1085-9195/04/41:41–62/$25.00

ORIGINAL ARTICLE

Self-Similar Mitochondrial DNA

Nestor N. Oiwa,[1] and James A. Glazier*[2]

[1]Institute of Physics, University of São Paulo, São Paulo, Brazil and Department of Physics,
University of Notre Dame, Notre Dame, IN; and [2]Department of Physics
and Biocomplexity Institute, Indiana University, Bloomington, IN

Abstract

We show that repeated sequences, like palindromes (local repetitions) and homologies between two different nucleotide sequences (motifs along the genome), compose a self-similar (fractal) pattern in mitochondrial DNA. This self-similarity comes from the looplike structures distributed along the genome. The looplike structures generate scaling laws in a pseudorandom DNA walk constructed from the sequence, called a Lévy flight. We measure the scaling laws from the generalized fractal dimension and singularity spectrum for mitochondrial DNA walks for 35 different species. In particular, we report characteristic loop distributions for mammal mitochondrial genomes.

Index Entries: Fractal; mitochondrial genome; scaling laws; Lévy flight; DNA walk; large-scale DNA analysis; palindromes; DNA loops; 16S rRNA.

INTRODUCTION

Repeated sequences, like palindromes and homologies, appear commonly in the literature on eukaryotic genomes *(1–3)*. These repeats can manifest in many ways. The most common case is when one genomic segment repeats in another part of the same nucleotide sequence. This repetition causes typical difficulty in recomposing large complete genomic sequences. Because complete nucleotide genomic sequences usually require assembling shotgun sequences by looking for overlapping *K*-letter words between sequences, we expect matching difficulties as a result of cloning bias and repeats in many genomes, including humans *(2,3)*. Repeat sequences are not restricted to pairs of genes placed in completely different parts of one chromosome, but include local repetitions too, both direct repeats and palindromes. Palindromes have many metabolic functions, including the control of the linkage number in circular DNAs *(1)* and the translational regulation of expression of the bacteriophage T4 lysozyme gene *(4)*. We also find motifs repeated throughout the genome, like the highly-repetitive *Alu* family, common in the human genome *(1–3)*. In nature, simply repeated structures often correspond to repetitions at different length scales (e.g., the

*Author to whom all correspondence and reprint requests should be addressed. E-mail: oiwa@fge.if.usp.br vascular system of a leaf is similar to the whole vascular system of a tree). In the case of nucleotide sequences, we consider the hypothesis that the organization of small subsequences could resemble the organization of the complete genome. The authors in the debate over the best large-scale DNA sequencing strategy (whole-genome or hierarchical shotgun methods) basically consider random or periodic nucleotide sequences using broad criteria (with or without repeats, *etc.*) *(5–7,9,10)*. In whole-genome shotgun sequencing, redundancy reduces misassembly *(2,5–7)*. In the hierarchical shotgun method, a large clone library (BAC library or map) with overlapped random sequences is created. The sequencing is performed locally over each clone contig. The whole nucleotide sequence is recovered, anchoring each BAC sequence in a previously defined map. In this way, the hierarchical shotgun method avoids long-range misassembly and reduces short-range misassembly *(3,8)*. Despite enormous efforts to complete the sequencing of the human genome, large gaps and many unplaced small islands of sequence remain *(2,3,9,10)*. Periodic or random simulated nucleotide sequences have been used to study the most efficient approach to complete the human genome sequence *(6,7,9,10)*. However, recent work suggests that the complete genome has neither random nor periodic organization, but something in between. The complete genome distribution of the protein coding and metabolic control sequences in higher eukaryotes exhibits scaling laws (fractality) *(11,12)*. This scaling might have evolved together with organism complexity *(11)*, where we understand complexity in terms of phenotype characteristics and environmental adaptation according to the definition provided by Haken and Nicolis and Prigogine *(13,14)*. Therefore, a *thermophilum* bacterium is simpler than a mammal because a superior eukaryote has a more elaborate life cycle than an archaebacterium. Oiwa and Goldman suggested that the coding-segment distribution of higher eukaryotes has long-range multifractal correlations, independent of the genetic message (coding content) *(11,12)*. Such order is absent in lower eukaryotes and bacteria. Many authors have found strong indications of multifractality in both genomic and mitochondrial nucleotide sequences *(15–23)*. However, the identification of the genomic structures responsible for the self-similarity became possible only recently with the refinement of multifractal analysis techniques (24). Our previous work linked DNA loops to the self-similarity of the mitochondrial genome (mtDNA) *(25)*. Here, we extend these results. We report scaling laws in mtDNA resulting from the self-similar distribution of loops, hairpins, and other double-stranded DNA (dsDNA) structures e.g., DNA inverted repeated sequences (palindromes). These scaling laws imply the repetition of similar sequences at the same scale and include rescaled self-similarities.

Our analysis concentrates on the circular mitochondrial deoxyribonucleic acid, instead of huge nuclear genomes. In nuclear genome each sequence is larger than 10^6 basepairs (bp), and most genes still await experimental confirmation (putative genes). On the other hand, mtDNA is small for mammals (around 16,000 nucleotides) and has relatively few genes, all of which are identified. Their biochemical pathways are also known, and the literature contains more than 300 complete mtDNA genomes *(26)*. The precise position of the nucleotide bases (adenine [A], thymine [T], guanine [G], and cytosine [C]), the molecular structure of their protein-coding sequences, the protein interactions with the DNA sequences that control gene expression, and so forth are comparable from organism to organism *(20–22,27)*. mtDNA is usually inherited maternally without recombination; thus, changes in mitochondrial DNA reflect mutations that occur in the mitochondria of maternal germ cells and accumulate continuously, so each mitochondrion preserves a history of prior mutations. These characteristics make mtDNA useful for phylogenetic analysis. In this article, we consider plant and fungal mtDNA too. Plant mtDNAs are large for higher plants, larger than 10^5 bp. Whereas the mtDNA of multicellular animals is almost completely coding, the mtDNA of plants presents many noncoding parts, including introns. mtDNA organization differs in plants and multicellular animals, and this disparity appears in our nonlinear analysis.

Despite the difficulties in linking statistical measures, like correlations, to functional aspects of nucleotide sequences *(28)*, many methods have been developed to search for patterns and organization in mitochondrial and nonmitochondrial genomes. In our previous work *(11)*, we studied the distribution of coding segments. Here, we are interested in nucleotide sequences.

The most common approach to studying nucleotide sequences is to search for simple similarities among subsequences *(29,30)*; for example, we might take from two species, proteins that perform similar functions and look for similarities in their DNA sequences. We might look for any two DNA sequences that are similar between organisms and try to infer whether the proteins have a similar function. We could look for subunits of proteins (modules) that might be conserved but might be organized differently in different genes and try to determine if they have similar functions (e.g., zinc-finger binding domains, membrane-crossing domains, β-sheet formers, or tyrosine–kinase-binding sites. We can construct dictionaries with such motifs or words, like the database LIGAND, the Database of Chemical Compounds and Reactions in Biological Pathways, from the Bioinformatics Center, Institute for Chemical Research, Kyoto University *(31)*, or we can build a motif dictionary and try to guess the meaning of the words *(32)*. For example, applying MobyDick, a dictionary-building algorithm, to the nucleotide sequence of the Tup1-Ssn6 repressor system localizes both known and putative binding sites for its regulatory elements *(33)*. MobyDick is an algorithm that builds k-word dictionaries according to patterns or motifs in the sequence of nucleotides. When the average number of occurrences of some particular motif exceeds a statistically significant threshold, we add these k-words to a dictionary. MobyDick weights each possible concatenation of words in this dictionary, looking for the prominent ones. If these combined words have a frequency above some statistical criterion, MobyDick includes them in the dictionary. We expect that the longer words of the dictionary reveal patterns of nucleotide sequences within genes which have particular functions. The main advantage of this method is that we do not need external reference data to build the dictionary, because the statistical significance of the longer words depends only on shorter words. Another proposal to describe DNA sequence content is the *K*-string matrix *(34,35)*. Each element of the 2^K vs 2^K matrix represents the number of repetitions of one particular sequence with *K* nucleotides. When *K* is large, the elements of the *K*-string matrix compose a fractal, indicating a possible self-similar pattern for genomes. We also mention Mantegna et al. *(36)*, where *n*-tuple Zipf analysis of noncoding subsequences exhibits a power law, coding sequences have logarithmic behavior, and noncoding segments have lower *n*-gram entropy than coding regions. However, many authors question the reliability of using Zipf's law to distinguish a hidden language from noise *(37)*.

Although these portraits based on small nucleotide sequences (up to 1000 bp) are popular, histogram or sequence similarities do not completely describe the genome from the thermodynamic point of view. They do not give us information on the structure of the whole genome (canonical ensemble), but only local information (microcanonical ensemble). Contingency tables are efficient for searching for specific motifs and patterns. However, they do not give statistics on the global genomic organization, because, by definition, these tables measure local nucleotide sequences. Thus, a sequence homogeneous at large scales (a random sequence, for example) will look inhomogeneous at small sequence lengths.

The most common large-scale genomic analysis is the cytosine and guanine variation content ratio $f(i)$ *(38–42)*. This analysis seeks contiguous regions with little variation in their C + G content, called isochors. The C + G-rich and C + G-poor domains compose the mosaic structure of the genome. Many authors try to correlate these isochors with genes, CpG regions, chromosomal bands, and so forth *(43)*. The C + G content ratio variation is the differential form of a one-dimensional walk $u(i)$, where cytosine and guanine result in an up step, $u(i + 1) = u(i) + 1$, and adenine and thymine in a down step, $u(i + 1) = u(i) - 1$. When we notice that $f(i) = (1/2) – [u(i + \Delta/2) – u(i – \Delta/2)]/(2\Delta)$, we have the C + G content ratio variation at the i-th nucleotide with a window of width Δ. Another popular prescription for one-dimensional walks uses the relative frequency of purines and pyrimidines *(16–19)*. Now, we have an up step for pyrimidines (C + T), $u(i + 1) = u(i) + 1$, and a down step for purines (G + A), $u(i + 1) = u(i) - 1$. The purine–pyrimidine walks define an exponent β, that is $[C(k)]^2 = <[y(k)]^2> – <y(k)>^2 \sim k^{2\beta}$, where $C(k)$ is the correlation function, $y(k) = u(i) – u(i + k)$, and **<···>** indicates mean value. For simple Brownian motion, β is just $\frac{1}{2}$. However, β differs from $\frac{1}{2}$ for DNA sequences, indicating a power-law decay (long-range correlation) for $C(k)$, Thus, DNA walks have no preferred scale and are fractal.

Although such DNA walks are useful for identifying interesting regions *(43)*, the hidden DNA structure is not one-dimensional *(25)*. Previous works did not detect the self-similar structures reported in this article because the hidden fractal structure is two-dimensional, not one-dimensional *(25)*. The purine–pyrimidine walk neglects the keto-amine variation. The C + G content ratio neglects the pyrimidine (C + T) variation. However, when we consider two–dimensional or higher-dimensional DNA walks, we can easily identify structures, as we will discuss.

MATERIALS

GenBank provides mtDNA nucleotide sequences for many species. GenBank, an Internet database under the responsibility of the National Center for Biotechnology Information, United States of America, stores biological information submitted by individual laboratories worldwide and shares databases with other institutions, including the European Molecular Biology Laboratory and the DNA Database of Japan *(27)*. Table 1 lists the 35 complete mtDNA genomes selected for this article.

Many approaches can be used to study genome content. The literature for large-scale DNA analysis (sequences larger than 1000 base pairs) sometimes treats complete genomes (entire chromosomes or genomes) as continuous nucleotide chains without identification of genes, and it sometimes examines just the protein-coding sequences, organized as exons and introns *(15–23,39,41,42)*. Nucleotide sequence analysis with protein-coding *locus* discrimination looks for patterns distinguishing noncoding nucleotides from protein-coding basepairs. However, the genome has many other structures affecting biological function, such as promoters, protein binding sites, and so forth. Because we are interested in *patterns of only coding sequences*, we must select the *loci* to study. Although the concept of genes as coding nucleotide sequences is precise from a Mendelian (or cytological) point of view, this concept is less useful when we examine a raw nucleotide sequence. We need to choose the appropriate coding and related control sequences. We take as protein-coding and control regions of DNA sequences the following *loci* (in bold) of GenBank flat files:

a. Amino-acid-coding sequences, **CDS**, including stop codons, **mRNA**s, **exon**s, as well as mature peptides **mat_peptide,** and signal-coding regions, **sig_peptide**. This work considers introns as noncoding regions and does not distinguish them from the noncoding sequences between genes.
b. Many types of RNA: transfer RNAs, **tRNAs**; ribosomal RNAs, **rRNAs**; diverse transcripts not defined by the cited fields for RNAs, **misc_RNA** (from miscellaneous RNAs), which require individual checking because of their ambiguous definition. We also consider RNA introns as noncoding regions, as for **CDS** fields.
c. Sites with specific biochemical functions such as stem loops in DNA or RNA,

Table 1
Organism and GenBank Accession Numbers of the mtDNAs Used

Name and GenBank accession number	L_{CD}	L	Box counting	Moving box	Sandbox
Balaenoptera physalus, finback whale, NC001321	16,372	16,398	1.14±0.01	1.15±0.01	1.064±0.005
Balaenoptera musculus, blue whale, NC001601	16,377	16,402	1.11±0.01	1.13±0.01	1.076±0.005
Mus musculus, house mouse, NC001569	16,263	16,295	1.13±0.01	1.14±0.01	1.103±0.006
Rattus norvegicus, Norway rat, NC001665	16,273	16,300	1.10±0.02	1.11±0.02	1.093±0.006
Homo sapiens, human, NC001807	15,883	16,569	1.116±0.008	1.120±0.009	1.070±0.004
Bos Taurus, cow, NC001567	16,309	16,338	1.14±0.01	1.15±0.01	1.124±0.006
Phoca vitulina, harbor seal, NC001325	15,409	16,826	1.14±0.02	1.14±0.02	1.112±0.006
Ornithorhynchus anatinus, duckbill platypus, NC000891	17,008	17,019	1.20±0.02	1.19±0.02	1.169±0.006
Alligator mississippiensis, American alligator, NC001922	16,515	16,646	1.12±0.01	1.12±0.01	1.080±0.005
Gallus gallus, chicken, NC001323	16,727	16,775	1.107±0.008	1.093±0.009	1.080±0.004
Xenopus laevis, African clawed frog, NC001573	17,500	17,553	1.21±0.01	1.20±0.01	1.193±0.008
Cyprinus carpio, common carp, NC001606	15,581	16,575	1.13±0.01	1.18±0.01	1.214±0.008
Crossostoma lacustre, tasseled-mouth loach, NC001727	16,488	16,558	1.188±0.009	1.194±0.009	1.204±0.008
Mustelus manazo, gummy shark, NC000890	16,686	16,707	1.17±0.01	1.16±0.02	1.110±0.006
Petromyzon marinus, sea lamprey, NC001626	16,110	16,201	1.12±0.02	1.14±0.02	1.149±0.006
Branchiostoma lanceolatum, amphioxus, NC001912	14,992	15,076	1.18±0.01	1.21±0.01	1.133±0.007
Balanoglossus carnosus, acorn worm, NC001887	15,335	15,708	1.09±0.02	1.12±0.01	1.058±0.006
Paracentrotus lividus, common urchin, NC001572	15,446	15,696	1.28±0.01	1.31±0.01	1.33±0.01
Strongylocentrotus purpuratus, purple sea urchin, NC001453	15,566	15,650	1.29±0.01	1.30±0.02	1.360±0.009
Drosophila yakuba, fly, NC001322	15,838	16,019	1.19±0.01	1.20±0.02	1.149±0.006
Drosophila melanogaster, fruit fly, NC001709	19,334	19,517	1.12±0.02	1.10±0.02	1.149±0.007
Ceratitis capitata, mediterranean fruit fly, NC000857	15,774	15,980	1.18±0.01	1.22±0.02	1.20±0.02
Apis mellifera, honey bee, NC001566	15,530	16,343	1.21±0.01	1.23±0.02	1.166±0.007
Lumbricus terrestris, common earthworm, NC001673	14,537	14,998	1.22±0.02	1.21±0.02	1.224±0.009
Caenorhabditis elegans, worm, NC001328	13,648	13,794	1.10±0.02	1.09±0.02	1.024±0.004
Ascaris suum, pig roundworm, NC001327	14,079	14,284	1.047±0.006	1.07±0.01	1.050±0.005
Onchocerca volvulus, river blindness roundworm, NC001861	13,521	13,747	1.033±0.005	1.038±0.009	1.057±0.007
Saccharomyces cerevisiae, baker's yeast, NC001224	29,648	85,779	1.23±0.01	1.246±0.008	1.284±0.008
Hansenula wingei, fungus, HASMT	22,371	27,694	1.269±0.007	1.284±0.009	1.197±0.006
Schizosaccharomyces pombe, fungus, MISPCG	16,015	19,431	1.37±0.02	1.41±0.01	1.33±0.01
Podospora anserine, fungus, MTPACG	70,613	100,314	1.19±0.02	1.218±0.009	1.209±0.004
Arabidopsis thaliana, thale cress, MIATGENA and MIATGENB	90,461	366,924	1.537±0.006	1.547±0.006	1.559±0.004
Marchantia polymorpha, liverwort, MPOMTCG	71,094	186,609	1.224±0.009	1.26±0.01	1.261±0.006
Chlamydomonas eugametos, green algae, AF008237	16,409	22,897	1.208±0.009	1.23±0.02	1.204±0.006
Pedinomonas minor, green algae, AF116775	23,677	25,137	1.00±0.01	0.99±0.02	1.103±0.008

Note: Total length L and the sum of protein-coding and control segment L_{CD} of the mtDNA sequences in basepairs (bp). Fractal dimension D_{zero} for the two-dimensional walk using box counting, moving box, and sandbox methods.

stem_loop, initial replication points of the DNA double helix, **rep_origin,** and **D-loop**s.

d. Repeated sequences, **repeat_region**. **repeat_unit** indicates one repeated unit of a **repeat_region**. We must be careful with repeated sequences as *Alu*, because we are considering only repeated sequences with a potential function related to genes in the Mendelian sense.
e. Regions with biological meaning not covered by another *locus* definition, **misc_feature**, like A + T-rich regions. We checked **misc_feature** one-by-one to determine if they belonged to control or protein-coding regions, because their descriptions are ambiguous.

We suppress noncoding segments ("junkDNA"), because we seek only patterns associated with genetic transcription *(11)*. Noncoding DNA could regulate coding nucleotides by determining the position of coding sequences and promoters (relative to the histone wrap and supercoiling), but we restrict our analysis to nucleotides with some direct relation to information transmission (e.g., amino acid codons, etc.). We are not studying the spaces between words, but the letters of the words. Oiwa and Goldman studied and reviewed the meaning of spaces in ref. *12*.

We could extract these regions of interest manually, because the number of fields is small, typically from 20 to 40 for mtDNA, but we use programs developed previously to simplify data manipulation *(11)*. Because gene overlapping is common, we consider as protein-coding and control sequences regions that belong to at least one field previously described *(11)*. We could manipulate these overlaps manually too, because only a few segments have such problems. However, we use our previously developed code for convenience *(11)*.

METHODS

Our starting point defines the DNA walk, known generally as a Lévy flight. Although the Lévy flight treats the DNA sequence diffusively, we reject the term "random" in the literature, because we can identify patterns, as we discuss below. The embedding space for the DNA walk is a space of possible sequences (not the phase space of dynamical systems theory, because we are not describing movements or oscillations in time). The walks in this work have the following symmetries, based on the biochemical characteristics of the genome *(20)*:

i. Complementarity: Because the information is duplicated on the two DNA strands, the content must be equivalent for both strands.
ii. Reflection: We read some genes from their 5′ end to their 3′ end and others from 3′ to 5′, so the reading direction should not influence our estimates.
iii. Substitution: We identify two nucleotide groups: T and A are weak-bonding bases (with two hydrogen bonds), whereas G and C are strong-bonding bases (with three hydrogen bonds). The substitution of T for A or C for G must preserve the walk shape.
iv. Compatibility: The walk must be the same for different embedding dimensions *d*. We assume an object of finite dimension can represent the data. Further, we hope this dimension is small, because current nonlinear methods are reliable only for objects with dimension below eight. We will check this assumption by validating our fractal analysis in successively higher dimensions.

When we combine these arguments in two dimensions, each nucleotide represents a vector in the space of possible sequences. Complementarity, reflection, and compatibility are satisfied by any single-base representation. However, we expect that {T} = – {A} and {G} = –{C} from substitution symmetry. So, we have

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

For dimers, complementary requires that {TT} = –{AA}, {TG} = –{AC}, {TC} = –{AG}, {TA} = –{AT}, {GT} = –{CA}, {GG} = –{CC}, {GC} = –{CG}, and {GA} = –{CT}, as well as {TG} = {GT}, {TC} = {CT}, {TA} = {AT}, {GC} = {CG}, {GA} = {AG}, and {CA} = {AC} because of reflection symmetry. So, we can group {GT}, {TG}, {AC}, and {CA} on one axis, as well as {CT}, {TC}, {AG}, and {GA} on the other. Moreover, {GT} = {TG} = –{AC} = –{CA}, and {CT} = {TC} = –{AG} = –{GA} by substitution, and they must differ from zero for compatibility. Then, we have in four dimensions:

Axis 1: {TT} = (1,0,0,0) and {AA} = (–1,0,0,0);
Axis 2: {GG} = (0,1,0,0) and {CC} = (0,–1,0,0);
Axis 3: {GT} = {TG} = (0,0,1,0)
and {AC} = {CA} = (0,0,–1,0);
Axis 4: {CT} = {TC} = (0,0,0,1)
and {AG} = {GA} = (0,0,0,–1);
No axis: {AT} = {TA} = {GC} = {CG} = (0,0,0,0).

Here, axes 1 and 2 represent dimers with four and six hydrogen bonds, respectively, because the basepair A and T has two hydrogen bonds and C and G has three hydrogen bonds. Axes 3 and 4 are dimers with five hydrogen bonds. Finally, {AT} and {TA} have four hydrogen bonds and {GC} and {CG} have six hydrogen bonds. So, in six dimensions, two choices are possible: We can split the five hydrogen bonds, axes 3 and 4, in four or we can construct the new axes from the dimers {AT}, {TA}, {GC}, and {CG}. Because the choice is arbitrary, we will use the Berthelsen et al. prescription for the six-dimensional walk *(20)*:

Axis 1: {TT} = (1,0,0,0,0,0) and
{AA} = (–1,0,0,0,0,0);
Axis 2: {GG} = (0,1,0,0,0,0) and
{CC} = (0,–1,0,0,0,0);
Axis 3: {CA} = (0,0,1,0,0,0) and
{AC} = (0,0,–1,0,0,0);
Axis 4: {TG} = (0,0,0,1,0,0) and
{GT} = (0,0,0,–1,0,0);
Axis 5: {GA} = (0,0,0,0,1,0) and
{AG} = (0,0,0,0,–1,0);
Axis 6: {TC} = (0,0,0,0,0,1) and
{CT} = (0,0,0,0,0,–1),
No axis: {AT} = {TA} = {GC} = {CG} =(0,0,0,0).

Finally, we construct axes 7 and 8 in eight dimensions from the dimers {AT}, {TA}, {GC}, and {CG}:

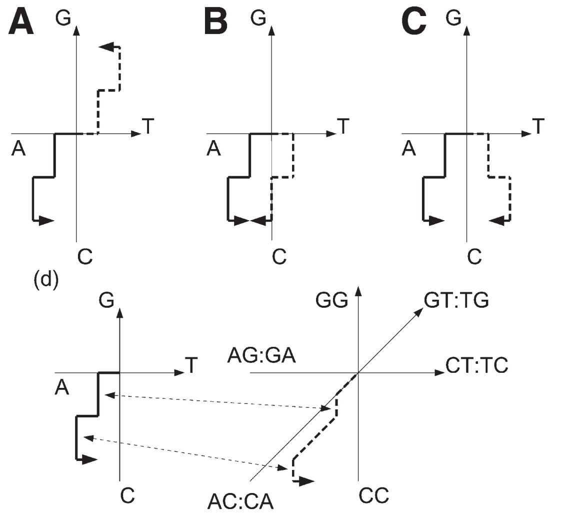

Fig. 1. For the sequence 5′→ACCACCT→3′, we have the following symmetries in the walk: **(A)** complementarity, 5′→TGGTGGA→3′; **(B)** reflection, 5′→TCCACCA→3′; **(C)** substitution of A for T and vice versa, 5′→TCCTCCA→3′; and **(D)** and compatibility between two and three dimensions.

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

To illustrate the DNA walk symmetries, consider the sequence 5′→ACCACCT→3′. Here, 5′→3′ indicates that we are reading the sequence from the fifth to third carbon of the sugar deoxyribose of the DNA chain. Figure 1A represents the complementary sequence by dashed

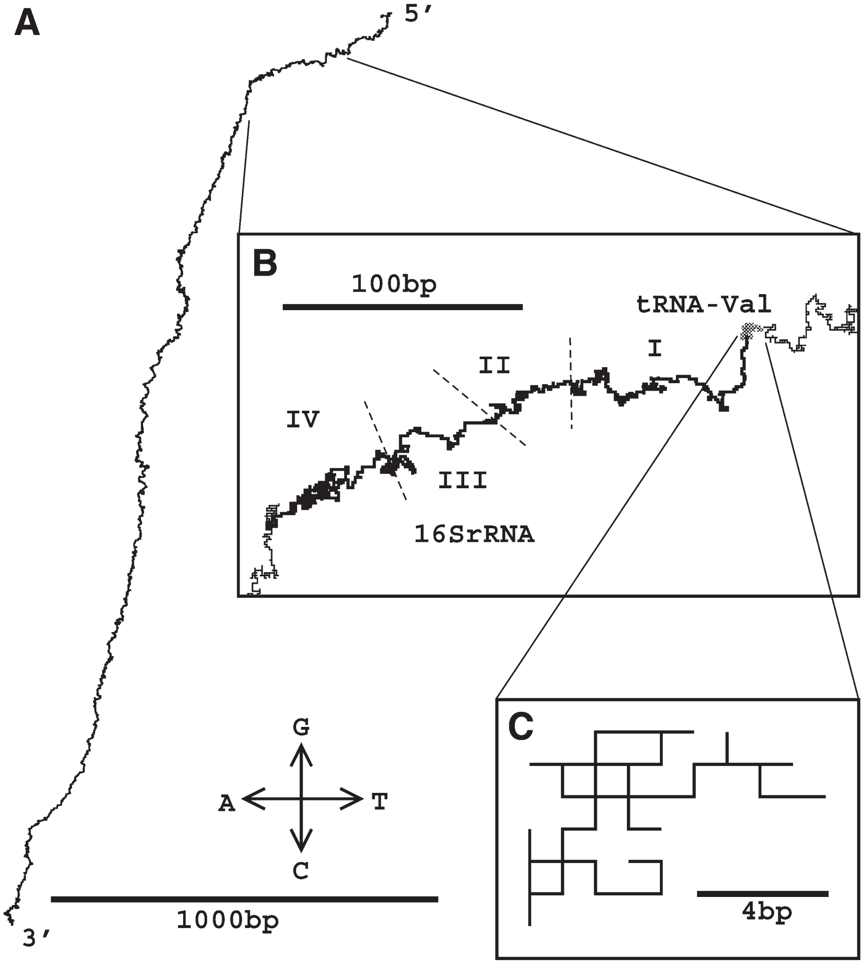

Fig. 2. **(A)** Two-dimensional walk for the mtDNA of the finback whale. We expand the segment around the 16S rRNA in **(B)**. **(C)** Shows the two-dimensional walk for the tRNA for valine, marked in gray in **(B)**.

lines. Figure 1B represents the reflected sequence. If we replace A by T and vice versa, we have Fig. 1C, where the substituted image is a mirror reflection of the studied object along the chosen axis. In all three cases, the symmetry arguments preserve the shape of the original object. Finally, Fig.1D shows the compatibility of ACCACCT from two to three dimensions. We omit the axis AA:TT because of the absence of these dimers in ACCACCT. Although the three-dimensional object (dashed lines) distorts along the axis AC:CA, the shape is the same as the two-dimensional representation of the sequence.

Applying these criteria to mitochondrial DNA, we obtain walks like those in Fig. 2A (e.g., the two-dimensional walk for the finback whale complete mitochondrial genome). If we blow up the walk around the 16S rRNA, we obtain profiles like those in Fig. 2B. If we again enlarge the walk at the beginning of the 16S rRNA sequence, we find structures like those in Fig. 2C for the valine tRNA. Gates reported these clusterlike structures in 1986 *(44)*. Not coincidently, the curly structures along the walk for complete mtDNA in Fig. 2A resemble the entanglements in Fig. 2B. In the following, we measure the scaling laws (fractal dimensions) for these clusterlike structures from their Lévy flights (DNA walks).

The fractal dimension is an efficient tool for characterizing objects with scaling laws *(44–47)*. We define the fractal dimension of an object D_{zero} as $N \approx \ell^{-D_{\text{zero}}}$, where N is the number of boxes of side ℓ required to cover the object. D_{zero} measures how one piece resembles another or the whole. The fractal dimension gives the maximum number of degrees of freedom (number of equations in a hypothetical model in a physical sense) via Takens' theorem *(48)*, as we will see later. These methods also allow us to identify the structures responsible for the self-similarity. To avoid biases typical of particular calculation methods, we apply three different algorithms to our data: box counting, moving boxes, and sandbox.

The box-counting algorithms estimate the fractal dimension by covering the walk with a set of boxes fixed to a grid *(49–51)*. Box counting usually overestimates the number of boxes required to cover, yielding the capacity dimension rather than the fractal (or Hausdorff) dimension *(45)*. Only when the covering is ideal, with the minimum number of boxes, does the capacity dimension coincide with the fractal dimension. However, for simplicity, we speak of the generalized capacity as a generalized fractal dimension because this error is small in the cases we examine.

The moving-box algorithm, which improves on box-counting, also covers the walk with a set of boxes, but independent of a grid *(24,52,53)*. For example, in Fig. 3, we have the same walk as Fig. 2A, but covered by boxes of size 16 bp using moving-boxes. We show the number of points in each box along the z-axis. This three-dimensional picture is a histogram in which dense regions appear as spikes along the walk. Instead of a qualitative and subjective description of this irregular object, we can quantify it using multifractal analysis: fractal dimensions and singularity spectra.

We estimate generalized fractal dimensions using box-counting and moving-boxes using a simple linear fit to:

$$\frac{1}{(q-1)} \ln \sum_{j=1}^{N} p_j^q = D_q \ln \ell_o + b, \tag{1}$$

where D_q is the generalized fractal dimension, N is the number of boxes, $p_j = m_j/n$ is the probability density in box j, m_j is the number of points in box j, n is the total number of steps in the walk, ℓ=$L_o/L_{\max}$, L_o is the box size in bp, $L_{\max}$ is the size of the walk (the largest minus the smallest values in the walk), and b is a parameter, which we will discuss later. D_q in Eq. (1) is computed using the least squares method *(54)*. q is a variable that weights the probabilities of the rarefied or heavy parts of the DNA walk. Negative q selects the sparse pieces of the Lévy flight. Positive qs, select the densest parts. q= 0 weights all parts the same, treating only the shape of the walk. The light and heavy segments of the Lévy flight need not have the same scalings. In this case, we have a multiscaling object, called a multifractal.

Returning to Eq. (1), if we fix q = 0, D_{zero}=–ln(Ne^b)/ln ℓ_o. Comparing this expression with the usual D_{zero} definition *(45)*, we have an extra term, e^b, in the sum. On the other hand, we know the number of boxes N in box-counting methods is usually larger than the ideal number of boxes N'. Thus, the capacity dimension D_{zero} will be the Hausdorff fractal dimension only if $N' = Ne^b$, where $b \le 0$. b moves closer to zero as the box covering becomes more efficient. When the efficiency of the algorithm is maximal and the studied object is covered with the smallest number of boxes, $b = 0$. This variable b, called "the covering efficiency of the algorithm," measures how near we are to the ideal covering *(24)*. Our numerical tests reveal that the linear fit for D_q works well for $-1 \le b \le 0$.

The numerical precision n_{bits} (the number of bits that we use to represent the walk) and the number n of steps also influence the choice of the region used for the fit in Eq. (1). These limits, shown in Fig. 4, are given by:

$$n_{bits} \le x \le 0, \quad -\log_2 n \le y \le 0. \tag{2}$$

The third method (sandbox) takes random samples from the walk *(54)* and computes D_q using a linear fit (55) to:

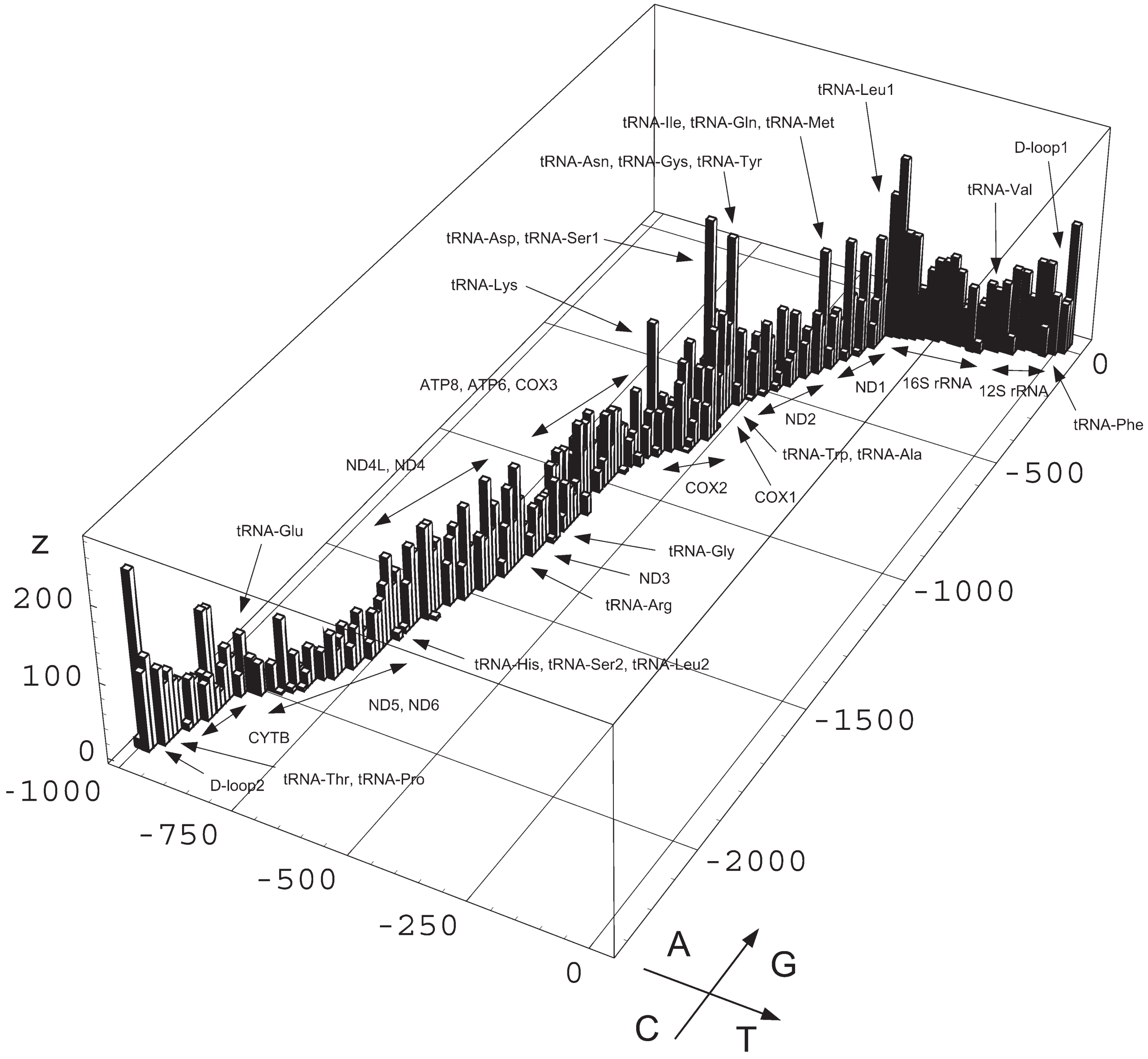

Fig. 3. The finback whale mtDNA two-dimensional walk in Fig. 2A covered by boxes of side 16 bp, using the moving-box algorithm *(24)*. Here, the *z*-axis represents the number of steps in each box.

$$\frac{1}{(q-1)}\ln\left\langle p_j^q \right\rangle D_q \ln \ell_o + b, \qquad (3)$$

where $<p_j^q> = (1/N') \sum_{j=1}^{N'} p_j^{q-1}$, N' is the number of random samples, $p_j = m_j/n$ and m_j is the number of steps within a circle of radius ℓ_o in bp centered around the jth sampled point. For the sandbox method, D_q is an average slope of $\ln \ell_o$ vs $\ln<p_j^q>$. We also consider the numerical precision n_{bits} and the number of steps n in the choice of the region used for the linear fit in Eq. (2).

Figure 4 shows a typical graph for estimating D_{zero}, applying the moving-box (squares) and

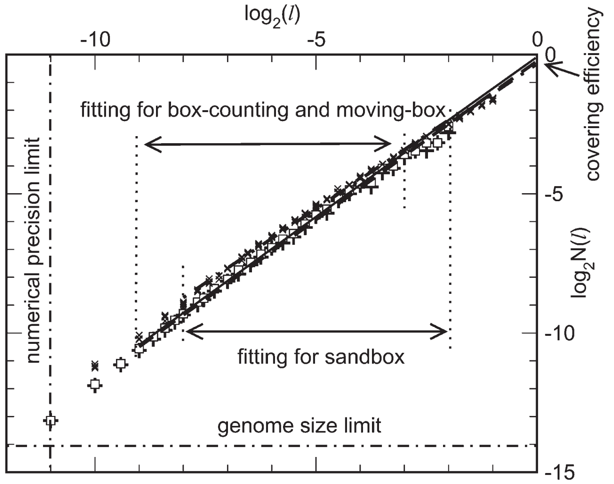

Fig. 4. $-\log_2 N(l)$ vs $\log_2 (l)$ plot for the two-dimensional walk of finback whale mtDNA. We use squares for moving-box and +'s for box-counting and x's for sandbox. The linear-slope fit gives D_{zero}. The intersection of the fit with the y-axis gives the covering efficiency b of the algorithm, and the dot-dashed lines show the limits of the applicability of the methods owing to the numerical precision n_{bits} and genome size n and the dotted lines are the chosen regions for the linear fitting at the fractal dimension estimates.

box-counting (+'s) and sandbox (x's) methods to the two-dimensional walk for the finback whale (Fig. 2). Because this walk has 16,372 steps, the lower limit on the y-axis in Fig. 2 is $\log_2(16{,}372) \approx 14.0$ and the numerical precision limit on the x-axis is 11, because the walk along the C:G axis in Fig. 2 needs 11 bits to represent all steps. Next, we look for the linear scaling region of the logarithmic graph, considering that the covering efficiency factor b must be between -1 and 0 for all methods. b near zero means excellent covering of the Lévy flight for the multifractal analysis, whereas b below -1 indicates that we must be cautious interpreting our results because of poor covering. The covering for the finback whale mtDNA is excellent because $b = -0.23 \pm 0.07$ (box-counting), -0.08 ± 0.06 (moving-box), and -0.22 ± 0.03 (sandbox). The linear scaling region ranges from –3 to –9 in $\log_2(\ell_o)$ for finback whale mtDNA. So, the smallest and biggest box sizes are respectively 1/512 and 1/8 (i.e., we have linear scaling for two decades in the two-dimensional walk of the finback whale mtDNA). This range is unusually large for experimental data, where we usually accept the fit if it applies over a range of more than one decade. In addition, $D_{\text{zero}} = 1.14 \pm 0.01$, 1.15 ± 0.01, 1.064 ± 0.005 for box-counting, moving-box, and sandbox, respectively. The difference of 7.5% between the sandbox and other methods is the result of the algorithm peculiarities and limitations, measured by b. These errors for D_{zero} are very small for experimental data. We can obtain such precision only for a true self-similar object.

Finally, when we change q, the quality of the linear fit decreases, as seen in the D_q and b curves in Fig. 5A,D. The best fits are near $q = 0$, because b is near zero. We can calculate D_qs for $q < 0$. However, the error revealed by b shows that only sandbox results are meaningful.

Although we could obtain all scaling laws for rarified (negative qs) and dense parts of the DNA walk (positive qs) using the D_q vs q curve, we introduce the concept of a singularity α and the singularity spectrum $f(\alpha)$. Like D_{zero}, the singularity is a scaling law, $p_j \approx \ell_o^{-\alpha_j}$, where p_j is the number of points or probability density within a

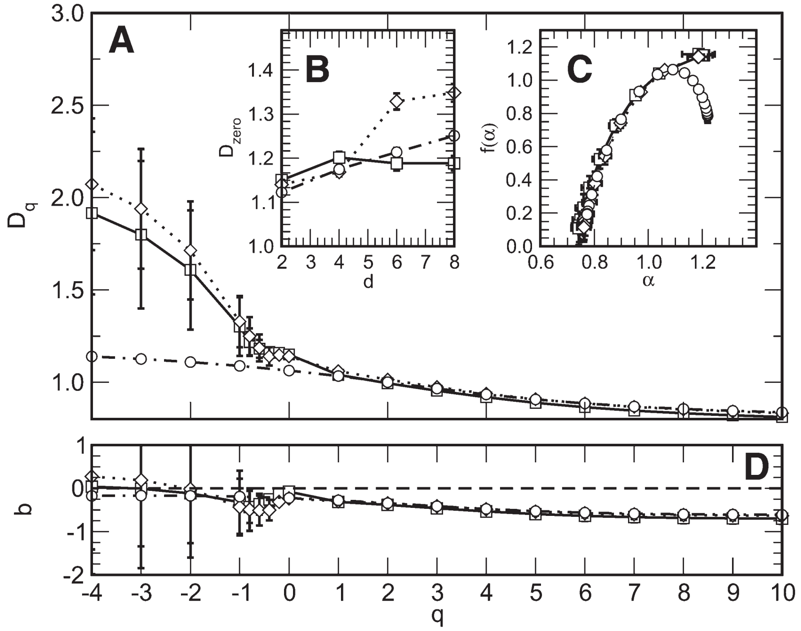

Fig. 5. **(A)** Generalized fractal dimension D_q, **(B)** fractal dimension D_{zero} varying the embedding dimension d, **(C)** singularity spectrum $f(\alpha)$, and **(D)** covering efficiency b for finback whale mtDNA using box-counting (dotted lines and diamonds), moving-box algorithms (solid lines and squares), and sandbox algorithm (dotted-solid lines and circles).

box of side ℓ_o around a given point. We define the number of boxes with scaling α as $N(\alpha)$ and define $N(\alpha) \approx \ell_o{}^{-f(\alpha)}$, $f(\alpha)$=–ln $N(\alpha)$/ln ℓ_o. $f(\alpha)$ is an homogenous function, with important thermodynamic consequences *(45–47)*. In the generalized fractal dimension, the relation between p_j and D_q depends on the sum of N boxes in Eq. (1), so we cannot identify the boxes responsible for one particular value of D_q.

We could estimate the singularity spectrum $f(\alpha)$ using the Legendre transform of the curve $(q - 1)D_q$,

$$\begin{aligned} f(\alpha) &= \alpha q - (q-1)D_q \\ \alpha &= \frac{\partial[(q-1)D_q]}{\partial q}. \end{aligned} \tag{4}$$

We call this approach microcanonical in the thermodynamic sense, because we compute D_q from the local probabilities p_j. However, we obtain better results using a canonical approach *(56,57)*, instead of the Legendre transform in Eq. (4). Thus, we compute α and $f(\alpha)$ directly using either box counting or moving-box using a linear fit to:

$$\begin{aligned} \sum_{j=1}^{N} \mu_j \ln p_j &= \alpha \ln \ell_o + b + (q-1)\frac{\partial b}{\partial q}, \\ \sum_{i=1}^{N} \mu_j \ln \mu_j &= f(\alpha) \ln \ell_o + b + q(q-1)\frac{\partial b}{\partial q}, \end{aligned} \tag{5}$$

where $\mu_j = p_j{}^q \;/\; \sum_{j'=1}^{N'} p_{j'}{}^q$. Using Eq. (5) avoids the numerical estimation errors of the slope of the $(q$ - $1)D_q$ curve in Eq. (4). For the sandbox algorithm, we have a linear fit to:

$$\begin{aligned} \langle \mu_j \ln p_j \rangle &= \alpha \ln \ell_o + b + (q-1)\frac{\partial b}{\partial q}, \\ \langle \mu_j \ln \mu_j \rangle &= f(\alpha) \ln \ell_o + b + q(q-1)\frac{\partial b}{\partial q}, \end{aligned} \tag{6}$$

where $\mu_j = p_j{}^q / {<}p_{j'}{}^q{>}$, ${<}\mu_j$ ln $p_j{>}$ = $(1/N')\ \sum_{j=1}^{N'} \mu$ ln p_j and ${<}\mu_j$ ln $\mu_j{>}$ =$(1/N')\ \sum_{j=1}^{N'} \mu_j$ ln μ_j. The limits in Eq. (2), used for the linear fit for D_q, also work here and can be calculated in the same way.

When we apply box-counting, moving-box, and sandbox methods to finback whale mtDNA, we find the singularity spectrum shown in Fig, 5C. Results using all methods agree, including the positions of the qs along the curve $f(\alpha(q))$. α is a function of q, $\alpha(q)$. So, $f(\alpha(q))$ also provides information about q. The smallest αs represent the largest q*s* (i.e., the densest parts of the walk). In the case of finback whale mtDNA, $f(\alpha)$ goes to zero at $\alpha_{\max}$=0.76 ± 0.03, 0.74 ± 0.03, 0.77 ± 0.06 respectively for box-counting, moving-box, and sandbox, indicating that the dense parts of the walk are fractals resembling a generalized Cantor set *(45)*. On the other hand, the maximum of the curve is always at $f(\alpha(q = 0))$ and tells us about the general shape of the walk. In particular, $D_{\text{zero}}=f(\alpha(q = 0))$. In Fig. 5C, $f(\alpha(q = 0))$ = 1.14 ± 0.01, 1.15 ± 0.01, and 1.064 ± 0.005 for box-counting, moving-box, and sandbox, respectively, and they coincide with the values of D_{zero}, as expected. Again, the value for $f(\alpha(q = 0))$ using sandbox is smaller than for other algorithms. We interpret these values in the following way: Because D_{zero} is near unity, the walk is lineshaped, with some extra structure because D_{zero} is larger than 1. We can analyze the sparse parts of the walk by studying negative qs. This information is only available using the sandbox method, and the values are $(\alpha_{\min}, f(\alpha_{\min}))$ = (1.220 ± 0.009, 0.79 ± 0.04). The $f(\alpha)$ curve does not reach the x-axis for negative q; therefore, either the most rarefied parts of the finback whale mtDNA walk do not scale or we simply lack sufficient statistics for the analysis. For box counting and moving box, b indicates that these methods are not reliable for qs much below zero. Thus, the $f(\alpha)$ curve in Fig. 5C shows definitively that the content of finback whale mtDNA is multifractal.

RESULTS AND DISCUSSION

Our previous work showed that the multifractality of mtDNA comes from the self-similar distribution of inverted repeat nucleotide sequences (palindromes) *(25)*. These palindromes might fold the single stranded DNA (ssDNA) in hairpins, loops, and other unnamed structures. Next, we discuss this conclusion in more detail using the results in Table 1 and Fig. 6. We also estimated the generalized fractal dimensions D_q vs q for all mtDNA in Table 1, but we omit them, because the $f(\alpha)$ spectra (Fig. 6) provide complete multiscaling information. To reduce clutter, Fig. 6 shows only sandbox-derived $f(\alpha)$ spectra. Box-counting and moving-box results are similar.

We can estimate D_q for any object, fractal or not. However, we obtain high-quality linear fits only if the object is a true fractal. For Brownian noise, D_{zero} is close to the embedding dimension d, but for the mtDNA walks, D_{zero} saturates around 1.2 as the embedding dimension increases in Fig. 5b. Finally, for simple Euclidian objects like a line or plane, D_{zero} is an integer, independent of d. If the walk were not a Euclidian geometric object, fractal, or Brownian noise, any attempt to estimate D_q or $f(\alpha)$ would result in large errors or unreliable values for b.

Because the walk might cross or follow a previously traced path multiple times, we might overestimate the number of steps in a box in our D_q estimates, Eq. (1). Checking the saturation of D_{zero} with d is particularly important to avoid false neighbors in fractal analysis *(58)*. A point could accidentally be close to another point: for example, when we fold a circle along its diameter to produce a one-dimensional object, we bring diametrically opposite points into correspondence. On the other hand, these two points will be far apart if we embed the same circle in two or more dimensions, eliminating the false neighbors.

For mtDNA walks, we can estimate D_q and $f(\alpha)$ with errors much lower than those usually associated with experimental data (less than 1%). Furthermore, the fits of D_{zero}, as in Fig. 4, have a large linear scaling region. D_{zero} always lies between 1 and 2. Takens and Mañé's theorem suggests that we need to embed our DNA walk in a maximum of three or four dimensions, as $d < 2D_{\text{zero}} + 1$ *(48,59)*. Finally, the saturation of D_{zero} around 1.2 in Fig. 5B also indicates that the hidden structure has at most

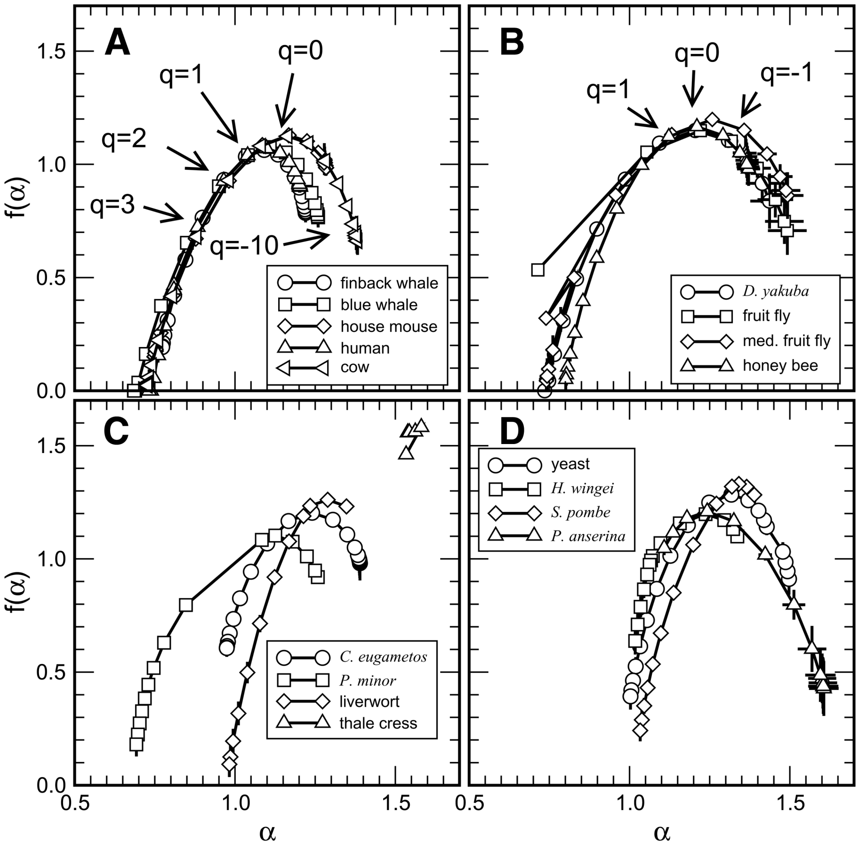

Fig. 6. Singularity spectra $f(\alpha)$ using the sandbox algorithm for mammals **(A)**, insects **(B)**, plants **(C),** and fungi **(D)**. q ranges from 0 to 3 for thale cress. The qs lie within intervals of one unit for other spectra.

two dimensions. Therefore, $d = 2$ suffices to describe the hidden structure.

We can identify this hidden two-dimensional DNA organization. According to Table 1, the two-dimensional walk for mtDNA is line-shaped, because D_{zero} is near 1, as we see in Fig. 2. The D_{zero} values around 1.2 (except for thale cress mtDNA) indicate some extra structure in addition to the line-shaped walk. We can get more information about this extra structure by analyzing the distribution of dense parts of the walk in Fig. 3. We always find spikes in Fig. 3 for each tRNA and rRNA because these molecules have a huge number of cross-shaped structures and loops in their spatial structure *(1,27)*. Looplike structures are two-dimensional double-stranded domains responsible for secondary bonds and for the ribosomal spatial conformation. Double-stranded DNA (dsDNA) does not change the shape of the walk because the displacement of one strand cancels that of the opposite strand. (For example, the complementary sequence 3′→TGGTGGA→5′ in Fig.1A cancels the direct sequence 5′→ACCACCT→3′) but increases the local density, resulting in spikes. In addition, a naïve random walk for a dsDNA reveals a clusterlike dispersion along the Lévy flight with a variance σ given by, $\sigma = \sqrt{3n_{steps}}/2$ where n_{steps} is the number of steps in the walk *(11)*. So each loop creates a clusterlike section in the DNA walk. Thus, valine tRNA, a cross-shaped 67-length DNA sequence

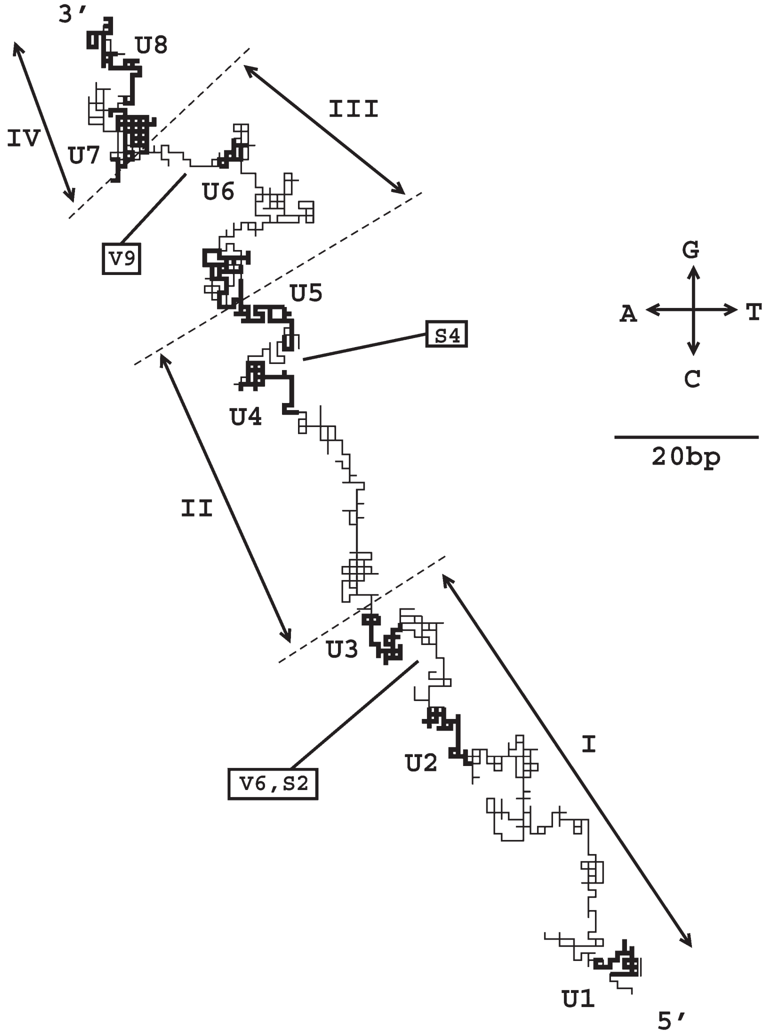

Fig. 7. Two-dimensional DNA walk for 16S rRNA in the bacterium *E. coli*, gene rrsH. The four domains of rRNA are indicated by I, II, III and IV. U1 to U8 are the universal regions in SSU rRNA. We also indicate the variable segments V6 and V9, as well as semiconserved regions S2 and S4. 5′ indicates the beginning of the walk; 3′ indicates its end.

with three loops and one dsDNA *(1)*, is confined to a box of side 9 bp (Fig. 2C).

Ribosomal structures are highly organized, because they are the key molecules for DNA transcription and becase the RNA sequence functions directly, not via translation into protein. Because tRNA is very small (around 70 bp) and its Lévy flight is curly (Fig. 2C), we could not distinguish between different tRNAs by looking at their DNA walks. However, the larger ribosomal DNA sequences have distinctive and highly conserved Lévy flights. The similarity between the 16S rRNA for finback whale mtDNA (Fig. 2B), for the bacterium *Escherchia coli* (Fig. 7) and for *Mus musculus* (Fig. 8) walks is not coincidental. The 16S rRNA of the bacterium *E. coli* always consists of four domains (I to IV) with eight universal regions (U1 to U8).

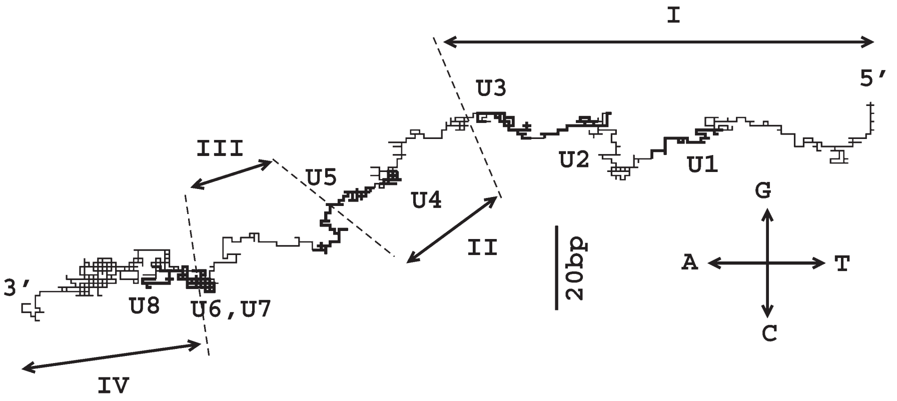

Fig. 8. Two-dimensional DNA walk for mitochondrial 16S rRNA in *M. musculus*. The four domain of the rRNA are indicated by I, II, III, and IV. U1 to U8 are the universal regions in SSU rRNA. The walks for U6 and U7 overlap. The walk starts at 5′ and ends at 3′.

This structure is equivalent in function to the 16S rRNA mtDNA in mammals *(27)*, despite some variation: The *E. coli* variable (V6, V9) and semivariable (S2, S4) structures are absent in *M. musculus*; on the other hand, we easily note the extra sequences between 5′ and U1 and 3′ and U8 in the mouse mtDNA. This structural similarity is the reason that the DNA walk for *E. coli* bacterial 16S rRNA, gene rrsH, is very similar to the 90° rotated walk of the 16S rRNA of finback whale (Fig. 2b) and *M. musculus* mtDNA (Fig. 8). This rotation is equivalent to exchanging G with A, A with C, C with T, and T with G between Fig. 7 and Fig. 8. We do not know the reason for the rotation symmetry in the Lévy flight, but this circular replacement does not affect the function of the rRNA. Obviously, the 90° rotation of the walk will result in different base pairing. As a result, we will not see similarities using more crude methods, like counting single base substitutions. We plan to track such differences in future phylogenetic analysis.

DNA walk multifractal analysis applies to other DNA as well *(25)*. We computed D_q and $f(\alpha)$ for the complete bacterial genome of *E. coli*, strain K-12 MG1655, GenBank accession number U00096, 4,639,221 bp. D_{zero} is 1.29 ± 0.02 (box counting) and 1.29 ± 0.03 (moving box) with $b \approx 0.0$. Unfortunately, we could not estimate D_{zero} using the sandbox algorithm, because it demands too much computation time. We are currently analyzing nonmitochondrial DNA walks. The small error in D_{zero} indicates bacterial nucleotide sequence self-similarity.

The rRNAs are not exceptional in having a looplike organization. Protein-coding DNA sequences can also resemble two-dimensional rRNA-like structures, indicating the presence of a second organization in addition to the amino acid coding (codons). Table 2 shows the DNA sequence of the nicotine adenine dinucleotide dehydrogenase (NADH) subunit 1 protein for finback whale mtDNA. NADH subunit 1 belongs to the NADH coenzyme Q reductase complex, responsible for oxidation reactions. Table 2 shows palindromes at positions 724–733 bp, a hairpin at 166–174 bp, and a loop at 560–583 bp. Our previous article presented a very simple search for complementary *K*-letter words that identified repeats *(25)*. Given a *K*-nucleotide sequence, we wish to find the nearest and largest reflected complementary sequence within the DNA chain. In an infinite random DNA chain, we always find it, because the matching sequence must appear somewhere. However, we are looking for a nonran-

Table 2
DNA Sequence for NADH Dehydrogenase Subunit 1 of Finback Whale mtDNA and Its Reflected Complement

```
1   atgtttataa ttaacattct aacactcatt ctccccatcc tcctagccgt agcattccta
        +++--- ^^vv  >>>> >                      ^^vv   + ++      ---
61  acgctagtag aacgcaaaat tctaggctat atgcagttcc gaaaggggcc aaacatcgta
       ^^vv<<< <<     ^^v v^^vv  ^^v v^^vv +++  ---   ^^vv ---
121 ggcccacatg gcttactcca accctttgcc gatgcaatta aattattcac taaagaaccc
    ^^vv  ^^vv         ++ +    ---     ^^vv ^^v v +++---
181 ctacggccag ctacatcctc aactactatg tttatcattg caccagtact agccctaacc
        ^^vv^^ vv                +++      ---^^ vv  +++---
241 ctggccctca ctatatgaag ccccctaccc ataccatacc ccctcattaa cataaaccta
      ^^vv +++  ^^vv---                               ^^vv  **** +++-
301 ggagtattat tcatattagc aatatccagc ctagccgtct actccatcct atgatcaggc
    --    ====   ^^vv      ^^vv      ^^vv                   +++---***
361 tgagcctcca actcaaaata cgcactaatt ggagccctac gagcagtagc acaaacaatc
    * ====                      ^^vv       ****      ====           >
421  tcatatgagg taacactagc cattatcctc ctatcagtac tcctaataaa cggctcctac
     >>>><<<<<       ^^vv                  +++-- -    ****   ****
481  accttatcaa cattagccac aacacaagaa caactatgat tactattccc atcatgaccc
        ====        ====                     >>> >>          +++---
541  ttagccataa tgtgattcat ctccacccta gcagaaacta atcgagctcc ttttgatcta
          +++ - --+++ ---> >>>>   ^^v v       << <<<+++---      ^^vv
601  acagagggag aatcagaact cgtatcaggc ttcaacgtag aatatgcagc aggccctttc
           <<<< <                         ^^vv    ^^vv       ^^vv
661  gccctattct tcctggcaga atacgccaac atcattataa tgaatatact cacagccatt
         ****     +++ ---= ===         aaaaaaAAA AAA^^vv
721  ttattcctag gaacattcca caaccctcat aacccagaat tgtacacagc aaaccttatt
        >>>>><< <<< ****           ** **    ==== +++---          ====
781  atcaagacac tactactcac aatatccttc ctatgaattc gagcatccta cccccgattc
```

Note: The first column is the position of the first nucleotide of the respective line in bp. The direct and complementary reflected sequences are indicated by + and – for 3 bp; * and = for 4 bp; > and < for 5 bp; a and A for 6 bp.

dom reflected complementary sequence. So, the probability that the right sequence is close to the original sequence by chance is very low. Given a sequence, the probability of the reflected complementary sequence appearing is, $P(r) = (1 - 1/4^K)^r\ 1/4^K$ where r is the number of nucleotides between the given sequence and the expected one, $1/4^K$ is the probability of finding the right sequence, and $(1 - 1/4^K)^r$ is the probability of the nonappearance of the reflected complementary sequence. Table 2 shows the DNA sequence marking complementary sequences between positions $K + 1$ and L, taking sequences with K-nucleotides that start at position i of the DNA chain. Here, i ranges from 1 to $L - K$. We intend to improve our search for palindromes in the future.

Unfortunately, we cannot estimate fractal dimensions or singularity spectra for single isolated rRNAs or genes like NADH subunit 1, because our methods do not give reliable results for short sequences (around 1000 bp). We need at least complete mitochondrial genomes (more than 10,000 nucleotides) for conclusive results.

When we systematically apply fractal analysis to our selected mitochondrial genomes, we obtain the D_{zero}s in Table 1, which measure how curly the Lévy flights are. In Table 1, mammals (finback and blue whales, mouse, rat, human, cow, and seal) have fractal dimension around 1.15, except for the duckbill platypus ($D_{\text{zero}} \approx 1.20$). This difference comes from an extra control region in positions 15,461 bp to 17,019 bp, near phenylalanine tRNA (phe-tRNA), absent in other mammals. The total coding segment length L_{CD} for mammals is around 16,000 bp, but that for the duckbill platypus is 1000 bp larger because of this peculiarity. The Lévy flight for the duckbill platypus is more entangled, with more DNA loops, perhaps, reflecting metabolic rate differences. Although the chicken and alligator also have a D-loop and a control region near phe-tRNA, they are fractal, with D_{zero} around 1.15, indicating that their general organization does not differ dramatically from mammals. However, D_{zero} for *Xenopus laevis* is around 1.20, because of the longer D-loop at the beginning of the DNA sequence, 2134 bp long, near phe-tRNA. Similar to the platypus, the length of coding segments L_{CD} for this frog is 1000 bp larger than for mammals. Fishes (amphioxus, sea lamprey, gummy shark, *Crossotoma lacustre*, and carp) have all D_{zero} around 1.20, independent of the absence (carp, sea lamprey, amphioxus) or presence of the D-loop (*C. lacustre*, gummy shark) near phe-tRNA. Finally, the urchins have fractal dimension around 1.30. Apparently, the fractal dimensions decreases as the animals become more complex and have heavier metabolic demands. Mammals have more elaborate life cycles with longer childhoods than other organisms, and they consume more calories than reptiles, amphibian or fish, because the first group are endothermal and the second group are ectothermal.

Perhaps because flies, honeybees, and earthworms share a common Metazoan ancestor, their D_{zero} is always around 1.20. In contrast, the fractal dimension of nematodes (*Caenorhabditis elegans*, pig roundworm, river blindness worm) is near 1, reflecting their differences from arthropods and annelids.

When we look at more ancient phylogenetic ramifications, we see fungi D_{zero}s are always higher than 1.2; that is, their genome is richer in looplike structures. Finally, the fractal dimension for plants ranges from 1.0 to 1.5. In particular, the fractal dimension of the thale cress is unusually high, around 1.5.

We do not need to restrict our analysis to the shape of the Lévy flight. We can study the distribution of the steps along the walk. If we have a simple fractal with a uniform step distribution along the walk, like a simple Cantor set, $f(\alpha)$ will collapse to a single point, and all qs will cluster together. In practice, most points of the singularity spectrum will concentrate around $q = 0$, but $f(\alpha)$ will display arms for large negative and positive qs, $|q| >> 1$, because of the presence of noise and other variables. For *Arabidopsis thaliana* (*see* Fig. 6c), $f(\alpha)$ collapses around $(\alpha, f(\alpha)) \approx (1.5, 1.5)$ for $0 \leq q \leq 3$. The spike distribution and heights along the DNA walk, in a Fig. 3-like plot, are uniform. We observe such monofractality in the liverwort, *Chlamyclomonas eugametos* and *Pedinomonas minor* $f(\alpha)$s: The negative q branches (right arms of the curves) are shorter than for mammals or insects; the positive q branches (left arms of the curves) do not reach the x-axis. The distribution and heights of spikes along the DNA walk are neither completely uniform nor multiscaling. The spectra for plants do not coincide with each other as do those for mammals. (Fig. 6A), because the chosen species are very distant phylogenetically (two green algae, a liverwort, and a higher plant). Perhaps evolutionary pressures are weaker or metabolism is simpler in plants, because we do not observe this tendency to monofractality in animals.

On the other hand, Fig. 6a shows that mammalian mtDNA sequences are multifractal for both dense $q > 3$ (left side of spectra) and average $0 \leq q \leq 3$ weights, because the positive qs span a large range of αs. $f(\alpha)$s for mammals are remarkably similar to each other, because the curves and the qs coincide with each other for $0 \leq q \leq 3$. Thus, the spike and step distributions are very similar in these five species. Because $(\alpha_{\min}, f(\alpha_{\min})) \approx (1.3, 0.7)$, not all scaling laws occur for these Lévy flights; that is, the boxes for nonloop segments contain a minimum number of steps, and the distribution of the rarefied boxes along the walk has a fractal dimension $D_{-\infty} = \alpha_{\min}$ around 1.3. Below this minimum scaling $f(\alpha)$, nothing occurs. However, the spectra reach the x-axis for large qs (dense parts of the DNA walk), indicating that we have all scaling laws; that is, we have looplike structures of all lengths. The distribution of such loops along the walk is a fractal with dimension around 0.7, similar to dotted lines or a Cantor set *(45)*.

Insects spectra also resemble those of mammals (Fig. 6B), except for *Drosophila melanogaster*. The $f(\alpha(q))$ curves of insects except for *D. melanogaster* coincide with each other, and they also exhibit multifractality. $D_{\text{zero}} = f(\alpha(q = 0)) \approx 1.2$ is higher than for mammals, indicating that the entanglement of loops is more complex. In the case of the fruit fly, $f(\alpha)$ is very broad, because of an extra structure in its genome. We can see in Table 1 that the sequence (L = 19,517 bp) is longer than for other flies and mammals (L around 16,000 bp for both). An extra $(AT)_n$ region contains the replication origin and two deoxythymidylate stretches, 4601 bp long, at the end of the mtDNA. This $(AT)_n$ sequence gives us a left–right movement, resulting in a spike in n at this point of the walk. This spike destroys the multiscaling law for dense parts of the fruit fly DNA walk, resulting in an $f(\alpha)$ curve where the left arm does not reach the x-axis in Fig. 6b. We do not know the function of this $(AT)_n$ sequence, but *Drosophila yakuba* has a similar structure: 1076 bp long. In the case of *D. yakuba*, the region is smaller and does not dramatically change the singularity spectrum, indicating that $f(\alpha)$ could be useful for detecting anomalous repeats like $(AT)_n$ domains.

Fungi are usually intermediate between animals and plants (*see* Fig. 6D), as we expect phylogenetically. Their $f(\alpha)$s sometimes look like an animal spectrum, as we see for *Schizosaccharomyces pombe*; that for yeast and *H. wingei* resembles that of plants. However, the *Podospora anserine* spectrum resembles neither that of animals nor plants. The branch for the positive qs is truncated, indicating few DNA looplike structures, and we observe a well-developed negative-q arm in the spectrum. Because we have $f(\alpha)$ spectra for only a few fungi, we cannot yet identify a characteristic fungal spectrum.

mtDNA walks of plants are simpler (more homogeneous) than those of mammals because the $f(\alpha)$s of plants look like simple monofractals near $q = 0$: that is, they have just one scale if we weight the rarefied and dense parts of the DNA walk in the same way. We do not know the reason for this behavior, especially because most of the genes are the same. Possibly plant mtDNA organization is simpler than in animals, because the metabolic rates in plants are smaller than in animals or because plants are less sensitive to random mutations in their mitochondria.

When we mention simplicity or complexity of the genetic organization in this work, we are talking about the scaling laws (fractality) of the distribution of nucleotide repetitions, measured by $f(\alpha)$s. Our next challenge is to identify the other structures responsible for these characteristic repeat distributions, beyond the DNA loops.

That the mitochondrial genome is a true fractal is evidence of a simple genomic grammar (rules for DNA coding), because we can build elaborate sequences by duplication of simple nucleotide repeats (palindromes). Duplications are a common mutation mechanism. Although we observe single-nucleotide replacements in long DNA sequences, we expect that these duplications will leave a scar in the genome. When we study $f(\alpha)$s, we can characterize the self-similarity of the distribution of the repeats resulting from these duplications.

Self-similarity in nucleotide sequences has consequences for shotgun sequencing too. Simulated nucleotide sequences are usually random or periodic, but not fractal *(5,7,9,10)*. All shotgun sequencing is based on the assumption of a random distribution of nucleotides. Redundant sequencing usually reduces misassembly as a result of repeats. However, we observe self-similarities resulting from the nonperiodic repetitions of palindromes. These repetitions are not uniformly distributed along the nucleotide sequence, but compose clusters that appear as spikes along the DNA walk (Fig. 3). Because all shotgun sequencing methods require sequence overlaps, we expect gaps in highly-repetitive sequences, as we indeed observe for long eukaryotic nucleotide sequences. Because these self-similarities result in gaps and clonally poorly-covered nucleotide sequences, we must consider them in developing new strategies for better sequence assembly.

CONCLUSION

The distribution of looplike spatial structures in mitochondrial genomes creates a self-similar (fractal) pattern. This fractal organization produces well-defined DNA walks (Lévy flights), like those of 16S rRNA in *E. coli* and the finback whale and *M. musculus* mitochondrial genomes. The Lévy flight technique clearly reveals the similarities and differences between rRNA sequences, even when the nucleotide sequences are reordered, mirrored, or have consistent substitutions of nucleotides, and so forth. Mammals' mtDNA flights have the same fractal dimension (around 1.15) and a characteristic singularity spectrum, indicating that their genomes have distributed looplike structures in all lengths. However, we do not expect multiscaling for the smallest unidentified features of mtDNA, because $f(\alpha)$ is truncated for negative qs. Plants tend to have monofractal $f(\alpha)$. Fungi have intermediate spectra, as expected from their phylogeny.

Multifractal analysis of DNA walks reveals a nonlinear organization in the mitochondrial genome. However, this work did not incorporate information from other traditional sequence analysis methods, like the C + G content variation ratio or purine–pyrimidine walks. We hope to find additional hidden genomic features by applying multiple techniques simultaneously.

ACKNOWLEDGMENTS

This work was supported by the following: Fundação de Amparo à Pesquisa do Estado de São Paulo, Brazil, under grants 1999/08479-4 and 2001/05839-1; National Science Foundation and Conselho, Division of Integrative Biology, United States of America, IBN-0083653; NASA Glenn Research Center, USA, NAG3-2366; Department of Energy, USA, DE-FGO299ER45785; National Science Foundation and Conselho Nacional de Desenvolvimento Científico e Tecnológico International Award, USA–Brazil, INT98-02417.

REFERENCES

1. Watson, J. D. Hopkins, N. H., Roberts, J. W., Steiz, J. A., and Weiner, A. M. (1987) *Molecular Biology of the Gene*, 4th ed., Benjamin/Cummings, San Francisco CA.
2. Venter, J. C., Adams. M. D., Myers, E. W. et al. (2001) The sequence of the human genome. *Science* **291,** 1304–1351.
3. Lander, E. S., Linton, L. M., Birren, B., et. al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921.
4. McPheeters, D. S., Christensen, A., Young, E. T., Stormo, G., and Gold, L. (1986) Translational regulation of expression of the bacteriophage T4 lysozyme gene. *Nucleic Acids Res*. **14,** 5813–5826.
5. Venter, J. C., Adams, M. D., Sutton, G. G., Kerlavage, A. R., Smith, H .O., and Humkapiller, M. (1998) Shotgun sequencing of the human genome. *Science* **278,** 1547–1551 .
6. Myers, E. W., Sutton, G. G., Smith, H. O., Adams, M. D., and Venter, J. C. (2002) On the sequencing and assembly of the human genome. *PNAS* **99,** 4145–4146.
7. Weber, J. L. and Myers, E. W. (1997) Human whole-genome shotgun sequencing. *Genome Res*. **7,** 401–409.
8. Waterston, R. H., Lander, E. S., and Sulston, J. E. (2002) On the sequencing of the human genome. *PNAS* **99,** 3712–3716.
9. Green, P. (1997) Against a whole-genome shotgun. *Genome Res.* **7,** 410–417.
10. Green, P. (2002) Whole-genome disassembly. *PNAS* **99,** 4143–4144.
11. Oiwa, N. N. and Goldman, C. (2000) Phylogenetic study of the spatial distribution of protein-coding and control segments in DNA chains. *Phys. Rev. Lett*. **85,** 2396–2399.
12. Oiwa, N. N. and Goldman, C. *Cell Biochem. Biophys.*, in press.

13. Nicolis, G. and Prigogine, I. (1989) *Exploring Complexity*, W. H. Freeman, New York.
14. Haken, H. (1988) *Information and Self-Organization: A Macroscopic Approach to Complex Systems*, Spring-Verlag, Berlin.
15. Arnéodo, A., d'Aubenton-Carafa, Y., Bacry, E., Graves, P. V., Muzy, J. F., and Thermes, C. (1996) Wavelet based fractal analysis of DNA sequences. *Physica D* **96,** 291–320.
16. Peng, C.-K., Buldyrev, S. V., Goldberger, A. L., et al. (1992) Fractal landscape analysis of DNA walks. *Physica A* **191,** 25–29.
17. Peng, C.-K., Buldyrev, S. V., Goldberger, A. L., et al. (1992) Long-range correlations in nucleotide sequences, *Nature* **356,** 168–170.
18. Buldyrev, S. V., Goldberger, A. L., Havlin, S., et al. (1993) Fractal landscapes and molecular evolution: modeling the myosin heavy chain gene family., *Biophys. J.* **65,** 2673–2679.
19. Buldyrev, S. V., Goldberger, A. L., Havlin, S., Peng, C.-K., Simons, M., and Stanley, H. E. (1993) Generalized Lévy-walk model for DNA nucleotide sequences. *Phys. Rev. E* **47,** 4514.
20. Berthelsen, C. L., Glazier J. A., and Skolnick, M. H. (1992) Global fractal dimension of human DNA sequences treated as pseudorandom walks. *Phys. Rev. A* **45,** 8902–8913.
21. Berthelsen, C. L., Glazier J. A., and Raghavachari, S. (1994) Effective multifractal spectrum of a random walk. *Phys. Rev. E* **49,** 1860–1864.
22. Glazier, J. A., Raghavachari, S., Berthelsen, C. L., and Skolnick, M. H. (1995) Reconstructing phylogeny from the multifractal spectrum of mitochondrial DNA. *Phys. Rev. E* **51,** 2665–2668.
23. Purugganan, M. D. (1993) Scale-invariant spatial patterns in genome organization. *Phys. Lett. A* **175,** 252–256.
24. Oiwa, N. N. and Fiedler-Ferrara, N. (1998) A moving-box algorithm to estimate generalized dimensions and the f(α) spectrum. *Physica D* **124,** 210–224.
25. Oiwa, N. N. and Glazier, J. A. (2002) The fractal structure of the mitochondrial genomes. *Physica A* **311,** 221–230.
26. Benson, D. A, Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A., and Wheeler, D. L. (2000) GenBank. *Nucleic Acids Res.* **28,** 15–18.
27. Gray, M. W., Sankoff, D., and Cedergreen, R. J. (1984) On the evolutionary descent of organisms and organelles: a global phylogeny based on a highly conserved structural core in small subunit ribosomal RNA. *Nucleic Acids Res.* **12,** 5837–5852.
28. Li, W. (1997) The study of correlation structures of DNA sequences: a critical review. *Computers Chem*. **21,** 257–271.
29. Setubal, J. and Meidanis, J. (1997) *Introduction to Computational Molecular Biology*, PWS Publishing, Boston.
30. Baxevanis, A. D. and Ouellete, B. F. F., eds. (2001) *Bioinformatics*, 2nd ed., Wiley, New York.
31. Goto, S., Nishioka, T., and Kanehisa, M., (1998) LIGAND: Chemical Database for Enzyme Reactions. *Bioinformatics* **14,** 591–599.
32. Bussemaker, H. J., Li, H., and Siggia, E. D. (2001) Regulatory element detection using correlation with expression. *Nature Genet*. **27,** 167–171.
33. Bussemaker, H. J., Li, H., and Siggia, E. D. (2000) Buiding a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *PNAS* **97,** 10,096–10,100.
34. B.-L. Hao, - Lee, H. C., and Zhang, S.-Y. (2000) Fractals related to long DNA sequences and complete genomes. *Chaos Solitons Fractals* **11,** 825; Yu, Z.-G., Hao, B. L., Xie, H. M., and Chen, G. Y. (2000) Dimensions of fractals related to languages defined by tagged strings in complete genomes. *Chaos Solitons Fractals* **11,** 2215.
35. Kolwalczuk, M., Gierlik, A., Mackiewicz, P., Cebrat, S., and Dudek, M. R. (1999) Optimization of gene sequences under constant mutational pressure and selection. *Physica A* **273,** 116.
36. Mantegna, R. N., Buldyrev, S. V., Goldberger, A. L., et al. (1994) Linguistic features of noncoding DNA sequences. *Phys. Rev. Lett.* **73,** 3169–3172; Mantegna, R. N., Buldyrev, S. V., Goldberger, A. L. et al. (1995) Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics. *Phys. Rev. E* **52,** 2939–2950; Mantegna, R. N., Bulyrev, S. V., Goldberger, A. L. et al.(1996) Reply. *Phys. Rev. Lett.* **76,** 1979–1981.
37. Israeloff, N. E., Kaganlenko, M., and Chan, K. (1996) Can Zipf distinguish language from noise in noncoding DNA.*Phys. Rev. Lett.* **76,** 1976; Bonhoeffer, S., Herz, A. V. M., Boerlijst, M. C., Nee, S., Nowak, M. A., and May, R. M. (1996) No signs of hidden language in noncoding DNA. *Phys. Rev. Lett.* **76,** 1977; Voss, R. F. (1996) Comment on "Linguistic features of noncoding DNA sequences." *Phys. Rev. Lett*. **76,** 1978.

38. Bernardi, G., Olofsson, B., Filipski, J., et al. (1985) The mosaic genome of warm-blooded vertebrates. *Science* **228,** 953–958.
39. Churchill, G. A. (1989) Stochastic models for heterogeneous DNA sequences. *Bull. Math. Biol.* **51,** 79–94.
40. Bernardi, G. (2000) Isochores and the evolutionary genomics of vertebrates. *Gene* **241,** 3–17.
41. Oliver, J. L., Bernaola-Gálvan, P., Carpena, P., and Román-Roldán, R. (2001) Isochore chromosome maps of eukaryotic genomes. *Gene* **276,** 47–56.
42. Li, W. (2001) Delineating relative homogeneous G + C domains in DNA sequences. *Gene* **276,** 57–72.
43. Bernardi, G (1989) The isochore organization of the human genome. *Annu. Rev. Genet*. **23,** 637–661.
44. Gates, M. A. (1986) A simple way to look at DNA. *J. Theor. Biol.* **119,** 319–328.
45. Halsey, T. C., Jensen, M. H., Kadanoff, L. P., Procaccia I., and Shraiman, B. I. (1986) Fractal measures and their singularities: the characterization of strange sets}. *Phys. Rev. A* **33,** 1141–1151.
46. Hao, B.-L. (1989) *Elementary Dynamics and Chaos in Dissipative Systems*, World Scientific, Singapore.
47. McCauley, J. L. (1993) *Chaos, Dynamics and Fractals*, Cambridge University Press, Cambridge.
48. Takens, F. (1981) Detecting strange attractors in turbulence. *Lect. Notes Math*. **898,** 366–381.
49. Block, A., von Bloh, W., and Schellnhuber, H. J. (1990) Efficient box-counting determination of generalized fractal dimensions. *Phys. Rev. A* **42,** 1869–1874.
50. Hou, X.-J., Gilmore, R., Mindlin G. B., and Solari, H. G. (1990) An efficient algorithm for fast $O(N + \ln(N))$ box counting. *Phys. Lett. A* **151,** 43–46.
51. Meisel, L. V., Johnson M., and Cote, P. J. (1992) Box-counting multifractal analysis. *Phys. Rev. A* **45,** 6989–6995.
52. Yamaguti, M. and Prado, C. P. C. (1997) A smart covering for a box-counting algorithm. *Phys. Rev. E* **55,** 7726–7732.
53. Oiwa, N. N. and Fiedler-Ferrara, N. (2002) Lyapunov spectrum from time series using moving boxes. *Phys. Rev. E* **65,** 036702/1-10.
54. Press, W. H., Flannery, B. P., Teukolsky S. A., and Vetterling, W. T. (1989) *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, Cambridge.
55. Tél, T., Fülöp A., and Vicsek, T. (1989) Determination of fractal dimensions for geometrical multifractals. *Physica A* **159,** 155–166.
56. Chhabra, A. B., Meneveau, C., Jensen R. V., and Sreenivasan, K. R. (1989) Direct determination of the $f(\alpha)$ singularity spectrum and its application to fully developed turbulence. *Phys. Rev. A* **40,** 5284–5294.
57. Yamaguti, M. and Prado, C. P. C. (1995) A direct calculation of the spectrum of singularities $f(\alpha)$ of multifractals. *Phys. Lett. A* **206,** 318–322.
58. Abarbanel, H. D. I. and Kennel, M. B. (1993) Local false nearest neighbors and dynamical dimensions from observed chaotic data. *Phys. Rev. E* **47,** 3057–3068.
59. Eckmann, J.-P. and Ruelle, D. (1985) Ergodic theory of chaos. *Rev. Mod. Phys.* **57,** 617–656.