsley, Michael. 1988. *Fractals everywhere.* San Diego: Academic Press, Inc.

Benson, Donald C. 1990. Fourier methods for biosequence analysis. *Nucleic Acids Research* 18: 6305-6310.

Bishop, D. Timothy, John A. Williamson, and Mark H. Skolnick. 1983. A model for restriction fragment length distributions. *American Journal of Human Genetics* 35: 795-815.

Blaisdell, B. Edwin. 1983. A prevalent persistent global nonrandomness that distinguishes coding and noncoding eucaryotic nuclear DNA sequences. *Journal of Molecular Evolution* 19: 122-133.

Blaisdell, B. Edwin. 1985. Markov chain analysis finds a significant influence of neighboring bases on the occurrence of a base in eucaryotic nuclear DNA sequences both protein-coding and noncoding. *Journal of Molecular Evolution* 21: 278-288.

Block, A., W. von Bloh, and H. J. Schellnhuber. 1990. Efficient box-counting determination of generalized fractal dimensions. *Physical Review A* 42: 1869-1874.

Churchill, Gary A. 1989. Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology* 51: 79-94.

Dewey, T. G., and M. M. Datta. 1989. Determination of the fractal dimension of membrane protein aggregates using fluorescence energy transfer. *Biophysics Journal* 56: 415-420.

Dix, Daniel B., and Robert C. Thompson. 1989. Codon choice and gene expression: Synonymous codons differ in translational accuracy. *Proceedings of the National Academy of Science, USA* 86: 6888-6892.

Dvorak, Ivan, and Jan Klaschka. 1990. Modification of the Grassberger-Procaccia algorithm for estimating the correlation exponent of chaotic systems with high embedding dimension. *Physics Letters A* 145: 225-231.

Elton, R. A. 1975. Doublet frequencies in sequenced nucleic acids. *Journal of Molecular Evolution* 4: 323-346.

Falconer, Kenneth. 1990. *Fractal geometry: Mathematical foundations and applications.* Chichester, England: John Wiley and Sons.

Fickett, James W. 1982. Recognition of protein coding regions in DNA sequences. *Nucleic Acids Research* 10: 5303-5318.

Fuchs, Camil. 1980. On the distribution of the nucleotides in seven completely sequenced DNAs. *Gene* 10: 371-373.

Garden, Peter W. 1980. Markov analysis of viral DNA/RNA sequences. *Journal of Theoretical Biology* 82: 679-684.

Gates, Michael A. 1986. A simple way to look at DNA. *Journal of Theoretical Biology* 119: 319-328.

Gatlin, L. L. 1972. *Information theory and the living system.* New York: Columbia University Press.

Glazier, James A., and Albert Libchaber. 1988. Quasi-periodicity and dynamical systems: An experimentalist's view. *IEEE Transactions on Circuits and Systems* 35: 790-807.

Gould, Harvey, and Jan Tobochnik. 1990. More on fractals and chaos: Multifractals. *Computers in Physics* Mar/Apr: 202-207.

Grassberger, Peter, and Itamar Procaccia. 1983. Characterization of strange attractors. *Physical Review Letters* 50: 346-349.

Grassberger, Peter. 1990. An optimized box-assisted algorithm for fractal dimensions. *Physics Letters A* 148: 63-68.

Greenside, H. S., A. Wolf, J. Swift, and T. Pignataro. 1982. Impracticality of a box-counting algorithm for calculating the dimensionality of strange attractors. *Physical Review A* 25: 3453-3456.

Hakansson, Jan, and Gunnar Russberg. 1990. Finite-size effects on the characterization of fractals sets: f(alpha) construction via box counting on a finite two-scaled Cantor set. *Physical Review A* 41: 1855-1861.

Halsey, Thomas C., Mogens H. Jensen, Leo P. Kadanoff, Itamar Procaccia, and Boris I. Shraiman. 1986. Fractal measures and their singularities: The characterization of strange sets. *Physical Review A* 33: 1141-1151.

Helman, J. S., Antonio Coniglio, and Constantino Tsallis. 1984. Fractons and the fractal structure of proteins. *Physical Review Letters* 53: 1195-1197.

Hentschel, H. G. E., and Itamar Procaccia. 1983. The infinite number of generalized dimensions of fractals and strange attractors. *Physica* 8D: 435-444.

Hong, Juan. 1990. Prediction of oligonucleotide frequencies based upon dinucleotide frequencies obtained from the nearest neighbor analysis. *Nucleic Acids Research* 18: 1625-1628.

Isogai, Yoshinori, and Toshiyuki Itoh. 1984. Fractal analysis of tertiary structure of protein molecule. *Journal of the Physical Society of Japan* 53: 2162-2171.

Jeffrey, H. Joel. 1990. Chaos game representation of gene structure. *Nucleic Acids Research* 18: 2163-2170.

Kemeny, John G., and J. Laurie Snell. 1976. *Finite Markov chains.* New York: Springer-Verlag.

Kimura, T., T. Takeya, and M. Takanami. 1989. Reconstitution of nucleosomes in vitro with a plasmid carrying the long terminal repeat of Moloney murine leukemia virus. *Biochimica et Biophysica ACTA* 1007: 318-324.

Kleffe, Jurgen, and Uwe Langbecker. 1990. Exact computation of pattern probabilities in random sequences generated by Markov chains. *Computer Applications in the Biosciences* 6: 347-353.

Konopka, A. 1984. Is the information content of DNA evolutionarily significant? *Journal of Theoretical Biology* 107: 697-704.

Kuhn, Thomas S. 1970. *The structure of scientific revolutions*, 2d ed., enlarged. Chicago: University of Chicago Press.

Lauc, Gordan, Igor Ilic, and Marija Heffer-Lauc. 1992. Entropies of coding and noncoding sequences of DNA and proteins. *Biophysical Chemistry* 42: 7-11.

Lewis, M., and D. C. Rees. 1985. Fractal surfaces of proteins. *Science* 230: 1163-1165.

Lim, H. A. 1991. A fractal representation approach to classify the functional regions of DNA sequences. In *The DOE Human Genome Program contractor-grantee workshop, February 17-20, 1991*. Santa Fe, New Mexico: 86, photocopied.

Luo, Liaofu, and Lu Tsai. 1988. Fractal dimension of nucleic acid sequences and the relation to evolutionary level. *Chinese Physical Letters* 5: 421-423.

Mandelbrot, Benoit B. 1983. *The fractal geometry of nature.* New York: W. H. Freeman and Co.

Mandelbrot, Benoit B. 1989. Fractal geometry: What is it, and what does it do? *Proceedings of the Royal Statistical Society, London A* 423: 3-16.

Nussinov, Ruth. 1980. Some rules in the ordering of nucleotides in the DNA. *Nucleic Acids Research* 8: 4545-4562.

Nussinov, Ruth. 1981. The universal dinucleotide asymmetry rules in DNA and the amino acid codon choice. *Journal of Molecular Evolution* 17: 237-244.

Nussinov, Ruth. 1984a. Doublet frequencies in evolutionary distinct groups. *Nucleic Acids Research* 12: 1749-1763.

Nussinov, Ruth. 1984b. Strong doublet preferences in nucleotide sequences and DNA geometry. *Journal of Molecular Evolution* 20: 111-119.

Ohno, Susumu. 1988. Codon preference is but an illusion created by the construction principle of coding sequences. *Proceedings of the National Academy of Science, USA* 85: 4378-4382.

Peng, C. K., S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley. 1992. Long-range correlations in nucleotide sequences. *Nature* 356: 168-170.

Pennings, S., S. Muyldermans, G. Meersseman, and L. Wyns. 1989. Formation, stability and core histone positioning of nucleosomes reassembled on bent and other nucleosome-derived DNA. *Journal of Molecular Biology* 207: 183-192.

Phillips, Gregory J., Jonathan Arnold, and Robert Ivarie. 1987. Mono- through hexanucleotide composition of the *Escherichia coli* gnome: A Markov chain analysis. *Nucleic Acids Research* 15: 2611-2626.

Pool, Robert. 1990. Fractal fracas. *Science* 249: 363-364.

Press, William H., Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. 1988. *Numerical recipes in C: The art of scientific computing.* Cambridge, Mass.: Cambridge University Press.

Purugganan, M. D. 1989. The fractal nature of RNA secondary structure. *Naturwissenschaften* 76: 471-473.

Ragosta, Maria, Carmelina Cosmi, Vincenzo Cuomo, and Maria Macchiato. 1992. An application of maximum entropy techniques to determine homogeneous sets of nucleotide sequences. *Journal of Theoretical Biology* 155: 129-136.

Ramsey, James B. and Hsiao-Jane Yuan. 1989. Bias and error bars in dimension calculations and their evaluation in some simple models. *Physics Letters A* 134: 287-297.

Rowe, Glenn W., and L. E. H. Trainor. 1983. On the informational content of viral DNA. *Journal of Theoretical Biology* 101: 151-170.

Rudnick, Joseph, and George Gaspari. 1987. The shapes of random walks. *Science* 237: 384-389.

Satchwell, Sandra C., Horace R. Drew, and Andrew A. Travers. 1986. Sequence periodicities in chicken nucleosome core DNA. *Journal of Molecular Biology* 191: 659-675.

Shannon, C. E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27: 379-423.

Shrader, T. E., and D. M. Crothers. 1989. Artificial nucleosome positioning sequences. *Proceedings of the National Academy of Science USA* 86: 7418-7422.

Silverman, B. D., and R. Linsker. 1986. A measure of DNA periodicity. *Journal of Theoretical Biology* 118: 295-300.

Smith, Leonard A. 1988. Intrinsic limits on dimension calculations. *Physics Letters A* 133: 283-288.

Staden, R. 1984. Measurements of the effects that coding for a protein has on a DNA sequence and their use for finding genes. *Nucleic Acids Research* 12: 551-567.

Stapleton, H. J., J. P. Allen, C. P. Flynn, D. G. Stinson, and S. R. Kurtz. 1980. Fractal form of proteins. *Physical Review Letters* 45: 1456-1459.

Subba Rao, G., Z. Hamid, and J. Subba Rao. 1979. The information content of DNA and evolution. *Journal of Theoretical Biology* 81: 803.

Subba Rao, J., C. P. Geevan, and Giva Subba Rao. 1982. The significance of the information content of DNA in mutations and evolution. *Journal of Theoretical Biology* 96: 571.

Takahashi, Manabu. 1989. A fractal model of chromosomes and chromosomal DNA replication. *Journal of Theoretical Biology* 141: 117-136.

Tavare, Simon, and Brenda Song. 1989. Codon preference and primary sequence structure in protein-coding regions. *Bulletin of Mathematical Biology* 51: 95-115.

Taylor, Charles C., and S. James Taylor. 1991. Estimating the dimension of a fractal. *Journal of the Royal Statistical Society B* 53: 353-364.

Tel, Tamas, Agnes Fulop, and Tamas Vicsek. 1989. Determination of fractal dimensions for geometrical multifractals. *Physica A* 59: 155-166.

Tsonis, Panagiotis A., and Anastasios A. Tsonis. 1989. Chaos: Principles and implications in biology. *Computer Applications in the Biosciences* 5: 27-32.

Uberbacher, Edward C., Joel M. Harp, and Gerard J. Bunick. 1988. DNA sequence patterns in precisely positioned nucleosomes. *Journal of Biomolecular Structure and Dynamics* 6: 105-120.

Veljkovic, V., I. Cosic, B. Dimitrijevic, and D. Lalovic. 1985. Is it possible to analyze DNA and protein sequences by the methods of digital signal processing. *IEEE Transactions on Biomedical Engineering* BME-32: 337-341.

Viscek, T., F. Family, and P. Meakin. 1990. Multifractal geometry of diffusion-limited aggregates. *Europhysics Letters* 12: 217-222.

Volinia, S., R. Gambari, F. Bernardi, and I. Barrai. 1989. The frequency of oligonucleotides in mammalian genic regions. *Computer Applications in the Biosciences* 5: 33-40.

Wang, Cun Xin, and Yun Yu Shi. 1990. Fractal study of tertiary structure of proteins. *Physical Review A* 41: 7043-7048.

Weibel, Ewald R. 1991. Fractal geometry: A design principle for living organisms. *American Journal of Physiology* 261: L361-L369.

Weir, B. S. 1985. Statistical analysis of molecular genetic data. *IMA Journal of Mathematics Applied in Medicine and Biology* 2: 1-39.