Fractal Landscapes and Molecular Evolution: Modeling the Myosin Heavy Chain Gene Family

S. V. Buldyrev,* A. L. Goldberger,‡ S. Havlin,* C-K. Peng,* H. E. Stanley,* M. H. R. Stanley,* and M. Simons‡§
*Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215; ‡Cardiovascular Division, Harvard Medical School, Beth Israel Hospital, Boston, Massachusetts 02215; and §Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

ABSTRACT Mapping nucleotide sequences onto a "DNA walk" produces a novel representation of DNA that can then be studied quantitatively using techniques derived from fractal landscape analysis. We used this method to analyze 11 complete genomic and cDNA myosin heavy chain (MHC) sequences belonging to 8 different species. Our analysis suggests an increase in fractal complexity for MHC genes with evolution with vertebrate > invertebrate > yeast. The increase in complexity is measured by the presence of long-range power-law correlations, which are quantified by the scaling exponent α. We develop a simple iterative model, based on known properties of polymeric sequences, that generates long-range nucleotide correlations from an initially noncorrelated coding region. This new model–as well as the DNA walk analysis–both support the intron-late theory of gene evolution.

INTRODUCTION

The local and global properties of nucleotide organization in DNA sequences have been the focus of considerable attention (Tavaré and Giddings, 1989; Li and Graur, 1991; Beckmann and Trifonov, 1991; Fickett 1982). Short-range correlations with a characteristic length scale, such as tandem repeats, have been intensively studied. More recently, several groups of investigators (Peng et al., 1992; Li and Kaneko, 1992; Voss, 1992; Munson et al., 1992) have reported the unexpected discovery of long-range power-law correlations in certain DNA sequences. This finding is of interest (Maddox, 1992) because it indicates a type of scale-invariant (fractal) organization that may extend over 10^4–10^5 nucleotides (Munson et al., 1992). Furthermore, using "DNA walk analysis" (Peng et al., 1992), we noted that these long-range correlations were present in noncoding (intron and intergenomic) sequences, but not in coding regions.

The biological implications of these long-range correlations for genomic structure and function remain speculative (Peng et al., 1992; Li and Kaneko, 1992; Voss, 1992; Maddox, 1992; Grosberg et al., 1993; Li 1992; Buldyrev et al., 1993a,b). Possible clues to their origin might be obtained by studying the change in long-range correlations with gene evolution.

Molecular evolutionary relationships are usually inferred from comparison of coding sequences, conservation of intron/exon structure of related sequences, analysis of nucleotide substitutions, and construction of phylogenetic trees (Li and Graur, 1991). The changes observed are conventionally interpreted with respect to nucleotide sequence composition (mutations, deletions, substitutions, alternative splicing, transpositions, etc.) rather than overall genomic organization.

Received for publication 10 May 1993 and in final form 30 September 1993.
Address reprint requests to Dr. S. V. Buldyrev, Department of Physics, Center for Polymer Studies, Boston University, 590 Commonwealth Avenue, Boston, MA 02215.

0006-3495/93/12/2673/07 $2.00

In this work, we sought to assess the utility of DNA correlation analysis as a complementary method of studying gene evolution. In particular, we studied the changes in "fractal complexity" of nucleotide organization of a single gene family with evolution. A recent study by Voss (1992) reported that the correlation exponent derived from Fourier analysis was lowest for sequences from organelles, but paradoxically higher for invertebrates than vertebrates. However, this analysis must be interpreted with caution since it was based on pooled data from different gene families rather than from the quantitative examination of any single gene family.

The purpose of the present study is, therefore, threefold:

1. To introduce the technique of DNA walk analysis to the problem of graphically depicting and quantitatively analyzing gene evolution.
2. To test the hypothesis that the fractal complexity of genes from higher animals is greater than that of lower animals, using single gene family analysis. We focused our analysis on the genome sequences from the conventional (Type II) myosin heavy chain (MHC) family. Such a choice limits potential bias that may arise secondary to nonuniform evolutionary pressures and differences in nucleotide content between unrelated genes. We use also this technique to study the MHC gene family because of the availability of completely sequenced genes from a phylogenetically diverse group of organisms, and the fact that their relatively long sequences are well suited to statistical analysis.
3. To present a simple iterative model of gene evolution based on known properties of polymeric sequences that generates long-range correlations.

DNA WALK ANALYSIS

Each nucleotide sequence is represented as a string of purines and pyrimidines. For the present analysis, the "DNA walk" is initiated from the first nucleotide of the coding sequence and continued to the last nucleotide of the last exon. For each pyrimidine at position i, the walker takes a step up $[u(i) = +1]$, and for each purine, a step down $[u(i) = -1]$

(the purine-pyrimidine rule). This procedure generates an irregular graph resembling a "fractal landscape" (Shlesinger, 1993), whose altitude $y(n)$ is the "net displacement" of the walker after n steps:

$$y(n) \equiv \sum_{i=1}^{n} u(i).$$

The defining feature of such a landscape is the statistical self-affinity of the plots obtained at various magnifications. We studied 11 distinct genes and their cDNAs from eight species (Fig. 1).

The landscapes for the MHC gene family (from yeast to human) are presented in Fig. 1. Note that while the ones for the MHC cDNAs are remarkably consistent across the evolutionary spectrum (Fig. 1 *A*), the genomic landscapes show increasing "complexity" (Fig. 1 *B*) evidenced by an apparent increase in the roughness of the surface created by the DNA walk associated with fragmentation of exons by introns.

To quantitatively characterize these landscapes, we partition the entire landscape into boxes of length ℓ and compute the "detrended walk" as the difference between the original walk and the local trend (Peng et al., preprint). The "local trend" of a given box is obtained from a least squares fit for the DNA walk displacement in that box. We then calculate the variance about the local trend for each box, and compute the average of these variances over all the boxes of size ℓ, denoted $F_d^2(\ell)$. The detrended walk analysis gives similar results as the min/max method used in Peng et al. (1992), but has the advantage that it is somewhat better suited to compensate for nonstationarity effects.

The roughness of the landscape is closely related to the correlation properties of the sequence. If the sequence has correlations of only a finite range (or is altogether uncorrelated), then for large ℓ we must find the asymptotic behavior $F_d(\ell) \sim \ell^{1/2}$. On the other hand, if the correlation has "infinite" range (Peng et al., 1992; Li and Kaneko, 1992; Voss, 1992; Munson et al., 1992), i.e., if the range of the correlation is as long as the length of the DNA sequence, then $F_d(\ell) \sim \ell^{\alpha}$ where the exponent $\alpha \neq 1/2$.

The DNA walk provides a visual representation of the statistical properties of DNA, thereby facilitating analysis of its "global" properties (Fig. 1). For example, the landscape produced by DNA walk analysis reveals that each MHC cDNA consists of two roughly equal parts with significant differences in nucleotide content (Fig. 1 *B*). The first part that codes for the heavy meromyosin or "head" of the protein molecule has a slight excess of purines (52% purines and 48% pyrimidines); the second part that codes for the light meromyosin or "tail" has about 63% purines and 37% pyrimidines. The *absolute nucleotide* contents are not shown in the graphical representation of Fig. 1 *A* because we subtract the average slope from the landscape to make relative

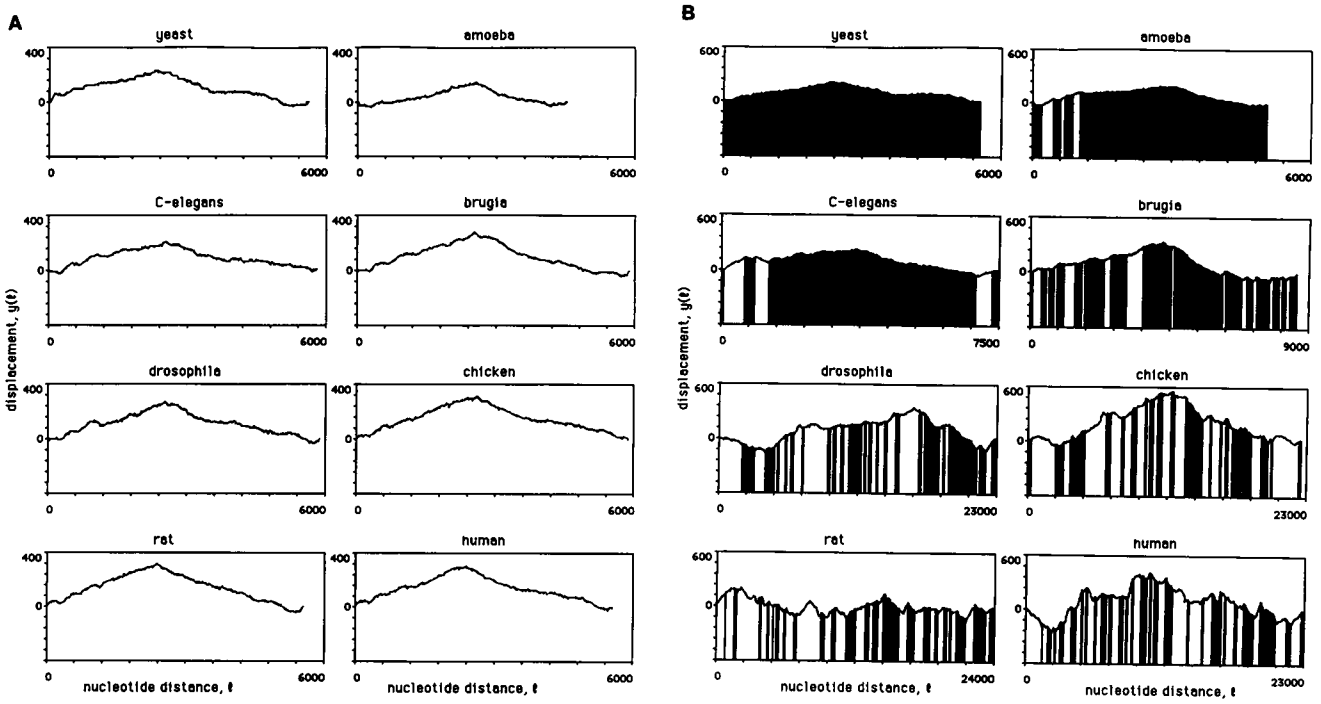

FIGURE 1 The DNA walk representations of (*A*) eight cDNA sequences from the MHC family and (*B*) the corresponding genes. DNA landscapes are plotted so that the end points have the same vertical displacement as the starting points (Peng et al., 1992). The graphs are for yeast, amoeba, *C. elegans*, brugia, drosophila, chicken, rat and human (from top to bottom, left to right). The shaded areas in (*B*) denote coding regions of the genes. The DNA walks for the genes show increasing "complexity" with evolution. In contrast, the cDNA walks all show remarkably similar crossover patterns due to sequential "up-hill" and "down-hill" slopes representing different purine/pyrimidine strand biases in the regions coding for the head and tail of the MHC molecule, respectively.

fluctuations around the average more visible. Indeed, one can easily see from Fig. 1 *A* that the relative concentration of pyrimidines in the first part ("uphill" region) of the myosin cDNA is much higher than in the second ("downhill" region).

As previously reported (Peng et al., 1992) we find that $\alpha \approx 1/2$ for all cDNAs (corresponding to no correlations or only short-range correlations), while all MHC genes containing introns have $\alpha > 1/2$, corresponding to long-range correlations. Data for all eight species are summarized in Table 1. Fig. 2 shows representative scaling plots of $F_d(\ell)$ versus ℓ. Of note, the value of α is not strongly related to the presence of exons since "stitching together" intron sequences (by removing exons) produces a value similar to that of the full gene. For example, for the human MHC gene the value of α after the exons are removed is 0.593 versus 0.586 for the complete sequence, further supporting the view that the composition of noncoding elements is the principal source of long-range correlations in genomic sequences.

INSERTION-DELETION MODEL

To gain some insight into possible evolutionary mechanisms that could increase the complexity of the landscapes and generate long-range correlations of DNA sequences, we next introduce a simple model that simulates conversion of originally coding regions into noncoding introns (cf. Buldyrev et al., 1993). The model is based on the hypotheses that genetic information was originally encoded in an mRNA molecule which was subsequently converted into a DNA sequence, and that this sequence underwent modifications due to mutagenesis and insertion of noncoding genetic material (introns) (Joyce, 1989).

(i) To simulate cDNA sequences, we start with a biased random walk of length L with an overall excess of purines over pyrimidines corresponding to that observed in the cDNA sequences.

(ii) At each time step, we "mutate" the sequence by the following procedure:

(*a*) Choose a random point in the sequence and cut a subsequence of length n starting from that point, where the length n is chosen from a power law distribution $\phi(n) \sim n^{-\beta}$ with $\beta \approx 2$ (between $L_o \approx 20$ and $L/2$). The reason for this power law distribution is that the cutting of a DNA segment most likely occurs when a loop is formed, and it is known that the distribution of loop sizes in a long polymer obeys a

TABLE 1 Long-range correlation analysis for myosin heavy chain gene family

Family Organism	GenBank accession #	Length analyzed (bp) (total length)	# of intron segments	% intron length	$\alpha \pm$ SE
Acanthamoeba castellanii (amoeba)	Y00624	5038 (5894)	3	10	0.51 ± 0.01 [0.51 ± 0.02]
Saccharomyces cerevisiae (yeast)	X53947	5552 (6108)	0	0	0.51 ± 0.01 [0.51 ± 0.02]
Drosophilia melanogaster (fruit fly)	M61229	19391 (22663)	16	66	0.52 ± 0.01 [0.50 ± 0.02]
Caenorhabditis elegans myo #1 (worm)	X08065	7464 (12241)	8	22	0.52 ± 0.01 [0.52 ± 0.02]
Caenorhabditis elegans unc-54 (worm)	J01050	7269 (9000)	8	19	0.53 ± 0.01 [0.53 ± 0.02]
Caenorhabditis elegans myo #3 (worm)	X08067	7332 (11621)	6	19	0.53 ± 0.01 [0.53 ± 0.02]
Brugia malayi (worm)	M74000	8599 (11766)	14	32	0.53 ± 0.01 [0.50 ± 0.02]
Caenorhabditis elegans myo #2 (worm)	X08066	6801 (10780)	11	14	0.54 ± 0.01 [0.52 ± 0.02]
Gallus gallus (chicken)	J02714	20238 (31111)	21	74	0.56 ± 0.01 [0.54 ± 0.02]
Rattus norvegicus (rat)	X04267	22052 (25759)	38	77	0.57 ± 0.01 [0.54 ± 0.02]
Homo sapiens (human)	M57965	20854 (28438)	39	74	0.59 ± 0.01 [0.53 ± 0.02]

Values of the correlation exponent (α) are presented in ascending order. The value of α is calculated by analyzing the GenBank sequence from the first to the last nucleotide of the protein coding sequence including all nucleotides in between. The values of α for the protein coding sequences (without introns), are given in brackets below the α value for gene sequence. The following myosin heavy chain (MHC) genes were studied: MHC type II from *A. castellani*, *S. cerevisiae*, *D. melanogaster*, and *B. malayi*; myo-1, myo-2, myo-3, and unc-54 MHC from *C. elegans*; embryonic skeletal (*G. gallus* and *R. norvegicus*), and beta cardiac (*H. sapiens*). The error bars, reported in the tables are due to the finite size of the analyzed sequences. Indeed, if one generates an artificial completely uncorrelated random sequence of *finite size L* and calculates its correlation exponent α, this value might differ from the expected value $\alpha = 0.5$. The standard deviation of this difference can provide a good estimation for the error bar in the calculated value of α for a real DNA sequence of the same size. As shown in Peng et al. (1993), the standard error of the measured value of α for finite sequences scales like $C(\ell/L)^{1/2}$, where $\ell = 100$ is the range of the linear fit, L is the length of the sequence analyzed, and C is a coefficient which in the case of the detrended analysis technique is approximately 0.1. We used this formula to estimate error bars in Tables 1 and 2.

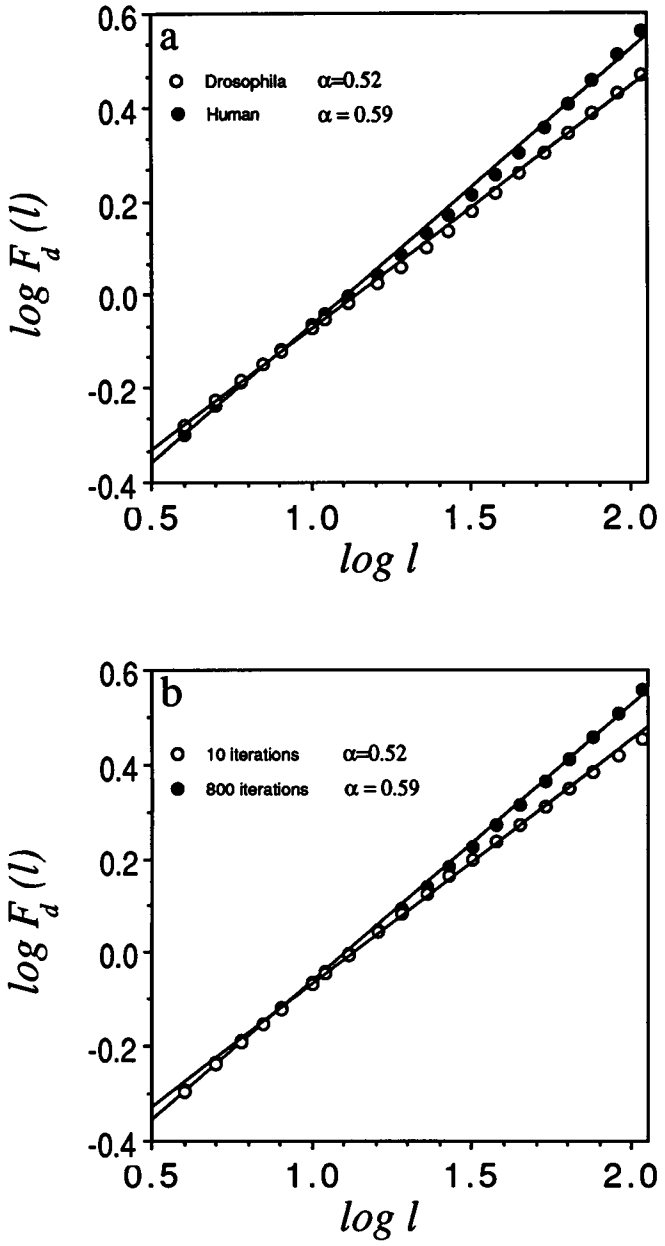

FIGURE 2 Double logarithmic plot $F_d(\ell)$ function versus ℓ for (*A*) full MHC genes (insect and human) and (*B*) nucleotide sequences generated by the model at different stages (number of iterations). The straight lines are linear regression fits from ℓ= 4 to 108.

power law (des Cloizeaux, 1980). Choose another random point in the sequence at which we insert this length, *n* subsequence.

(*b*) With probability 0.5, a *strand substitution* may occur in this subsequence (i.e., all purines are substituted by pyrimidines and vice versa, thereby inserting a complementary strand).

(*c*) To simulate retroviral insertions occurring, with some small probability p_i, the subsequence to be inserted is substituted by a random sequence of equal length with the same percentage of purines and pyrimidines as in the initial cDNA sequence.

Representative purine-pyrimidine landscapes generated by this model are shown in Fig. 3. After the first few iterations (Fig. 3 *A*), the landscape is remarkably similar to that of primitive organisms such as phages (see Fig. 1 of Peng et al., 1992) and *E. coli* which have only a small percentage of introns. The scaling behavior shows $F_d(\ell) \sim \ell^{1/2}$ as anticipated. After more iterations, the landscape seems visually more complex and the α measured for the first two decades increases, reminiscent of the increase noted in the MHC family from yeast to invertebrates to vertebrates. After roughly 1000 iterations, the value of α asymptotically approaches 0.60 as shown in Fig. 2 *B*. The landscapes for the model and for the rat MHC gene become quite similar (cf. Figs. 3 *B* and 3 *C*).

Of note, if the model is iterated without the insertion of random biased sequences as assumed in rule (iic), the value of α will return to 0.5, indicating a random sequence. Insertion of biased random regions (according to a power-law distribution) maintains the exponent $\alpha > 0.5$. The importance of rule (iic) of the model is consistent with the hypothesized role of retroviral insertions in the genomes of high animals (Hwu et al., 1986).

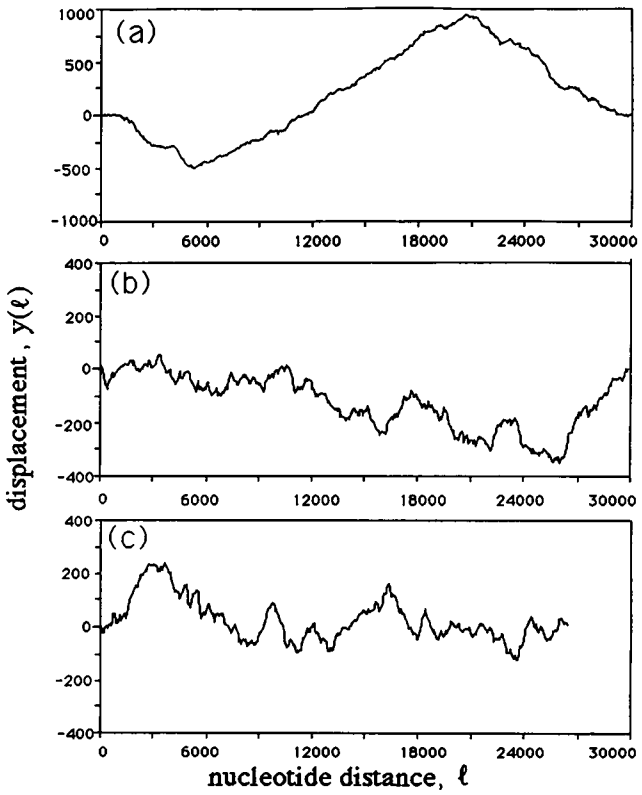

FIGURE 3 DNA walk representations of artificial sequences generated by the stochastic model described in the text. The parameters used in this simulation are: L = 30,000; p_i = 0.2; L_o = 20; and 62% purines in the initial sequence. (*A*) The early stage of "evolution" (after 100 iterations in the model simulation) shows a landscape with two to three large regions of different bias (the up-hill and down-hill regions). (*B*) After 1600 iterations, the landscape becomes visually more complex, resembling the actual DNA walk representation for the rat MHC sequence in (*C*). The values of p_i, L_o, and L given here are typical of those we used in our simulation.

Furthermore, without strand substitution as implemented by rule (iib), no long-range correlation will appear. This mirror-image replacement mimics molecular evolution occurring by partial gene duplication or transposition (Schleif, 1988) and the occurrence of "extinguished exons" (Jaworski et al., 1989). In order to test our assumption of strand substitution we also analyzed an alternative DNA landscape in which nucleotides cytosine (C) and guanine (G) result in an up step, while adenine (A) and thymine (T) correspond to a down step. Since such walks cannot be affected by strand substitution, our model would predict the absence of long-range correlations. Indeed, our analysis of the fluctuation $F_d(\ell)$ for this modified DNA landscape does not exhibit as robust a power law correlation as for the original purine-pyrimidine rule. Another crucial assumption is the existence of an overall bias (either of purines or of pyrimidines) in the initial sequence; it is this bias that enables strand substitution to produce differences in nucleotide content. This assumption is consistent with our observation that most coding regions exhibit overall bias in their purine-pyrimidine concentration.

The mechanism of generating power-law correlations in this insertion-deletion model is related to the competition between two countervailing "forces." The deletion and insertion of segments in rule (iia) and (iib) tends to *randomize* the sequence, while the insertion of biased segment implemented by rule (iic) tends to *organize* the system. As the iteration proceeds, the newly inserted biased segment is then broken into smaller pieces of different bias (according to a power-law distribution). After a large (but finite) number of iterations (which depends on the parameters of the model), these two competing effects will tend to balance each other. At this point the system will exhibit power-law correlation.

DISCUSSION

The existence of long-range correlations in DNA sequences was first demonstrated to occur in noncoding elements of eukaryotic genes (Peng et al., 1992; Li and Kaneko, 1992), and later confirmed by Voss (1992) and then extended to an entire chromosome (Munson et al., 1992). Despite the extensive demonstrations of their existence, little is known regarding the biological significance of these correlations. Such power-law behavior reflects a *scale-invariant* property of DNA and, therefore, cannot be attributed simply to the occurrence of nucleotide periodicities such as those associated with nucleosome packaging (Beckmann and Trifonov, 1991). Possibly these long-range correlations are related to higher order DNA/chromatin structure, to DNA bending or looping, or to HnRNA splicing. In this regard, it is interesting to note that long-range correlations are present in the intergenomic and intron but not exon (protein coding) sequences. This demonstration of a self-affine (fractal) structure of these elements points to an important although still undefined biological role.

We have presented here evidence suggesting that the fractal complexity of the MHC gene family increases with evolutionary order. In particular, we find vertebrate genes appear to be more complex than those of invertebrates or yeast, as quantified by the long-range correlation exponent α. In contrast to the landscapes of the full genes, the MHC cDNAs show remarkable preservation of organization (invariance) across the evolutionary spectrum. Indeed, the eight cDNAs in Fig. 1*A* are virtually indistinguishable by visual inspection and all have a value of $\alpha \approx 0.5$ indicating no long-range correlations for the coding sequences.

The use of α orders MHC genes family in the same order as that obtained by conventional sequence analysis Katsuragowa et al., 1989). Furthermore, we observed higher α for unc-54 MHC than for the rest of *Dict.* myosin, suggesting later evolution of this sequence. This is consistent with analysis in Katsuragowa et al. (1989) that also suggest late origin of unc-54 myosin.

We have focused on the MHC family because of the extensive nature of data available from a variety of invertebrate and vertebrate species, as well as the substantial length of the individual sequences (Peng et al., 1993). Ideally, one would want to study precisely similar genes from different species. For the present analysis we had to include closely related but not identical myosin genes (e.g., human beta cardiac myosin and rat and chicken embryonic skeletal myosin) because of the current limitations in the GenBank sequence records.

We also employed similar methodology to study three additional gene families: actin, cytochrome c, and lysozyme (Table 2). In all three cases, increase in the value of α roughly follows accepted evolutionary order of species development, consistent with the trends in the MHC family. Overall, for the gene families presented in Tables 1 and 2, bacteria, yeast, and plants have the lowest values of α, ($n = 8$; $\bar{\alpha} \pm \text{SE} = 0.504 \pm 0.014$) followed by invertebrates ($n = 10$; 0.525 ± 0.006) and then vertebrates ($n = 19$; 0.574 ± 0.006).

An important question is whether the observed increase in the value of α with evolution is simply a reflection of the increasing percentage of introns in higher species. We note that while the percentage of introns remains relatively constant for the most advanced three species studied in the MHC family (Table 1), the value of α is not constant. This finding, in accord with the insertion-deletion model, suggests that the organization of intron sequences, not only their length, is important in the process of evolution.

The observed trend of α to increase with evolutionary status for the MHC family is also consistent with the predictions of the model: "higher" species that appeared more recently will tend to generate long-range correlations with a larger value of the parameter α. Thus, vertebrate myosin is likely to be more "complex" than invertebrate myosin because the former incorporated genetic material from the latter species. This view of molecular evolution is consistent with the theory of punctuated equilibrium (Eldredge and Gould, 1972) that postulates rather rapid periods of change (occurring during speciation) followed by periods of stasis.

Our finding that α increases with evolution contradicts a recent study by Voss (1992) which paradoxically reported that the strength of nucleotide correlations (quantified by the

TABLE 2 Long-range correlations for cytochrome C, lysozyme, and actin gene families

Family organism	GenBank accession #	Length Analyzed (bp) (total length)	# of intron segments	% intron length	$\alpha \pm$ SE
Cytochrome C					
Bradyrhizobium japonicum (bacterium)	M77796	449 (703)	0	0	0.43 ± 0.05
Emericella nidulans (fungus)	M83141	446 (1310)	2	26	0.49 ± 0.05
Desulfovibrio vulgaris (bacterium)	M63807	1637 (2160)	0	0	0.5 ± 0.02
Oryza sativa (rice)	M63704	1903 (2408)	2	82	0.52 ± 0.02
Arabidopsis thaliana (plant)	X59459	475 (1263)	2	29	0.54 ± 0.05
Saccharomyces cerevisiae (yeast)	J01320	341 (869)	0	0	0.54 ± 0.03
Drosophilia melanogaster (fruit fly)	M11381	326 (1015)	0	0	0.56 ± 0.06
Gallus gallus (chicken)	K02303	490 (1620)	1	36	0.6 ± 0.05
Gallus gallus (chicken)	X00230 K02303	481 (1616)	1	35	0.6 ± 0.05
Mus musculus (mouse)	X01756	418 (1436)	1	25	0.63 ± 0.05
Homo sapiens (human)	M22877	418 (3088)	1	24	0.64 ± 0.05
Lysozyme					
Drosophilia melanogaster (fruit fly)	X58383	425 (1366)	0	0	0.49 ± 0.05
Hyalophora cecropia (silk moth)	M60914	2731 (3392)	2	84.7	0.51 ± 0.02
Rattus norvegicus (rat)	L12458	6594 (7951)	3	93.3	0.54 ± 0.02
Homo sapiens (human)	X14008	4746 (6807)	3	91.8	0.55 ± 0.01
Bos taurus B (cow 2)	M95098	5104 (10212)	3	91.4	0.56 ± 0.01
Bos taurus C (cow 3)	M95099	6489 (8051)	3	93.2	0.57 ± 0.01
Bos taurus A (cow 1)	M95097	7289 (12222)	3	94	0.57 ± 0.01
Gallus gallus (chicken)	X61001	2315 (4803)	4	92.7	0.57 ± 0.02
Actin					
Dictyostelium discoideum (slime mould A12 actin)	X03282	1127 (1131)	0	0	0.53 ± 0.03
Rattus norvegicus (rat, alpha-skeletal)	J00692	1723 (3200)	5	34.5	0.54 ± 0.02
Homo sapiens (human, beta cytoplasmic)	M10277	1907 (3646)	4	41	0.56 ± 0.02
Grassa idella (grass carp, beta cytoplasmic)	M25013	1720 (4243)	4	35	0.57 ± 0.02
Gallus gallus (chicken, type 5 cytoplasmic)	M10279	1130 (2620)	0	0	0.57 ± 0.03
Gallus gallus (chicken, alpha skeletal)	V01507	1855 (2426)	5	39	0.58 ± 0.02
Rattus norvegicus (rat, beta cytoplasmic)	V01217	1887 (4100)	4	41	0.60 ± 0.02

Correlation exponents α for genomic sequences of cytochrome C, lysozyme, and actin gene families are presented in the order of ascending α values. Sequence analysis was carried out in the same way as for the MHC gene family. However, the values of α for the coding sequences from these families are not calculated since these sequences are too short to provide sufficient accuracy.

power spectral scaling exponent β, which is uniquely related to α) increases from organelle to invertebrates but then *decreases* for primates. This apparent discrepancy is likely due to the facts that Voss (1992) (i) did not analyze *single* gene families with evolution, (ii) did not distinguish intron-containing versus intron-less sequences, and (iii) did not correct for large regions of "strand bias" (unequal numbers of purines and pyrimidines) as we do in the present article using detrended fluctuation analysis. We have found that if one does not take into account the *crossover* between two large (but uncorrelated) regions of strand bias as seen in all the MHC cDNAs (corresponding to the uphill and downhill regions in Fig. 1 *A*), one can obtain a spuriously large value of α.

Nee (1992) proposed that it is the alternation of introns and exons (regions containing different nucleotide content) which modulates the long-range correlations. This idea is somewhat similar to the proposed model, but its main conclusion–that the sequence from which all exons have been cut does not exhibit long-range correlations–appears to be incorrect. In fact, intron sequences show long-range correlations as robust as those of complete genes with approximately the same exponent α. In contrast, our model describes not only the intron-insertion process, but also the shuffling process within noncoding sequences (introns and intergenomic sequences). This shuffling process (not just the insertion of uncorrelated introns) leads to $\alpha > 0.5$ within single introns and intergenomic sequences, a fact that cannot be explained in the framework of the Nee hypothesis. Further, our model bears potential relevance to biological evolution, by providing a possible mechanism for transformation of primordial RNA molecules (currently considered to be the first to develop) into complex DNA sequences containing noncoding elements.

Finally, two major theories have been advanced to explain the origin and evolution of introns. One suggests that precursor genes consisted entirely of coding sequences and introns were inserted later in the course of evolution to help facilitate development of new structures in response to selective pressure, perhaps, by means of "exon shuffling" (Gilbert, 1978). The alternative theory suggests that precursor genes were highly segmented and subsequently organisms not requiring extensive adaptation or new development or, perhaps, facing the high energetic costs of replicating unnecessary sequences, lost their introns (Hagerman, 1990; Doolittle, 1990). Support for these hypotheses has remained largely conjectural; no models have been brought forward to support either process. The landscape analysis of the MHC gene family and the stochastic model presented in this study are most consistent with the former view.

We wish to thank F. Sciortino for important contributions in the initial stages of this project, and C. Cantor, C. DeLisi, J. M. Hausdorff, R. D. Rosenberg, J. J. Schwartz, M. Schwartz, and N. Shworak for valuable discussions. Partial support was provided to A. L. G. by the G. Harold and Leila Y. Mathers Charitable Foundation, the National Heart, Lung and Blood Institute, and the National Aeronautics and Space Administration, to M. S. by the American Heart Association to C.-K. P. by the National Institute of Mental Health, and to S. V. B., M. H. R. S., and H. E. S. by the National Science Foundation.

REFERENCES

Beckmann, J. S., and E. N. Trifonov. 1991. Splice junctions follow a 250-base ladder. *Proc. Natl. Acad. Sci. USA*. 88:2380–2390.

Buldyrev, S. V., A. L. Goldberger, S. Havlin, C-K. Peng, M. Simons, and H. E. Stanley. 1993a. Generalized Lévy walk model for DNA nucleotide sequences. *Phys. Rev. E.* 47:4514.

Buldyrev, S. V., A. L. Goldberger, S. Havlin, C-K. Peng, M. Simons, F. Sciortino, and H. E. Stanley. 1993b. Long-range power-law correlations in DNA. *Phys. Rev. Lett.* 71:1776.

Des Cloizeaux, J. 1980. Short range correlation between elements of a long polymer in a good solvent. *J. Physique (Paris)*. 41:223–238.

de Gennes, P. G. 1979. Scaling Concepts in Polymer Physics. Cornell University Press, Ithaca, NY.

Doolittle, W. F. 1990. Understanding introns: Origins and functions. *In* Intervening Sequences in Evolution and Development. E. Stone and R. Schwartz, editors. Oxford University Press, New York. 42–62.

Eldredge, N., and S. J. Gould. 1972. Punctuated equilibria: An alternative to phyletic gradualism. *In* Models in Paleobiology. T. J. M. Schopf, editor. Freeman and Cooper, Inc., San Francisco. 82–115.

Fickett, J. W. 1982. Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res.* 10:5303–5318.

Gilbert, W. 1978. Why genes in pieces? *Nature (Lond.)*. 271:501–501.

Grosberg, A. Yu., Y. Rabin, S. Havlin, and A. Nir. 1993. Self-similarity in DNA structure. *Biofisika (Russia)*. 26:1–6; *Europhys. Lett.* In press.

Hagerman, P. 1990. Sequence-directed curvature of DNA. *Annu. Rev. Biochem.* 59:755–781.

Hwu, R. H., J. W. Roberts, E. H. Davidson, and R. J. Britten. 1986. Insertion and/or deletion of many repeated DNA sequences in human and higher ape evolution. *Proc. Natl. Acad. Sci. USA*. 83:3875–3879.

Jaworski, C. J., and J. Piatigorsky. 1989. A pseudo-exon in the functional human A-crystallin gene. *Nature (Lond.)*. 337:752–760.

Joyce, G. F. 1989. RNA evolution and the origins of life. *Nature (Lond.)*. 338:217–224.

Katsuragowa, Y., M. Yanagisawa, A. Inoue, and T. Masaici. 1989. Two distinct nonmuscle myosin heavy chain mRNAs are differentially expressed in various chicken tissues. *Eur. J. Biochem.* 184:611-616.

Li, W-H., and D. Graur. 1991. Fundamentals of molecular evolution. Sinauer Associates, Sunderland, MA.

Li, W., and K. Kaneko 1992. Long-range correlations and partial $1/f^{\alpha}$ spectrum in a noncoding DNA sequence. *Europhys. Lett.* 17:655–658.

Maddox, J. 1992. Long range correlations within DNA. *Nature (Lond.)*. 358:103–103.

Munson, P. J., R. C. Taylor, and G. S. Michaels. 1992. Long range DNA correlations extend over entire chromosome. *Nature (Lond.)*. 360:636–636.

Nee, S. 1992. Uncorrelated DNA walks. *Nature (Lond.)*. 357:450–450.

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 (Lond.)*. 356:168–170.

Peng, C-K., S. V. Buldyrev, A. L. Goldberger, S. Havlin, M. Simons, and H. E. Stanley. 1993. Finite size effects on long-range correlations: implications for analyzing DNA sequences. *Phys. Rev. E.* 47:3730–3733.

Schleif, R. 1988. DNA-looping perspective. *Science (Washington DC)*. 240: 127–135.

Shlesinger, M. F. 1993. Random Walks. World Scientific, Singapore.

Tavaré, S., and B. W. Giddings. 1989. *In* Mathematical Methods for DNA Sequences. M. S. Waterman, editor. CRC Press, Boca Raton, FL. 117–132 and references therein.

Voss, R. 1992. Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences. *Phys. Rev. Lett.* 68:3805–3808.