Graphical analysis of DNA sequence structure: III. Indications of evolutionary distinctions and characteristics of introns and exons

A. Nandy

Computer Division, Indian Institute of Chemical Biology, 4 Raja S. C. Mullick Road, Calcutta 700 032, India

A technique for graphical representation of a gene sequence on a 2D plot has been shown by us earlier to be able to provide considerable information on new global sequence patterns and homologies, repeated structures, relative base abundances, probable evolutionary paths and evolutionary divergences. Here we report that at a more micro level the graphical representation reveals differences in the features of intron and exon segments of eukaryotic sequences. Detailed analyses of these differences show that base composition and distribution of exons have tended to become more homogeneous with evolution, whereas introns show no significant change for the classes of genes tested by this procedure.

PROTEIN-CODING gene sequences of eukaryotic organisms are generally fragmented into several coding and non-coding segments characterized as exons and introns[1,2]. Base compositions of exons and introns are known to be different[3] and thus form the basis of several computer programs that attempt to identify the coding regions within a stretch of DNA segment[4–10].

The origins of introns and exons, however, remain elusive with two main schools of thought proposing contrasting intron-early and intron-late models. Proponents of the first model[11–14] hypothesize that the primordial genes from which functional protein templates arose were themselves split and present-day exons and introns are the evolutionary products of the ancient coding regions and their spacers; introns are hypothesized to have been discarded in evolution of bacterial genomes as their transcription processes became streamlined. The alternative hypothesis[15] states that introns were late developments that arose out of insertion events into functional protein templates and which were never present in the ancestors of those organisms that now lack them. Evidence for either point of view is expected to be found in genes and proteins of ancient origin[16,17], but controversies remain. Support for the intron-early view has been claimed from a putative ancient correspondence between exons and units of protein structure and introns, e.g. in corresponding positions of globin genes[12], but a recent study of the presumed correspondence between the intron–exon positions and protein structures in four such genes reported negative results[18].

In this context it may be appropriate to examine how the introns and exons may have evolved over time to learn, if possible, about their earliest states. One line of investigation arising out of a suggestion of Gilbert that exon shuffling could lead to insertions of additional copies of exons in existing gene sequences, has found it to be indeed a favoured process in the development of complex eukaryotic genes. This was interpreted to be a vindication of the intron-early model for building up functional protein templates, but has since been considered as a pathway available to both models.

Taking up a different point of view, we have examined the base distribution characteristics of exons and introns of several ancient and late gene sequences such as kinetoplasts, globins, myosin heavy chain genes, etc. by using a graphical representation technique[19–21] which we have developed extensively[20–23]. We report here our findings that there is a fair amount of evidence for gradual change with evolution in exon base composition and distribution, but that the corresponding parameters for introns have remained fairly constant and in fact retain characteristics comparable to the earliest coding segments in bacteria and some eukaryotic genes.

Method

The designing of experiments was based on the observations[3,22] that intron and exon regions have significantly different base composition and distribution, with exons tending to have a fairly homogeneous base composition whereas the intron regions tend to have a high preponderance of A, T or G, C bases[22]. These show up as differences in their maps in a two-dimensional graphical representation[20] where we construct a symmetric purine–pyrimidine Cartesian co-ordinate system, with purines on the x-axis (A in the negative direction, G positive) and pyrimidines on the y-axis (T in the negative direction, C positive) and plot the sequence structure as a succession of points, one for each occurrence of each base. This will give rise to a map of the gene

sequence, whose structure then in terms of base distribution will be reflected from the progression of the points along the map. The details of such a representation and the rich diversity of information obtainable therefrom have been discussed in detail elsewhere[20–23]. We merely reiterate that in such a representation, conserved sequences such as the globin genes show uniquely recognizable patterns which provide a base for global homology search in an intuitively simple way and that the systematic differences observed in the patterns of these genes appear to be directly related to evolutionary divergence[20]. Further, the graphs generated in this representation provide direct visual identification of regions of different relative abundances of the various bases, and presence of long stretches of repeated structures[20–22]. Also, differences in base composition intrinsic to intron and exon sequences lead to marked differences in their graphical representations[22,23]. In this paper we investigate this aspect in greater detail.

In this representation, the intron segments generally show up as comparatively more open, filament-like structures than the exons, which latter tend to form more closed curves leading to formation of dense clusters of points, especially in the higher eukaryotic gene sequences[22]. This is readily apparent when individual intron–exon regions are plotted, as for example around the region of exon no. 21 of the chicken myosin heavy chain gene (Figure 1 *a*). Likewise, differences between introns and exons were observable also with varying degrees of complexity in other gene sequences: from the apparently similar maps in the heat shock proteins through e.g. the tubulin genes to the strong clustering in the globins (Figure 1).

While individual exon maps can form quite dense clusters, as for example in the case of the third exon of the globin gene (Figure 1 *b*), when the exons are viewed together in tandem, they may form more open curves depending on the overall base composition as seen in the case of the myosin heavy chain (MHC) genes[22]: small deviations from complete homogeneity in individual exons may get amplified as the exons are arranged in tandem. This, however, does not detract from the general observation that many of the individual exons tend to exhibit a more compact cluster formation on the graphical representation, while the introns most often map into long, filamentous structures. The clustering tendency in the exon maps is found to be more pronounced, as shown later, in some genes such as the myosin heavy chains and the globins, and less so in some others such as the tubulins, the heat shock proteins, etc. from which a tentative hypothesis can be made that the exons of the later genes, generally having longer introns, tend to form denser clusters. It seems that the relatively more ancient bacterial genes such as the intronless bacteriophage lambda have a rather open representation (for both the G/C and A/T rich halves) (Figure 1 *e*).

To get a quantitative measure of the cluster formation

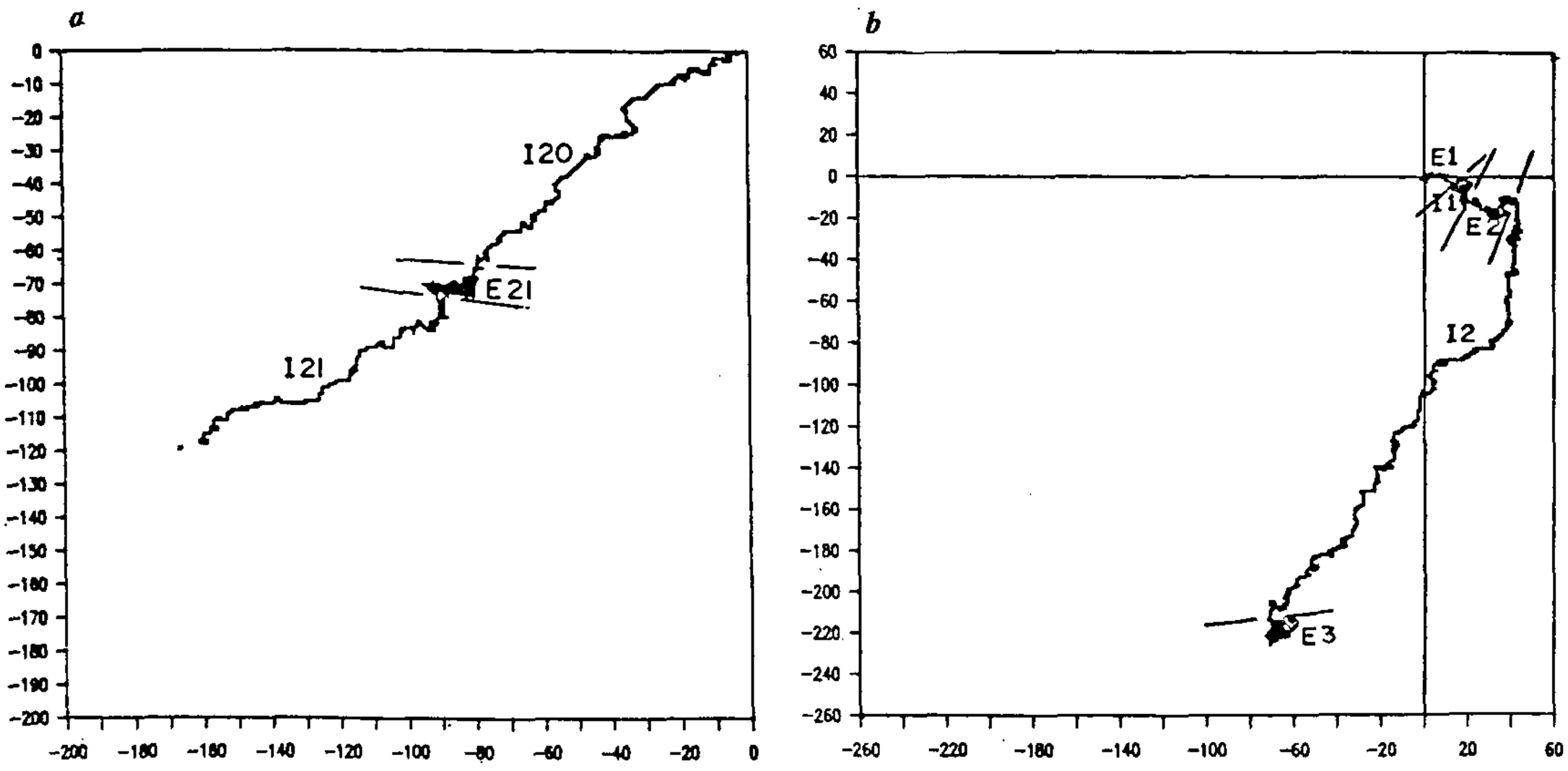


Figure 1 *a–b*. Maps of intron (I) and exon (E) segments from selected sequences plotted on an ACGT-axes system[20]. *a*, Chicken embryonic skeletal myosin heavy chain gene sequence (EMBL AC: M12086, ID: GGMYHE) from base number 13500 to 14800 comprising part of intron number 20 (514 bp), exon 21 complete (256 bp), beginning of intron 21 (531 bp). *b*, Human beta globin from chromosome 11: E1–92, I1–130, E2–223, I2–850, E3–129 bp.

as exhibited in our graphical representation, we have defined three different techniques. In the first, to get a measure of the density of clustering of an individual segment, the part of the gene map comprising the individual exon or intron is enclosed in a rectangle with sides parallel to the x- and y-axes and extent defined by the minimum and maximum co-ordinates of the segment on the two axes. The number of points mapped within the rectangle divided by the area provides a measure of the density of points within this sequence segment. In the case of the exon segments where the four bases are represented in almost equal strengths, the map is generally crowded in a small area and thus will have high density of points whereas intron segment

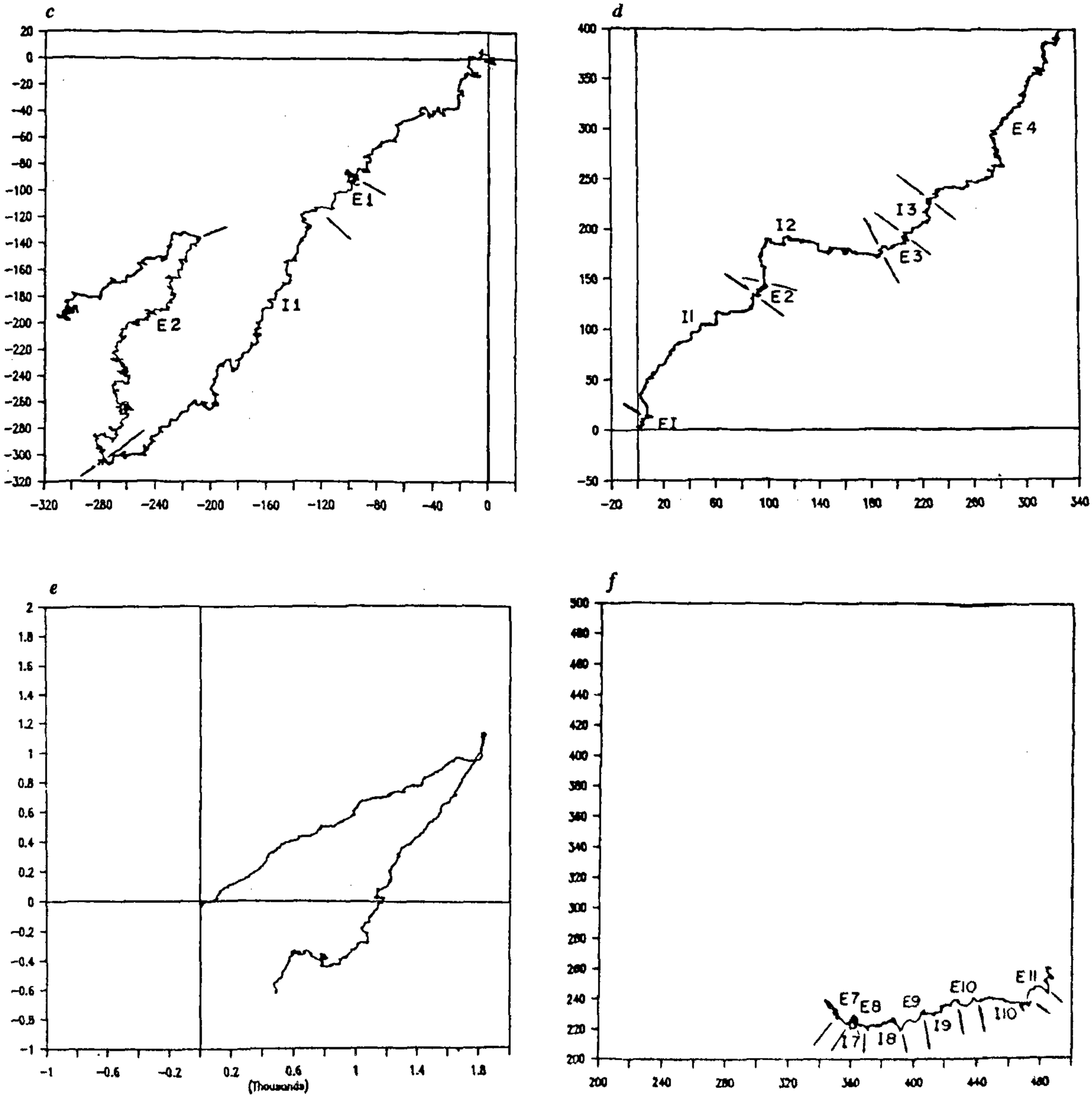


Figure 1 *c–f.* Maps of intron (I) and exon (E) segments from selected sequences plotted on an ACGT-axes system[40]. *c*, *D. melanogaster* heat shock protein 82 (EMBL AC: X03810) complete sequence including flanks (total 5024 bp comprising: e1–149, I1–1129, E2–2151 bp. *d*, Chicken beta-4-tubulin gene (EMBL ID: GGTUB4B), with E1–57, I1–404, E2–109, I2–513, E3 111, I3–139, E4–1073 bp. *e*, Bacteriophage lambda genome (48502 bp). *f*, Human alpha collagen type 1 gene, (EMBL AC: M20789) with exons 7 to 11 of sizes 54, 45, 54, 45, 54 bp alternating with introns of sizes 89, 116, 114, 178, respectively, and flanked by parts of introns 6 (153 bp) and 11 (94 bp).

maps being long and filamentous, will show low densities. However, a possible source of confusion may arise in those cases where a filament plot runs parallel to the x- or y-axis giving a rectangle with very short length or width and thus perhaps leading to high densities; however, such parallel runs require a continuous series of one base or another predominating significantly. While this is known to occur in short stretches, they seldom comprise a whole intron or exon; in general the intron segments have substantially large share of A/T or G/C bases and therefore will map into filament-like structures running more or less diagonally in our plot rather than parallel to the axes. This is more to the point since we are considering complete introns and exons in genome length sequences in this analysis and therefore can expect that the pathological cases mentioned earlier would be rather rare.

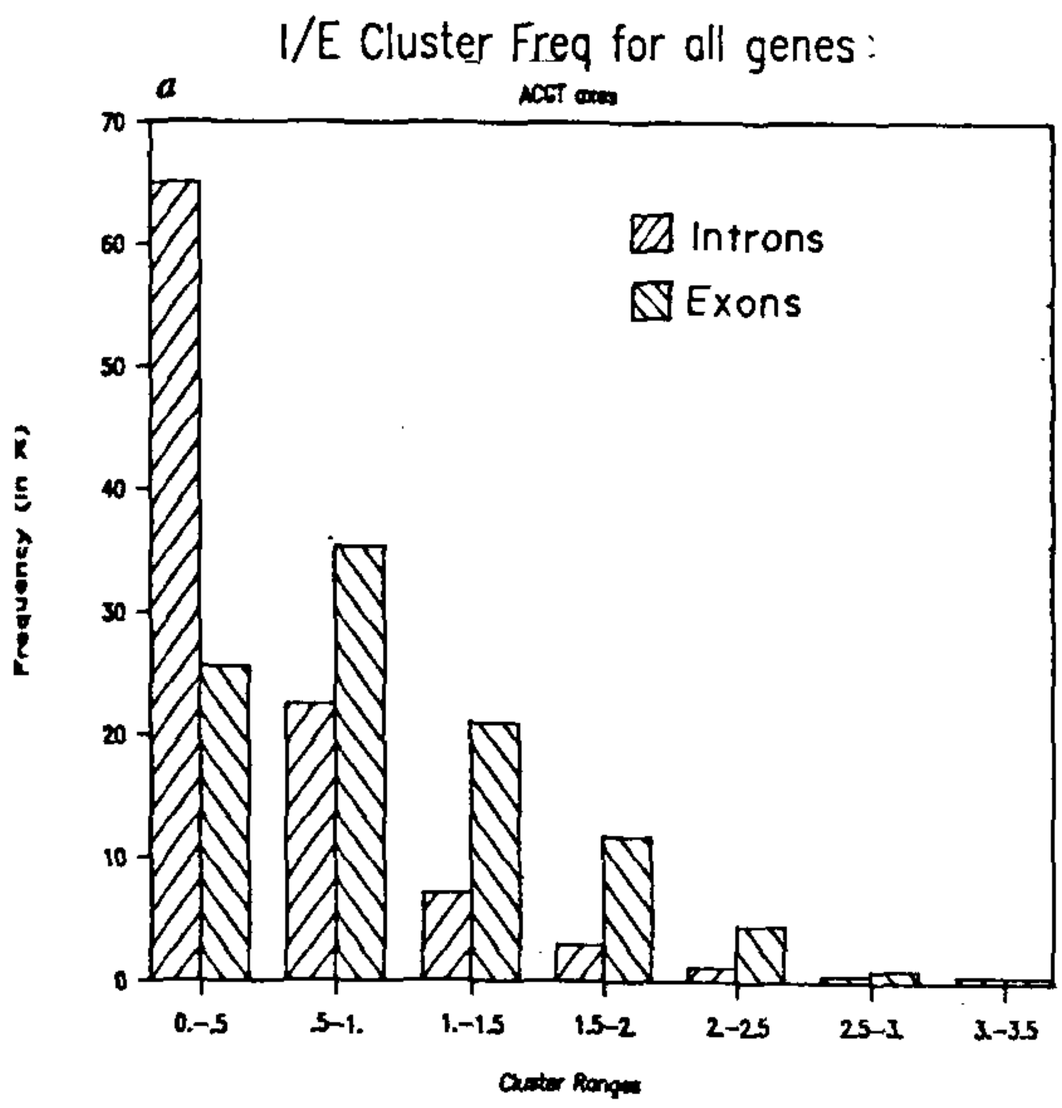
The second method is based on the observation that the compositional differences between exon and intron sequences lead, in this graphical representation, to intron maps generally having long runs in one direction at a stretch thus covering considerable lengths on this plot, whereas in the case of the exons there are frequent changes of directions leading to overall much smaller displacements for equivalent base lengths (see, e.g. the maps in Figure 1). We have shown elsewhere[23] that this feature provides a convenient handle for discriminating between intron and exon segments in intron-rich sequences. In the case of the problem at hand, we follow a slightly different tack and take the average displacement or its inverse as a measure of the difference between intron and exon segments. Thus, we take as the starting point the first point of the kth segment, measure the displacements of the subsequent points till the end of the segment, and then take the average displacement $\bar{d}_k$ as

$$\bar{d}_k = \frac{1}{n_k} \sum_i [(x_i - x_0)^2 + (y_i - y_0)^2]^{1/2},$$

where n_k is the total number of bases of the kth segment, (x_0, y_0) represents the co-ordinates of the starting point and i runs from 1 to n_k. Considering the inverse of the average displacement, the higher this number the smaller the displacement and, therefore, the denser the clustering. This method can lead to aberrant results in the case of very short introns or very long exons, but since generally introns are much longer than exons, especially in higher eukaryotic sequences, the method can be considered to be applicable in the majority of the cases.

One can also define a fractal coefficient alpha for the clusters by considering the maximum radial extent of a cluster in relation to the constituent bases of the segment as follows:

$$\text{alpha} = \frac{\log(\text{no of bases in segment})}{\log(\text{maximum extent of the segment})}$$



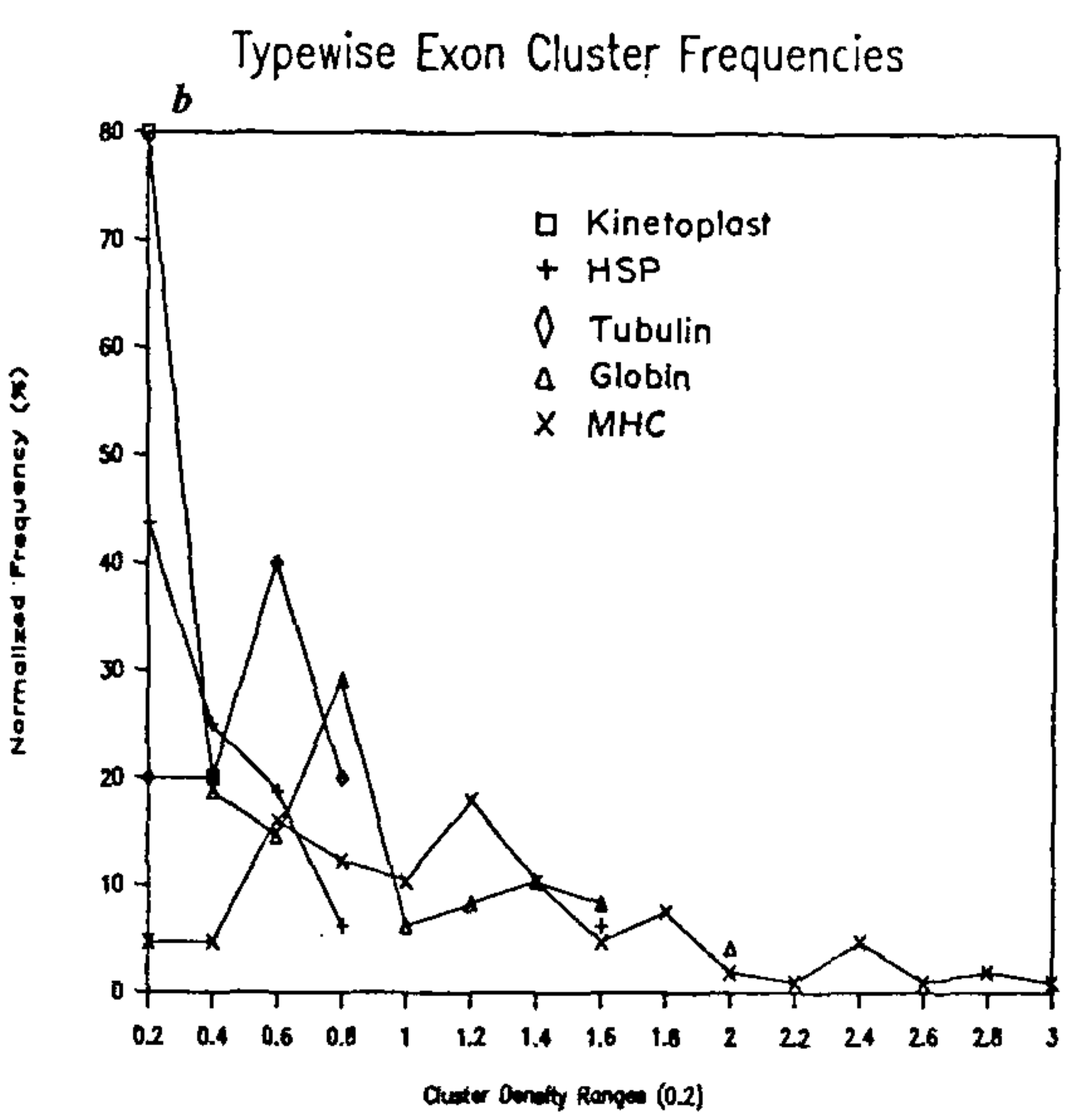


Figure 2 *a*, *b*. *a*, Intron, exon normalized frequency vs cluster density range for all genes in sample. *b*, Exon normalized frequency distribution vs cluster density range for different types of genes.

By this measure, the more tightly packed the bases are in the graphical representation, the larger will be the fractal coefficient. However, as in the previous two quantification techniques, here also erroneous results may arise in some cases, e.g. a variety of different base distributions of essentially similar maximum radial extent but different spreads will lead to essentially same values of the coefficient. However, if we apply all three

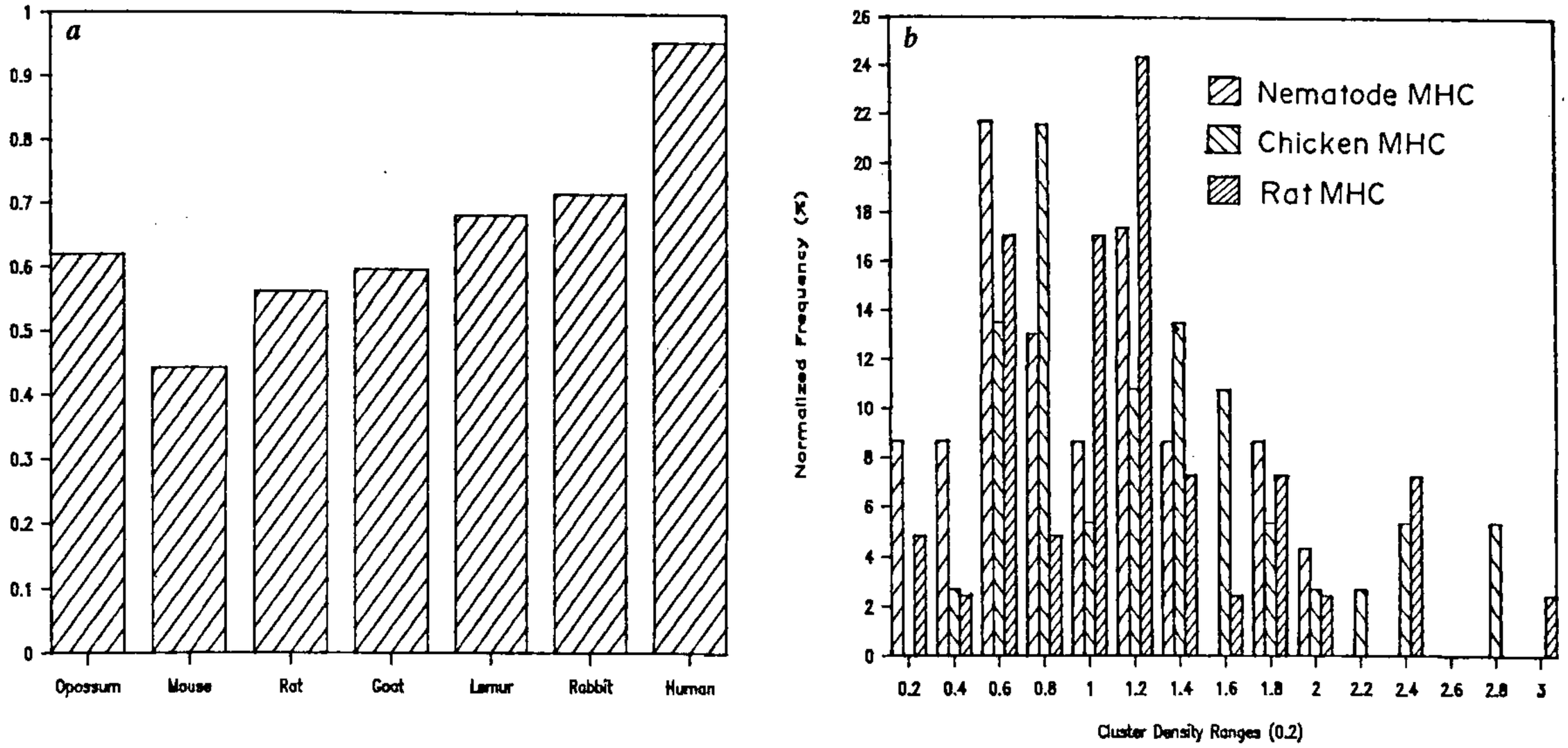


Figure 3 *a, b*. *a*, Cluster densities for third exon of beta globin for different species showing evolutionary changes. The arrangement corresponds to the phylogenetic sequences determined through various models for molecular evolution[26–28] using mitochondrial and globin sequences. *b*, Frequency distribution of cluster densities of exons for three species of myosin heavy chain genes: *C. elegans* MHCs 1, 2, 3; chicken embryonic skeletal MHC and rat embryonic skeletal MHC.

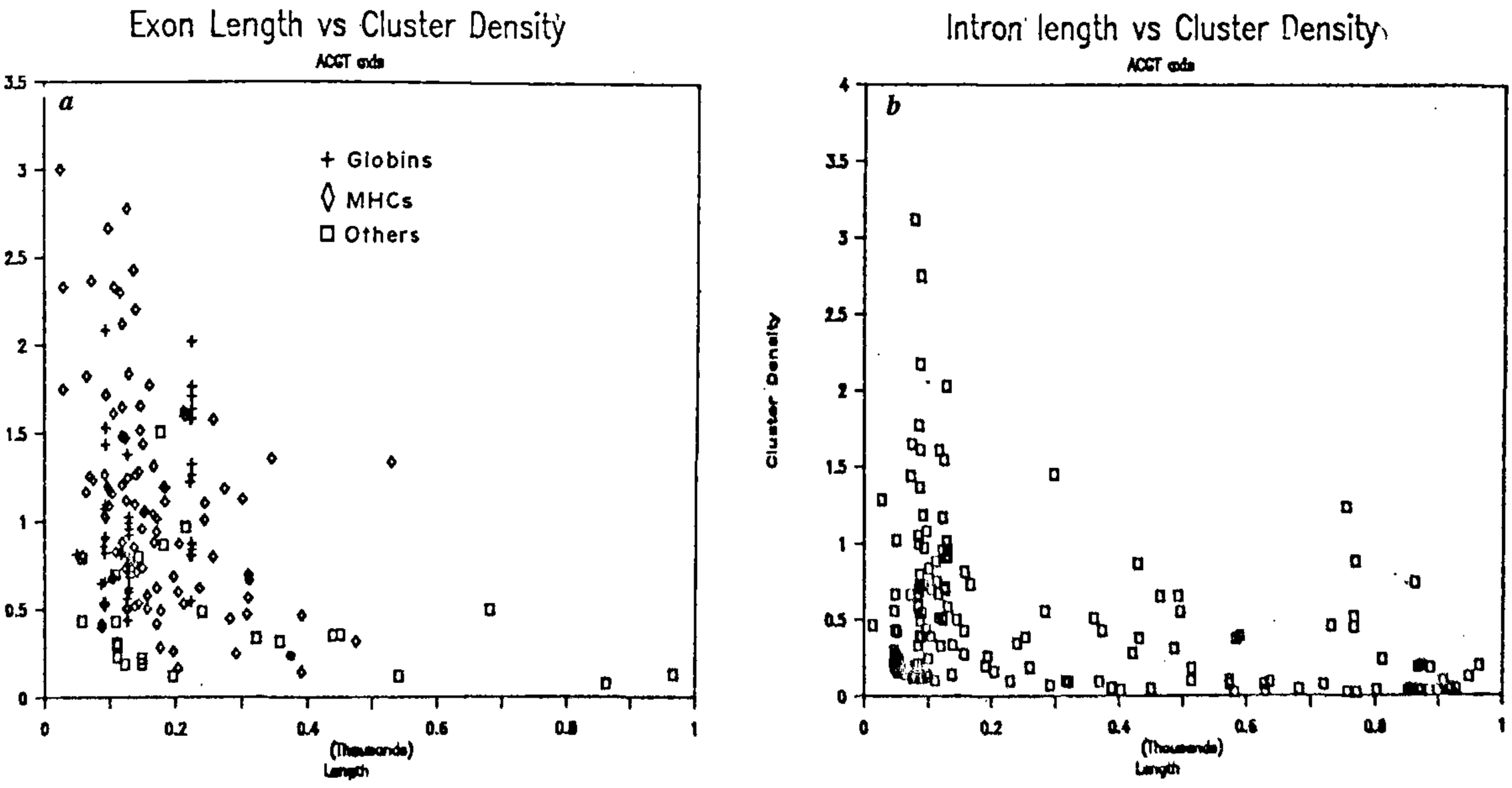


Figure 4 *a, b*. *a*, Scattergram of exon segment lengths (up to 1 kbp) vs cluster densities for different genes. *b*, Scattergram of intron segment lengths (up to 1 kbp) vs cluster densities for all genes.

measures simultaneously, the chances of generating erroneous results by all three techniques in any individual case would be minimal and conclusions derived from them can be considered sufficiently reliable.

Results and discussions

We have studied cluster analysis with various nematodes and vertebrate gene sequences that code for a wide variety of proteins such as globins, tubulins, myosin heavy chains, kinetoplasts, etc. Using a sample of 35 genes with 386 introns and exons covering 262,146 bases excluding all flanking regions, we find from a frequency plot that intron representations predominantly form clusters of very low densities, and the frequency of occurrence falls off exponentially rapidly with cluster density (Figure 2 *a*); the exons maps, on the other hand, grow in clustering density to around 0.8 per unit area and then fall off gradually. The exon sample comprising exon lengths from 9 bases (human collagen) to 6351 bases (*D. discoideum* MHC) have a mean length of 340 bases which is in conformity with the observation of Senapathy[24] that exons generally tend to form small segments. The introns in our sample set have an average spread of 1119 bases ranging from a 13-base intron for *M. fasciculata* to those of several kilobases length; our analysis of cluster densities also includes the intra-globin segments from the human beta globin gene cluster on chromosome 11 extending to over ten thousand bases and the sensitivity of the results to this group is also briefly given in Table 2.

Tables 1 and 2 list the normalized frequencies for clustering effects on the intron and exon plots as

Table 1. Exons: Normalized frequency table of exon clusters from various gene types

Range	Kineto-plast	Heat shock protein	Tubulin	Globin	Myosin	Collagen (human alpha)
Cluster density method						
0–0.2	80	43.75	20		4.72	
0.4	20	25.00	20	18.75	4.72	20
0.6		18.75	40	14.58	16.04	25
0.8		6.25	20	29.17	12.26	5
1				6.25	10.38	10
1.2				8.33	17.92	
1.4				10.42	10.38	5
1.6		6.25		8.33	4.72	20
1.8					7.55	5
2				4.17	1.89	
2.2					0.94	
2.4					4.72	5
2.6					0.94	
2.8					1.89	
2.8–3.0					0.94	5
	100	100	100	100	100	100
Inverse displacement method						
0.1	100	100	70	30.30	32.41	30
0.2			20	63.64	41.67	40
0.3			10		12.04	25
0.4					9.26	5
0.5				6.06	1.85	
0.6					0.93	
0.7					0.93	
0.8					0.93	
	100	100	100	100	100	100
Fractal coefficient method						
1.2	80					
1.4	20	33.33	20		3.04	50
1.6		44.45	50	36.11	25.49	15
1.8		22.22	30	36.11	49.02	25
2				25.00	18.62	10
2.2				2.78	3.92	
	100	100	100	100	100	100

Frequency ranges are at intervals of 0.2 units each for cluster density and fractal coefficient calculations, and at interval of 0.1 unit each for the inverse displacement method.

Table 2. Introns: Normalized frequency table of intron clusters from various gene types

Range	Kineto-plast	Heat shock protein	Tubulin	Globin	Myosin	Collagen (human alpha)
Cluster density method						
0–0.2	50	87.50	66.67	55.56*	30.77	
0.4	25	12.50	11.11	2.78*	26.92	36.84
0.6	25		11.11	5.56	13.46	31.58
0.8				11.11	10.58	10.53
1				11.11	5.77	5.26
1.2				8.33	3.85	
1.4			11,11		1.92	5.26
1.6					1.92	10.53
1.8				2.78	2.88	
2						
2.2				2.78	0.96	
2.4						
2.6						
2.8						
3						
3–3.2					0.96	
	100	100	100	100	100	100
Inverse displacement method						
0.1	50	100	88.89	58.33	66.10	84.22
0.2				25.00	27.97	5.26
0.3	50			16.67	4.24	5.26
0.4			11.11		1.69	5.26
0.5						
0.6						
0.7						
0.8						
	100	100	100	100	100	100
Fractal coefficient method						
1.2	25		33.33		2.04	
1.4	50	100	44,45	46.43	33.68	10.53
1.6	25		11.11	25.00	36.74	63.16
1.8			11.11	21.43	19.39	15.79
2				7.14	7.14	10.52
2.2					1.01	
	100	100	100	100	100	100

*The data for the globin introns includes intra-globin segments from the human globin cluster on chromosome 11; if these are deleted from the analysis, the contribution to the first two intervals changes to 50% and 5.88% with minor changes in the others.

measured by all the three techniques mentioned above. It is interesting to note that the clustering tendencies in exon plots, as shown by the calculated data, increase significantly from heat shock proteins to tubulin genes, globins and the myosins and collagens (Table 1): the peak frequencies shift to higher indices (i.e. cluster density, inverse displacement or fractal coefficient), as also the index range required to cover say 70% of the exons. In contrast, the majority of intron plots have clustering parameters mostly at the lowest range within each gene type (Table 2); here also, there is a progressive increase in the index range covered from one gene type to another, but considerably less so than for the exons. We note that these conclusions are uniform, within their applicable index ranges, irrespective of the technique used to parametrize the clustering propensity, and we, therefore, restrict our discussions in the rest of the paper to the cluster density method of parametrizing as being the most intuitively simple of the three methods.

Table 3. Cluster density analysis for phage M13 and lambda

Cl density range	Phage M13	Lambda
0.2	33.4	20.7
0.4	33.3	17.2
0.6	11.1	20.7
0.8	11.1	3.4
1.0	11.1	6.9
1.2		13.8
1.4		6.9
1.6		3.4
1.8		0.0
2.0		3.4
2.2		
2.4		
2.6		3.4
	100	100

In the light of the above observations on the clustering differences in the plots of the exons and introns, it is of interest to consider the plots of intronless genes like the phage M13 genome or the bacteriophage lambda (Figure 1 *e*). We find that 60% of the genes of these two sequences are covered by cluster density indices (referred to as cluster densities henceforth) of 0.6 and below (Table 3), thus closely paralleling the normal intron frequencies.

Since increasing density in our representation indicates greater homogeneity in base composition and distribution, the above tables imply that there is a systematic change in these features with gene type. Considering the sample genes mapped in Figure 1, comparison of equal-length exons (Table 4) shows that the standard deviation of base composition from the mean of 25% shows a small decrease from heat shock protein to tubulin to the beta globin, thus implying increasing homogeneity. In the cluster density consideration, heat shock proteins, which can be considered amongst the earliest genes, are almost totally weighted towards the low density ranges (Figure 2 *b*). The myosin heavy chain genes, responsible for muscle fibre proteins and which are of comparatively much recent origin, on the other hand, show a greater weightage towards and wider spread over higher densities; the tubulin and globin genes occupy middle ranges.

This change in cluster density of the exons with evolution is also seen within a gene type and, in some instances, in comparable exons. For example, the separation of the alpha, zeta and myoglobin chains is believed to have taken place between 400 and 500 million years ago, whereas the development of the beta, delta and epsilon globins occurred approximately 300 million years later[29]; the cluster densities of the exons of the two groups show a very clear difference reflecting, we

Table 4. Compositional variance of different exons of selected genes

EMBL AC (and ID)		Total	A(%)	C(%)	G(%)	T(%)	STD	$\frac{G+C-A-T}{A+C+G+T}$ (%)
M12086 (GGMYHE)	E16	310	26.13	28.39	24.84	20.65	2.82	6.45
Human beta globin	E1*	92	18.48	20.65	39.13	21.74	8.24	19.57
	E2	223	19.73	26.01	29.15	25.11	3.39	10.31
	E3*	129	20.93	28.68	27.13	23.26	3.07	11.63
	Avg E1 + E3	*111*	*19.70*	*24.67*	*33.13*	*22.50*	*5.01*	*15.60*
GGTUB4B	E1	57	28.07	10.53	33.33	28.07	8.63	12.28
	E2*	109	22.94	20.18	30.28	26.61	3.80	0.92
	E3*	111	14.41	22.52	28.83	34.23	7.38	2.70
	E4	1073	20.41	16.31	31.13	32.15	6.81	5.13
	Avg E2 + E3	*110*	*18.68*	*21.35*	*29.55*	*30.42*	*5.59*	*1.81*
X03810 (DMHSP82)	E1*	*149*	*37.58*	*15.44*	*16.78*	*30.20*	*9.28*	*35.57*
	E2	2151	25.38	27.06	28.41	19.15	3.54	10.93

Exon numbers given above refer to those appearing in the EMBL database. The starred items are the ones with base lengths of around 100. Where more than one such exon is available, the numbers are averaged and shown in italics.

Table 5. Average cluster densities of different globin gene types

Globin	Exon 1	Exon 2	Exon 3	Average
Alpha	0.51	0.21	0.47	0.40
Zeta	0.74	0.22	0.38	0.45
Myo	—	—	—	0.49
Beta	0.89	1.37	0.65	0.97
Delta	0.79	1.13	0.74	0.88
Epsilon	0.68	2.06	0.96	1.23

believe, the evolutionary time gap; the densities in the plots of the former group average between 0.4 and 0.5, whereas those for the latter group are around 0.9 to 1.2 (Table 5). Comparing the densities within one exon type, for instance the third exon of the beta globin genes, there is a progressively higher density with evolution (Figure 3 *a*); comparing the base composition in the first and second half of the exons separately in the samples tested, the standard deviation of the base composition shows a small but perceptible shift (more noticeable in the second half than in the first) towards smaller values as we go from mouse (std of second 63 bases is 6.8%) through rat (std: 5.2%), lemur (5.7%), rabbit (4.7%) to human (4.3%), implying growing homogeneity in distribution of bases within the exons. Similarly, comparison of the frequencies of occurrence of cluster densities of the myosin heavy chain genes of the nematode, chicken and rat shows a slight, but again perceptible, shift in the distribution pattern from nematode to chicken to rat MHCs (Figure 3 *b*); a Gaussian fit to the bar charts would give peaks at 0.65 (nematode), 0.85 (chicken) and 1.1 (rat) MHC cluster densities.

A closer examination of cluster densities vs lengths (Figure 4 a) indicates that the more recent genes show greater fragmentation and smaller lengths of the coding region segments while the earlier genes have a larger fraction of exons with large lengths, say above 500 bases. The scattergrams of cluster densities in the maps of the introns and exons with respect to their segment lengths for all the genes tested also show that introns generally plot to low density index of up to 0.5 irrespective of their lengths (Figure 4 *b*), while exons, on the other hand, show cluster densities spanning a wide range but mostly for segment lengths of up to about 400 bases (Figure 4 *a*), whereas the longer exons tend to open up to lower densities. These statistics correspond closely with the results of Naora and Deacon[25] on intron–exon frequencies vs lengths established some time ago.

Summary and conclusions

Comparing the intron–exon representations, therefore, one may conclude that (i) evolutionary growth leads to greater homogeneity in exon base composition and distribution in the eukaryotes, whereas introns show no such comparable changes; this is seen between gene types as well as within gene types for different species; (ii) that in the case of the bacteriophage lambda and the M13, the representation is primarily an open structure like those of the introns; and (iii) that there is a marked tendency for greater fragmentation in exons in later genes, the opposite to which would be expected in the intron-early theory that expects exons to be functional template units of proteins and introns to fall off with evolution.

1. Gilbert, W., *Nature*, 1978, **271**, 571.
2. Crick, F., *Science*, 1979, **204**, 264–271.
3. Nussinov, R., *Comput. Appl. Biosci.*, 1991, **7**, 287–293.
4. Staden, R., *Nucleic Acids Res.*, 1984, **12**, 551–567.
5. Ohshima, Y. and Gotoh, Y., *J. Mol. Biol.*, 1987, **195**, 247–259.
6. Staden, R., *Methods Enzymol.*, 1990, **183**, 163–180.
7. Holley, H. and Karplus, M., *Methods Enzymol.*, 1991, **202**, 204–224.
8. Reddy, B. V. B., Deshpande, M. and Pandit, M. W., *Computers in Biomedicine* (eds Held, K. D., Brebbia, C. A. and Ciskowski, R. D.), Computational Mechanics Publications, Boston, 1991.
9. Mural, R. J., Einstein, J. R., Guan, X., Mann, R. C. and Uberbacher, E. C., *Trends Biotechnol.*, 1992, **10**, 66–69.
10. Kel, A. E., Ponomarenko, M. P., Likhachev, E. A., Orlov, Yu. L., Ischenko, I. V., Milanesi, L. and Kolcahnov, N. A., *Comput. Appl. Biosci.*, 1993, **9**, 617–627.
11. Gilbert,W., *Science*, 1985, **228**, 823–824.
12. Go, M., *Nature*, 1981, **291**, 90–92.
13. Darnell, J. E., *Science*, 1978, **202**, 1257–1260.
14. Doolittle, W. F., *Nature*, 1978, **272**, 581–582.
15. Hickey, D. A., Benkel, B. F. and Abukashawa, S. M., *J. Theor. Biol.*, 1989, **137**, 41–53.
16. Blake, C., *Nature*, 1983, **306**, 535.
17. Sharma, A. K., *Curr. Sci.*, 1995, **68**, 801–806.
18. Stoltzfus, A., Spencer, D. F., Zuker, M., Logsdon, J. M. and Doolittle, W. F., *Science*, 1994, **265**, 202–207.
19. Gates, M. A., *J. Theor. Biol.*, 1986, **119**, 319–328.
20. Nandy, A., *Curr. Sci.*, 1994, **66**, 309–314.
21. Nandy, A., *Curr. Sci.*, 1994, **66**, 821.
22. Nandy, A. and Nandy, P., *Curr. Sci.*, 1995, **68**, 75–85.
23. Nandy, A., *Comput. Appl. Biosci.*, 1996, (in press).
24. Senapathy, P., *Proc. Natl. Acad. Sci. USA*, 1986, **83**, 2133–2137.
25. Naora, H. and Deacon, W. J., *Proc. Natl. Acad. Sci. USA.*, 1982, **79**, 6196–6200.
26. Kishino, H. and Hasegawa, M., *Methods Enzymol.*, 1990, **183**, 550–570.
27. Saccone, C., Lanave, C., Pesole, G. and Preparata, G., *Methods Enzymol.*, 1990, **183**, 570–583.
28. Czelusnik, J., Goodman, M., Moncrief, N. D. and Kehoe, S. M., *Methods Enzymol.*, 1990, **183**, 601–615.
29. Strickberger, *Genetics*, 3rd edn, Ch 36.

Received 25 May 1995; revised accepted 14 February 1996