Reconstructing phylogeny from the multifractal spectrum of mitochondrial DNA

James A. Glazier,[1,*] Sridhar Raghavachari,[1] Cheryl L. Berthelsen,[2] and Mark H. Skolnick[3]
[1]*Department of Physics, University of Notre Dame, Notre Dame, Indiana 46556*
[2]*Department of Health Information Management, University of Mississippi Medical Center, School of Health Related Professions, 2500 North State Street, Jackson, Mississippi 39216*
[3]*Department of Medical Informatics, Genetic Epidemology, University of Utah, 420 Chipeta Way, Suite 180, Salt Lake City, Utah 84108*
(Received 23 March 1994)

Conventional methods of phylogenetic reconstruction from DNA sequences require simplified models of evolutionary dynamics. We present a method based on fractal analysis to reconstruct the evolutionary history of organisms from mitochondrial DNA sequences. We map animal mtDNA into four-dimensional random walks and estimate their long range correlations using multifractal spectra. We see systematic changes in correlations in mtDNA sequences across taxonomic lines, which translate into changes in the scaling of the random walks. We use cluster analysis to group the multifractal spectra and obtain the phylogeny of the organisms. Though our method uses no *a priori* assumptions and is independent of gene order, it yields phylogenetic relationships broadly consistent with established results. Several recent papers have analyzed DNA using fractal analysis and have found long range correlations. However, no one has succeeded in using them to deduce biologically significant relationships.

PACS number(s): 87.10.+e, 87.22.As

Animal mitochondrial DNA (mtDNA) displays a variety of conserved and distinct features among different species, which provides a useful means of evaluating phylogeny [1]. The maternal inheritance of mtDNA and the absence of recombination imply that much of the evolutionary history of the organism is preserved in its mtDNA sequences [2]. mtDNA sequences show a pattern of species specific codon usage and base composition, indicating the presence of taxon dependent correlations between nucleotides at various length scales. We quantify these correlations by calculating the multifractal spectrum of the sequences and compare the trends across taxonomic lines.

We used 15 animal mtDNA sequences from GenBank with a wide taxonomic span (Table I) to generate pseudo-random walks on a unit step lattice using the embedding scheme of Berthelsen, Glazier, and Skolnick [3]. The axis assignments for the embedding are as follows (where A, C, T, and G are the bases adenine, cytosine, thymine, and guanine, respectively, that make up a DNA sequence):

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 $\{CT\}=\{TC\}=(0,0,0,1)$.

$\{AT\}=\{TA\}=\{CG\}=\{GC\}=(0,0,0,0)$.

The geometric representation of the sequences results in fractal structures which are highly inhomogeneous. Global measures such as correlation exponents have been used to study changes in nucleotide organization [4,5] with evolution, but they give limited information because much of the local structure averages out. We used the sandbox method [6] to calculate the multifractal spectrum of the walks to obtain information about their local organization:

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

where the square brackets indicate an average over a number of randomly sampled points on the walk, p_i is the number of points within a circle of radius R centered around the ith sampled point divided by the total number of points in the walk, and N the total number of points sampled on the walk. D_q is exact in the limit of zero radius. We follow Berthelesen, Glazier, and Skolnick [3] and calculate a least squares fit over the range from one lattice step to the average span of the walk which results in a multifractal spectrum independent of the walk length [7].

The multifractal spectrum of a random walk depends on the embedding scheme used to generate the walk. Different representations emphasize the frequency of different dimers and reveal different sets of correlations. We can maximize information about the sequence by using additional embedding schemes to generate multifrac-

*Corresponding author.

TABLE I. Mitochondrial sequences used in phylogenetic analysis.

Sequence	GenBank Code	Length
Homo sapiens (human)	HUMMTCG	16 569
Bos taurus (cow)	MIBTXX	16 338
Rattus norvegicus (rat)	MIRNXX	16 298
Mus musculus (mouse)	MUSMT	16 295
Phoca vitulina (harbor seal)	MIPVDNA	16 826
Balaenoptera physalus (fin whale)	MIBPCG	16 398
Xenopus laevis (toad)	XELMTCG	17 553
Cyprinus carpio (carp)	MICCCG	16 364
Crossostoma lacustre (fish)	CRQMTGENOM	16 558
Drosophila yakuba (fruit fly)	MIDYRRN	16 017
Apis mellifera (honey bee)	AMFGENOM	16 343
Strongylocentrotus purpuratus (sea urchin)	MISPXX	15 650
Paracentrotus lividus (sea urchin)	PALMTCG	15 696
Caenorhabditis elegans (nematode)	MTCE	13 794
Ascaris suum (nematode)	MTAS	14 284

tal spectra, as long as the resulting spectra are independent of each other. We used the alternate embedding scheme described in Ref. [3], which is as follows,

Axis 1. $\{AA\}=\{CC\}=(-1,0,0,0)$

and $\{TT\}=\{GG\}=(1,0,0,0)$.

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 scheme yields D_0 values for sequences that are uncorrelated with the D_0 values from our first embedding scheme. For a discussion of the significance of these differences, see [3]. The effect of sequence lengths on D_q is discussed in detail in [7].

In Fig. 1 we show the average multifractal spectra using the first embedding scheme of the six major groups. We find significant differences between the multifractal spectra for vertebrates and invertebrates. Generally, D_q decreases with increasing organismal complexity. The nematode spectrum, however, is different from the other invertebrates as well as from the vertebrates. The second embedding scheme produces slightly lower values of D_q but does not change the order of the D_q curves significantly. The mean error in estimating the D_q curves from the goodness of the log-log fit was 5%.

Vertebrate and invertebrate mitochondrial genomes have similar gene content but significant differences exist in the ordering of the genes among species. The variation in gene ordering does not explain the differences in the multifractal spectra, because a rearrangement of invertebrate mtDNA sequences to match the order in vertebrate sequences results in identical multifractal spectra [8]. Such a rearrangement does not remove the long range correlations present in the sequence, which are a result of codon usage, base frequency, and evolutionary pressure. The conserved nature of the mitochondrial genome is apparent from the narrow range of D_q values. We can perform a broad grouping of the organisms into vertebrates and invertebrates based on a visual analysis of Fig. 1. Thus the multifractal spectra allow us to estimate pairwise distances between sequences without any knowledge of evolutionary mechanisms.

To determine phylogenetic associations among the taxa, we need to recognize inherent groups in the data from the distances between sequences. A variety of multivariate statistical techniques can be used to partition data into groups. We used the hierarchical clustering program of Murtagh and Heck [9] which clusters data into groups using a selectable clustering criterion and used the (q,D_q) values as characteristic variables for the mtDNA derived random walks. The (q,D_q) values obtained from both embedding schemes were used as

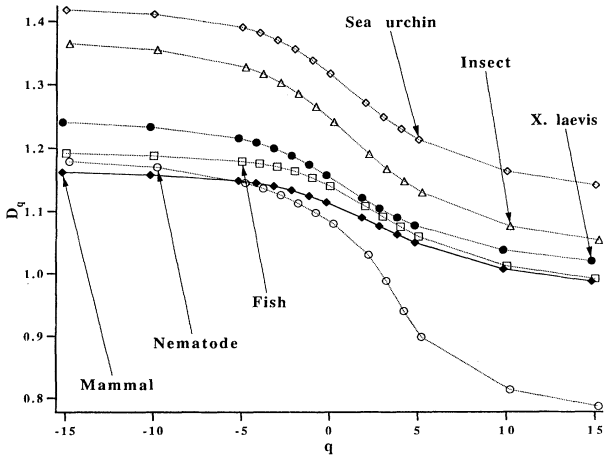

FIG. 1. Averaged D_q curves calculated using the sandbox method for six taxa reveal clear differences in information content and the correlation as a function of length scale of mtDNA sequences.

characteristic variables for clustering to maximize the available information about the sequences. Clustering algorithms treat the objects to be clustered as points in a multidimensional space, with the values of the characteristic variables as the coordinates. They compute the intercluster dissimilarities as Euclidean distances in the multidimensional space and join the two least dissimilar clusters. Errors in locating the coordinates of individual points in the space do not accumulate while computing the distance which makes cluster analysis superior to simple statistical techniques. We applied two commonly used clustering methods; the complete-link method and the minimum-variance method. The complete-link method computes intercluster dissimilarity as the largest dissimilarity among objects in two clusters at each step, i.e.,

$$d(A,B)=\max_{\substack{i\in A\\ j\in B}} d(i,j)\ . \tag{2}$$

The minimum-variance method defines the intercluster dissimilarity as

$$d^2(A,B)=\frac{n_A n_B}{n_A+n_B}|\bar{c}(A)-\bar{c}(B)|^2\ , \tag{3}$$

where n_A and n_B are the number of objects in a cluster, and $\bar{c}$ is the centroid of each cluster defined as

$$\bar{c}=\frac{1}{n}\sum_i \mathbf{i}\ , \tag{4}$$

where the ith object is a vector, the rank of which is the number of criteria used to represent the objects, the components being the values of each criterion for the object, and the sum is over the n objects in the cluster. This technique is similar to least squares minimization. Clusters which minimize this statistic are joined at each step. Both methods yield a single measure representing the dissimilarity between clusters, which indicates the values of the dissimilarities at which lineages diverge.

We show the nearly identical dendrograms resulting from the two methods in Fig. 2. The vertebrates and the invertebrates separate into distinct clusters at an early stage. Both methods clearly delineate the mammalian, amphibian, and piscine clusters. The two dendrograms, however, differ in the ordering of invertebrate divergences. Among the invertebrates, the nematodes, the echinoderms, and the insects cluster among themselves. The mammalian cluster is identical in both dendrograms with somewhat different estimates of the branching distances. Our method is thus fairly robust with respect to the clustering method used.

We compare our result with established lineages obtained using molecular and nonmolecular approaches and find the ordering of the vertebrates consistent with results obtained from earlier analyses of mtDNA [10,11], indicating that our method preserves the monophyletic origins of both the vertebrates and the mammals. Both methods assign the amphibian and piscine lineages to a sister group relationship. The divergence of the invertebrates in both dendrograms differs somewhat from the results of Wolstenholme [10]. The single-link method [Fig. 2(a)] places the nematodes as an outgroup, but places the arthropod divergence later than the echinoderm divergence with respect to the vertebrate invertebrate split. In contrast, the minimum-variance method [Fig. 2(b)] separates the arthropods and the echinoderms as a single unit which later bifurcates to separate the two taxa, indicating a paraphyletic relation between the nematodes and the other invertebrates. Broadly, the two dendrograms are consistent with morphological data. In both the dendrograms, the cow and the fin whale are correctly grouped in a distinct clade, and the carnivore (harbor seal) groups correctly with the primate (human) rather than the cow-whale clade, in contrast to ordering according to molecular data [12] or mtDNA sequences [11] using other methods.

The errors ($\sim 5\%$) in calculating D_q from the slope of the log-log curve (see above) were not included in the clustering because the largest such error was much smaller than the smallest distance between any two clades (vertebrates-invertebrates or mammals-amphibians-fish). This error does not change the order of branching of the lineages. For the worst case error, the order of branching of closely related organisms (e.g., the sea urchins) would switch, but the rest of the tree survives intact [13]. The clustering algorithm itself does not have any associated uncertainty. The dendrograms are not just the most probable ones, but the only ones.

Mitochondrial DNA can resolve evolutionary diver-

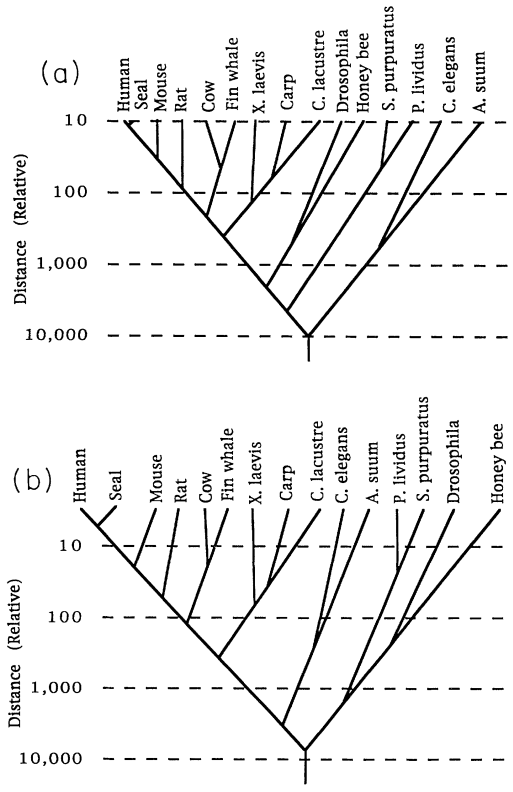

FIG. 2. Dendrogram calculated using (a) complete-link method, (b) minimum-variance method. All values of distance scaled by 10 000 for clarity.

gence of organisms over a wide range of time scales. Fractal analysis is sensitive enough to resolve short time changes such as mammalian evolution (millions of years), in addition to resolving long time evolutionary differences (hundreds of millions of years) between vertebrates and invertebrates. If mtDNA behaved as a molecular clock with a constant temporal rate of base substitutions, we would expect the distance measures given by clustering to correspond to an absolute time of speciation. Such a clock would not explain the apparent divergence of the piscine and amphibian lines after the mammalian lineage has been established. If the rate of evolution of mtDNA depends on the generation number, rather than time, since the generation times of the amphibians and fish are larger than from mammals, we would expect their time scales to be compressed with respect to the mammals, as we observe.

This research was supported by the NSF Grant No. DMR92-57011, Exxon Educational Foundation, the Ford Motor Company, and the American Chemical Society, Petroleum Research Fund.

[1] D. R. Wolstenholme, D. O. Clary, J. L. Macfarlane, J. A. Wahleithner, and L. Wilcox, in *Achievements and Perspectives of Mitochondrial Research,* edited by E. Quagliariello, E. C. Slater, F. Palmieri, C. Saccone, and A. M. Kroon (Elsevier, Amsterdam, 1985), Vol. II, pp. 61–69.

[2] G. A. Watterson and P. Donnely, Genet. Res. Cambridge **60**, 221 (1992).

[3] C. L. Berthelsen, J. A. Glazier, and M. H. Skolnick, Phys. Rev. A **45**, 8902 (1992).

[4] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).

[5] S. V. Buldyrev, A. L. Goldberger, S. Havlin, C-K. Peng, M. H. R. Stanley, and M. Simons, Biophys. J. **65**, 2673 (1993).

[6] T. Tel, A. Fulop, and T. Vicsek, Physica A **59**, 155 (1989).

[7] C. L. Berthelsen, J. A. Glazier, and S. Raghavachari, Phys. Rev. E **49**, 1860 (1994).

[8] C. L. Berthelsen, Ph.D. dissertation, University of Utah, 1992, p. 128.

[9] F. Murtagh and A. Heck, *Multivariate Data Analysis* (Kluwer, Dordrecht, 1987), pp. 55–109.

[10] D.R. Wolstenholme, Int. Rev. Cytol. **141**, 173 (1992).

[11] U. Arnason and E. Johnsson, J. Mol. Evol. **34**, 493 (1992).

[12] W-H. Li, M. Guoy, P. M. Sharp, C. O'Huigin, and Y-W. Yang, Proc. Nat. Acad. Sci. U.S.A. **87**, 6703 (1987).

[13] S. Raghavachari, J. A. Glazier, C. L. Berthelsen, and M. H. Skolnick (unpublished).