Physics Letters A 175 (1993) 252–256
North-Holland

PHYSICS LETTERS A

Scale-invariant spatial patterns in genome organization

Michael D. Purugganan
Botany Department, University of Georgia, Athens, GA 30602, USA

Received 24 August 1992; accepted for publication 10 February 1993
Communicated by A.P. Fordy

Eukaryotic genomes are characterized by coding regions interspersed within non-coding sequences, creating irregular dispersal patterns. Fractal analysis was applied to the study of this dispersed organization of eukaryotic genomes. The results show that eukaryotic genomes possess two length regimes – a short ordered length scale, and a fractal regime with fractal dimensions ranging from 0.21 ± 0.02 to 0.84 ± 0.02. Fractal scaling provides clues to the origin and evolution of sequence patterns within genomes, and provides us with tools necessary to characterize such patterns in detail.

1. Introduction

Spatial power-law correlation (fractal) analyses have proven useful in investigations on the nature of the processes that shape irregular structures. In biological systems, the organization of ramified physiological structures [1], protein backbones and surfaces [2,3] and RNA secondary structure [4] appears to be governed by power laws. Recently, there has been interest in the study of long-range correlations in DNA nucleotides sequence data [5–7]. These studies have demonstrated the existence of compositional correlations in DNA sequences approximated by the power law

$$F(l) \propto l^{\alpha} . \qquad (1)$$

Interestingly, these long-range sequence correlations appear prominently in the non-coding intron sequences of genes.

Compositional correlations are only one of several possible sequence correlations in organismal DNA; of interest, particularly to biologists, are spatial correlations of functional coding regions within genomes. Only a small portion of the eukaryotic genome encodes functional coding sequences [8,9]. Several studies estimate that only one-tenth to one-third of genomes are necessary for organismal viability [9–11]. Coding sequences are interspersed with chromosomal non-coding sequences that comprise that bulk of the genome, leading some workers to characterize genomes as consisting of "islands of transcribed sequences in a sea of silent DNA" [12].

The majority of non-coding DNA within eukaryotic genomes are found as introns which separate exons (coding sequences), or intergenic spacers between genes or gene clusters. Although non-coding sequences may contain regulatory signals, most are believed to be essentially non-functional [9]. Selective association of coding sequences are evident in many gene clusters, a result of functional or developmental constraints on gene expression. The spatial proximity of metabolically related genes in small eukaryotic genomes [13,14] and the β-globin cluster [15] are just a few examples. Most genes, however, are randomly dispersed within the genome. There is growing evidence from both experimental [16] and theoretical work [12,17,18] that stochastic sequence rearrangements may be responsible for the dispersal patterns of coding sequences. Scale-invariant properties of genome organization should exist if non-equilibrium dynamic processes are responsible for coding sequence dispersal.

The underlying distribution of coding sequence information within eukaryotic genomes may be derived using either correlation [7] or box-counting methods [19]. In the former, a function $g(x)$ is defined, which equals p_1 if the basepair at position x is

within a coding sequence and p_2 otherwise. The correlation function is

$$c(r) = \langle g(x)g(x+r) \rangle \propto r^{-(1-D)}, \quad (2)$$

for r within the fractal regime [20]. The exponent D is the Hausdorff (fractal) dimension. The value of D can also be determined using box-counting methods. A genomic sequence is divided into sequence blocks of length l and counting the number of blocks, $n(l)$, which contain coding sequences. By this method, $n(l) \propto l^D$.

2. Methods

A spatial analysis of genes, gene clusters and genomes was undertaken to investigate the properties of eukaryotic genome organization. Several eukaryotic genes or gene clusters were analysed, although prokaryotic, viral and organellar systems were included for comparison. The sequences were collected from the literature or Genbank; the extent and position of coding sequences in these genes were already previously determined. A box-counting algorithm was utilized in the analysis, and the logarithmic plot of $n(l)$ versus l generated for each sequence.

3. Coding sequence distribution in genomes show fractal properties

Representative logarithmic plots of $n(l)$ versus l for several genomic systems analysed are shown in fig. 1, and the results are summarized in table 1. Prokaryotic systems show a linear plot with one value of D characterizing the organizational pattern for the entire length scale studied. The fractal dimensions for prokaryotic genes and viral genomes are very near unity – the lowest value obtained is 0.97 ± 0.01 for the *E. coli lac* operon. This reflects the compact nature of these genomes, with little excess DNA to separate coding sequences. Evolutionary streamlining of these genomes to minimize the energetic burden or, for viral genomes, to allow for efficient viral packaging accounts for the paucity of non-coding sequence DNA in these systems.

The fractal plots for eukaryotic genomes, unlike the prokaryotic systems, display a pronounced con-

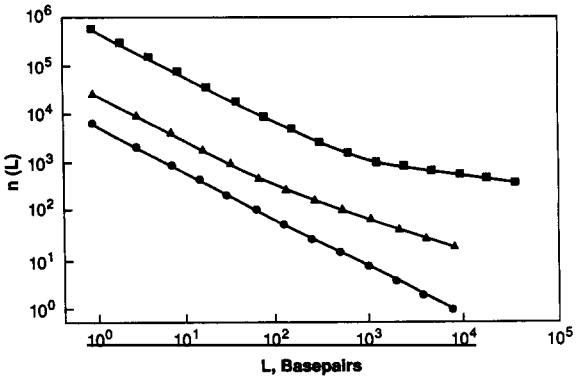

Fig. 1. Representative logarithmic plots of $n(l)$ versus l for several genes. Curve (●) is for the *E. coli trp* operon, (▲) for the human *apoAI* gene cluster, and (■) for the human estrogen receptor gene. Curves (▲) and (■) are vertically displaced one and two orders of magnitude, respectively.

vexity. A transition region separates two length regimes – a highly ordered domain at short length scales and a fractal regime at larger length scales. The short length scales, from 1 to around 19–74 basepairs, show D values of between 0.93 ± 0.01 to 0.99 ± 0.01, corresponding to the construction of exons from nucleotides. Beyond this is a fractal regime where coding sequence organization displays structure over a wide range of length scales. Calculated dimensions within the fractal regime, which spans two decades, vary widely from system to system. It is 0.21 ± 0.2 for the highly dispersed human 5S rRNA tandem gene cluster and 0.84 ± 0.02 for the highly compact *N. crassa qa* gene cluster. This range of values reflects the diverse organizational patterns of various gene systems and genomes.

The cross-over region between the two length regimes in eukaryotic systems is a useful indicator of system properties [21]. In several cases, transitions in scaling behaviour reflect fundamental differences in the underlying growth patterns that give rise to irregular structures. The presence of such transitions in eukaryotic genomes is vividly illustrated if one plots the fractal dimension against the length scale, l, as shown in fig. 2. The cross-over region for our systems ranges from 19 to 296 bps, which corresponds to the characteristic length scale for exons [22,23].

Table 1
Summary of results (n.a.: not applicable).

Gene/gene cluster/genome		Size (bp)	Total coding sequence size (bp)	D_1 [a]	D_2 [b]
viral genomes	lambda phage [c]	48502	40263	0.98	n.a.
	T7 phage [c]	39936	36751	0.99	n.a.
E. coli genes	*malb* operon [c]	6545	5599	0.98	n.a.
	trp operon [c]	7335	6562	0.98	n.a.
	lac operon [c]	7477	6024	0.97	n.a.
organelle genomes	human mitochondria [c]	16569	15361	0.99	n.a.
	mouse mitochondria [c]	16265	15352	0.99	n.a.
	tobacco chloroplast [27]	155844	95248	0.99	0.86
eukaryotic genes	*N. crassa qa* cluster [14]	17260	10420	0.99	0.84
	X. laevis globin cluster [15]	56615	2831	0.97	0.49
	mouse *hprt* gene [28]	33770	1318	0.97	0.31
	human *hprt* gene [28]	40477	1385	0.95	0.41
	human 5S rRNA cluster [29]	30608	1770	0.94	0.21
	human *apoA* cluster [30]	15709	2870	0.97	0.68
	human *apoB* cluster [30]	43108	14149	0.98	0.70
	human *apoCI* cluster [30]	26932	1495	0.95	0.42
	human Factor VII gene [31]	186128	9563	0.98	0.46
	human estrogen receptor gene [32]	144822	6269	0.98	0.29
	chicken progesterone receptor gene [33]	37964	4671	0.98	0.51
	chicken ovalbumin gene [34]	33933	6372	0.98	0.73
	chicken crystallin gene [35]	20120	3669	0.93	0.63
	artificial genome [12] [d]	63560	5040	0.92	0.58

[a] Dimension for ordered regime (short length scales for two-regime systems), with maximum error of ± 0.01. Error estimation indicates the least-squares analysis of best fit to a straight line for $\log n(l)$ versus $\log l$.
[b] Dimension for fractal regime, with maximum error of ± 0.03.
[c] Sequence from Genbank. [d] Each unit taken as ten basepairs.

4. Evolutionary comparisons of genome organization

The fractal dimension is high (>0.8) for gene clusters or genomes whose coding sequences comprise greater than 60% of the total sequence. Wide variation in fractal dimensions for those systems with less than 60% coding sequence information reflects the topological properties of their organizational patterns, which is dependent on the evolutionary forces that create and maintain them. The degree of clustering is one factor that is reflected in the fractal dimension – higher dimensions are associated with pronounced clustering of structural elements. Interestingly, the fractal properties of animal mitochondrial DNA are reminiscent of prokaryotic systems, while chloroplast genome more closely resemble eukaryotic systems.

One should note two other interesting results of these analyses. First, homologous genes in different species have dissimilar organizational patterns as a result of their divergent evolutionary histories. This is reflected in their fractal dimensions, which may assume values characteristic of the species – examples are the human and mouse *hprt* gene, which respectively have D values of 0.41 ± 0.02 and 0.31 ± 0.02. Furthermore, various regions of a species' genome have different fractal properties, pointing to the organizationally heterogenous nature of a genome. This implies that the extent of localized evolutionary structuring within genomes may differ,

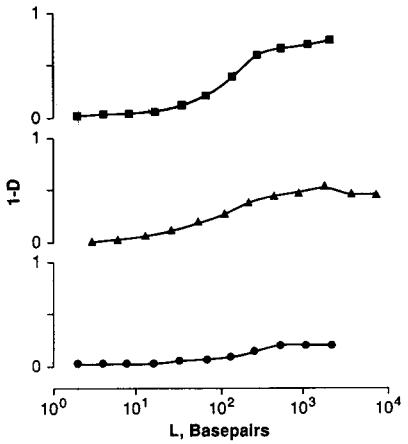

Fig. 2. The transition between the two length regimes is clearly displayed in these plots of $1-D$ versus l. Curve (●) is for the N. crassa *qa* gene cluster, (▲) for the *X. laevis* globin gene cluster, and (■) for the mouse *hprt* gene.

possibly reflecting differences in the susceptibility of various chromosomal regions to certain types of sequence rearrangements.

Given the results of these analyses, one can discern the nature of the dominant evolutionary forces that shape eukaryotic genome organization at short and long length scales. The exons serve as the fundamental structural element of genomes, providing ordered structure at short length scales. At larger length scales, integrity may be maintained even with spatial separation of coding sequences, and exons may be evolutionarily dispersed, clustered or shuffled. This is due in part to the evolution of splicing mechanisms to join exons in the mature mRNA transcript. Moreover, except for selectively clustered genes, dispersal along the chromosome may proceed without any significant adverse effects [9,12]. Stochastic processes such as duplications, deletions and insertions [24], as well as unequal crossing-over between repeat sequence families [25], may dominate in shaping genome organization at these larger length scales.

Such sequence rearrangements have been important, for example, in the organizational evolution of such genes as the β-globin cluster [15,26] and the myosin heavy-chain gene [26]. Theoretical studies have demonstrated that large quantities of non-coding sequences can evolve stochastically to disperse genes [12,17,18]. In one simulation study [12], random duplications and deletions were sufficient to generate vestigial sequences, and the resulting artificial genome has fractal properties similar to that of real eukaryotic genomes (see table 1). The relative importance of various molecular evolutionary mechanisms in generating dispersal patterns requires further exploration, and fractal analysis may allow investigators to address this issue more closely by providing mathematical tools that will allow detailed dissection of simulation results.

References

[1] B. West and A. Goldberger, Am. Sci. 75 (1984) 354.
[2] M. Lewis and D.C. Rees, Science 230 (1985) 1163.
[3] G. Wagner, J. Colvin, J. Allen and H. Stapleton, J. Am. Chem. Soc. 107 (1985) 5584.
[4] M.D. Purugganan, Naturwissenschaften 76 (1989) 471.
[5] W. Li and K. Kaneko, Europhys. Lett. 17 (1992) 655.
[6] C. Peng, S. Buldyrev, A. Goldberger, S. Havlin, F. Sciortino, M. Simons and H. Stanley, Nature 356 (1992) 168.
[7] R. Voss, Phys. Rev. Lett. 68 (1992) 3805.
[8] L. Orgel and F.H. Crick, Nature 284 (1980) 604.
[9] W.F. Loomis, Comput. Biochem. Physiol. 95B (1990) 21.
[10] H. Naora, K. Miyahara and R. Curnow, Proc. Natl. Acad. Sci. USA 84 (1987) 6195.
[11] M. Goebl and T. Petes, Cell 46 (1986) 983.
[12] W.F. Loomis and M.E. Gilpin, Proc. Natl. Acad. Sci. USA 83 (1986) 2143.
[13] N.H. Giles, Am. Natur. 112 (1978) 641.
[14] J. Greener et al., J. Mol. Biol. 207 (1989) 15.
[15] H. Hosback, T. Wyler and R. Weber, Cell 32 (1983) 45.
[16] N.V. Fedoroff, Cell 16 (1979) 697.
[17] T. Ohta and M. Kimura, Proc. Natl. Acad. Sci. USA 78 (1979) 1129.
[18] T. Ohta, in: Oxford surveys in evolutionary biology, eds. P. Harvey and L. Partridge (Oxford Univ. Press, Oxford, 1982).
[19] B. Mandelbrot, Fractal geometry of nature (Freeman, New York, 1982).
[20] H. Takayasu, Fractals in the physical sciences (Manchester Univ. Press, Manchester, 1990).
[21] T. Freltoft, J. Kjems and S. Sinha, Phys. Rev. B 33 (1986) 269.
[22] S.K. Holland and C. Blake, in: Intervening sequences in evolution and development, eds. L. Stone and R. Schwartz (Oxford Univ. Press, Oxford, 1990).
[23] J. Hawkins, Nuc. Acids Res. 16 (1988) 9893.
[24] M. Nei, Molecular evolutionary genetics (Columbia Univ. Press, New York, 1987).
[25] G. Dover and R. Flavell, Genome evolution (Academic Press, New York, 1982).

[26] M. Edgell et al., in: Evolution of genes and proteins, eds. M. Nei and R. Koehn (Sinauer, Sunderland, 1983).
[27] E. Strehler, V. Mahdavi, M. Periasamy and B. Nadal-Ginard, J. Biol. Chem. 260 (1985) 468.
[28] K. Shinozaki, EMBO J. 5 (1988) 2043.
[29] D. Melton, in: Oxford surveys of eukaryotic genes, ed. N. Maclean (Oxford Univ. Press, Oxford, 1987).
[30] R. Little and D. Braaten, Genomics 4 (1989) 376.
[31] A. Frossad, Nuc. Acids Res. 14 (1986) 8699.
[32] B. Blackhart et al., J. Biol. Chem. 261 (1986) 15364.
[33] J. Scott, in: Oxford surveys of eukaryotic genes, ed. N. Maclean (Oxford Univ. Press, Oxford, 1987).
[34] J. Gitschier et al., Nature 312 (1984) 326.
[35] S. Green et al., Nature 320 (1986) 134.