strategy) and transcription (antigene strategy) by forming duplex and triplex with the complementary isomeric 3′,5′-strand.

1 Padget, R A., Kornarska, M M , Grabowski, P J , Hardy, S H and Sharp, P A , *Science*, 1984, **225**, 898–903
2 Kerr, I M and Brown, R E , *Proc Natl Acad Sci USA*, 1978, **75**, 256–260
3 Lesiak, K , Imai, J , Smith, G F and Torrence, P F., *J Biol Chem* , 1983, **258**, 13082–13088
4 Lohrman, R and Orgel, L E , *Tetrahedron*, 1978, **34**, 853–855
5 Dhingra, M M and Sarma, R H , *Nature*, 1978, **272**, 798–801
6 Parthasarathy, R , Malik, M and Fridey, S M , *Proc Natl Acad Sci USA*, 1982, **79**, 7292–7296
7. Ball, A L., in *Enzymes* (ed Boyer, P D), Academic Press, New York, 1982, Chap 11, pp 281–313
8 Shefter, E , Barlow, M , Sparks, R A and Trueblood, K N , *Acta Cryst B*, 1969, **25**, 895–909
9 Kondo, N S , Holmes, H M , Stempel, L. M and Ts'o, P. O P , *Biochemistry*, 1970, **9**, 3479–3498
10 Doornbos, J , Charubala, R , Pfleiderer, W and Altona, C , *Nucleic Acids Res* , 1983, **11**, 4569–4581
11 Gopalakrishna, V , Ghadage, R S and Ganesh, K N , *Biochem Biophys Res. Commun.*, 1991, **180**, 1251–1257
12 Kierzek, R., He, L and Turner, D H , *Nucleic Acids Res* , 1992, **20**, 1685–1690
13 Krishnan, R and Seshadri, T. P , *J Biomol Struc Dyn* , 1993, **10**, 727–745
14 Jin, R , Chapman, Jr. W H , Srinivasan, A R , Olson, W K , Breslow, R and Breslauer, K J , *Proc Natl Acad Sci USA*, 1993, **90**, 10568–10572
15 Srinivasan, A R and Olson, W K , *Nucleic Acids Res* , 1986, **19**, 379–384
16 Anukanth, A and Ponnuswamy, P K , *Biopolymers*, 1986, **25**, 729–752
17 Krishnan, R , Seshadri, T P and Viswamitra, M A , *Nucleic Acids Res* , 1992, **19**, 379–384
18 Westhof, E and Sundaralingam, M , *Proc Natl Acad Sci USA*, 1980, **77**, 1852–1856
19 Arnott, S and Hukins, D W J , *Biochem Biophys Res Commun* , 1972, **47**, 1504–1510
20 Weiner, S J , Singh, U C , Kollman, P A and Case, D A , 'Molecular Mechanics and Dynamics Program – AMBER, Ver 3 0', University of California, San Francisco, USA
21 Sekharudu, C Y , Yathindra, N and Sundaralingam, M , *J Biol Mol Struc Dyn* , 1993, **11**, 225–244
22 Kennard, O and Hunter, W N , *Q R Biophys* , 1989, **22**, 327–379
23 Wang, A H -J , Quigley, G. J , Kolpak, F. J , Crawford, J L., Van Boom, J H , Van der Marel, G and Rich, A , *Nature*, 1979, **283**, 680–686

ACKNOWLEDGEMENTS. V L thanks DAE for the award of K. S Krishnan fellowship We thank Bioinformatics Group, MK University Madurai, for the colour pictures

Received 17 October 1994, revised accepted 15 November 1994

Graphical analysis of DNA sequence structure: II. Relative abundances of nucleotides in DNAs, gene evolution and duplication

A. Nandy* and P. Nandy†

*Computer Division, Indian Institute of Chemical Biology, Calcutta 700 032, India
†Department of Physics, Jadavpur University, Jadavpur, Calcutta 700 032, India

A technique for graphical representation of gene sequences on a two-dimensional cartesian coordinate system is shown to highlight visually the relative abundances of nucleotides along a DNA sequence on a global scale. In some sequences such as the rat myosin heavy-chain gene this reveals a rich structure in the DNA map. In several cases the gene sequences are shown to map to an almost uniform linear structure. The possibility that this may be due to gene evolution by gene duplication, as hypothesized by Ohno, or due to extensive repetitive segments in the sequence is discussed with reference to myosin heavy-chain genes, where the rod-encoding part is known to have a large number of repeats, and the kinetoplast genes of e.g. *L. tarentolae*, where almost the entire gene fragment is composed of six repeating sequences. This feature of the graphical representation provides an easy analytical tool to identify the parts of a gene sequence with large repetitive segments. A comparison is made with the chaos generator diagrams of Jeffrey[1] and the two methods are shown to complement each other in the analysis of gene sequences.

ONE of the interesting problems in molecular biology concerns the interpretation of base composition and distribution in long DNA sequences. While several approaches have led to identification of small segment motifs[2, 3] such as signal sequences, the TATA box, repeats and hairpin loops, techniques for analysis of the total span of a DNA sequence still remain elusive. There has been a renewal of interest in this problem in recent times, brought about by the chaos generator technique of

Jeffrey[1], where long DNA sequences of vertebrates have been shown to form characteristic patterns and which Burma *et al.*[4] have extended to gain new insights into genome structure. Peng *et al*[5] have demonstrated through an application of the random walk model that long DNA sequences with sizeable intron content have long-range correlations which, significantly, are not evident in intronless coding sequences. In an exhaustive study of over 25,000 DNA sequences, Voss[6] found long-range fractal correlations and prominent short-range periodicities for different genus and species.

To search for global patterns in the total span of a DNA sequence, we have proposed a graphical technique that has the advantage of enabling visualization of the nucleotide distribution pattern of individual gene sequences and comparison of different sequences to determine areas of approximate visual similarities. It also appears to be possible to extend the technique for a rapid search of megabase sequences for sequence patterns for some classes of conserved gene sequences[7]. The technique is useful in highlighting global features of base distribution and arrangement that may not be readily apparent in the letter series representation of the gene sequence, as we have seen in the case of the globin genes presented in our earlier paper[7]. A similar approach had been presented earlier by Gates[8]. Both methods give rise to sequence maps that reflect sequence distribution, but our choice of axes introduces subtle differences in patterns and interpretation that provide new insights into genome structure analysis as e.g. concerning the existence of long-range correlations in gene sequences[9].

Here we show that this graphical method also enables identification of sections of a sequence that have distinctly different base compositions and get an idea of probable stages in sequence evolution. In particular, it is interesting to observe that portions of a gene that have evolved through extensive gene duplication show up as a pattern that is almost linear on the scale of the plot. We also compare our results with the chaos generator diagrams of Jeffrey[1] to show that the two methods reflect different aspects of gene sequences and thus provide complementary analytical tools.

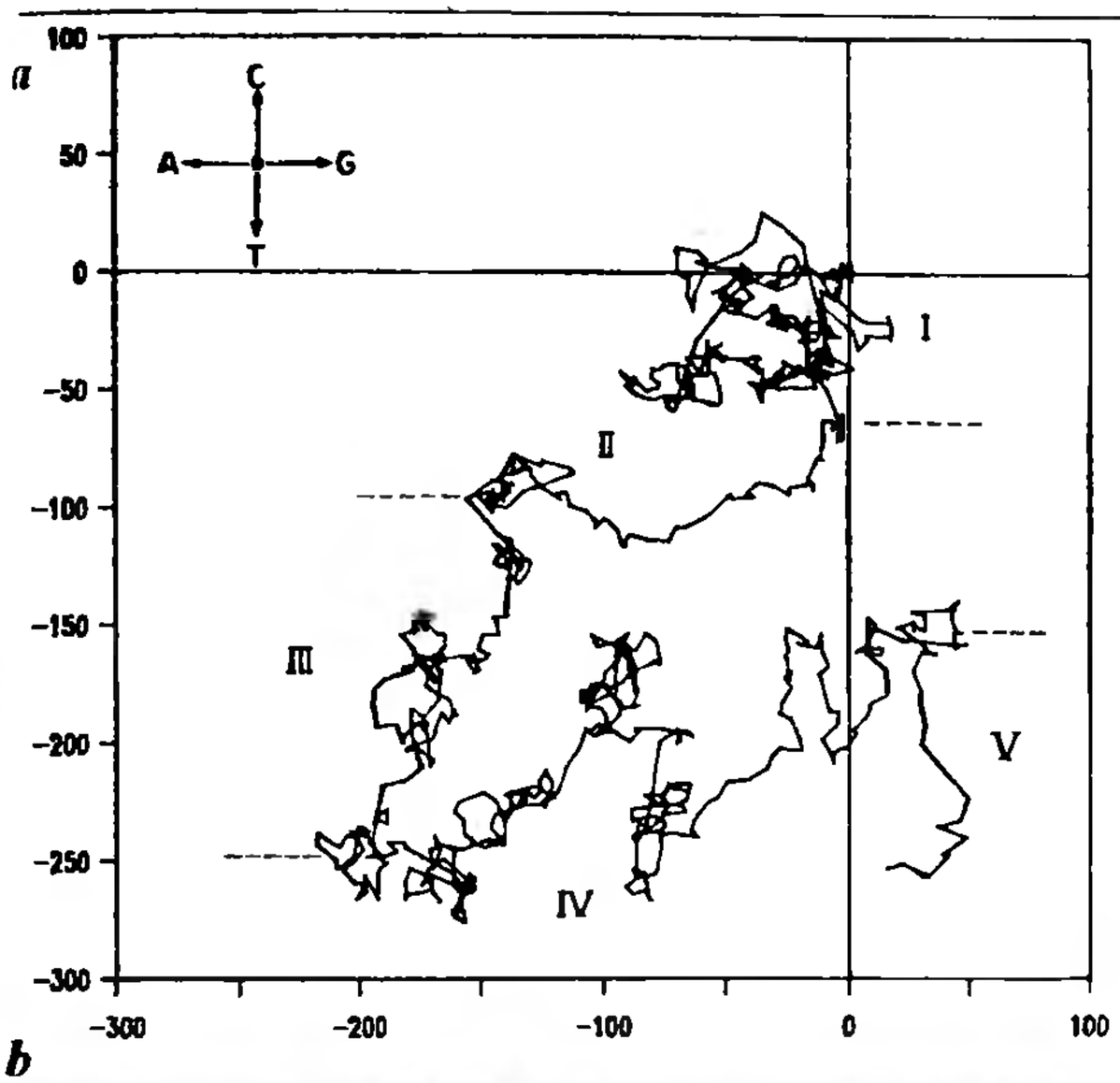
Method

In our approach to representation of DNA sequences, discussed in detail elsewhere[7], we construct a symmetric purine–pyrimidine graph on the cartesian coordinate system, with purines on the x-axis (A in the negative direction, G in positive) and pyrimidines on the y-axis (T in the negative direction, C in 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 in terms of sequence distribution will be reflected from the progression of the points along the map.

At a detailed level, the fine structure of such a map can serve to highlight the relative local abundance of one nucleotide over another within the sequence from a study of the twists and turns along the sequence map (keeping the same scale along each axis). Thus, a segment lying parallel to the x-axis would imply that the difference of the C and T bases remains constant along that stretch of the sequence; a vertical segment would imply a similar constancy between the A and G bases along the stretch. A repeated motif relevant to the scale of the plot would show up as a repeated pattern on the map. However, dyads like AG and CT will appear as a single point on the ACGT-axes system; to display the AG- and CT-rich and other combination sequences, one could construct complementary plots by exchanging the G- and T-axis, or the C- and G-axis. We shall refer to the three plots as the ACGT-, ACTG- and AGCT-axes systems by taking the names clockwise from the x-axis. The preferred choice of axes would normally be dictated *a priori* by the two most dominant bases – e.g. for the chicken myosin heavy-chain gene introns with A, T dominance (as shown later) it would be ACGT-axes system; for the corresponding exons with A, G dominance the preferred axes would be ACTG. (One may also consider the dinucleotide frequencies to decide on pairs of bases for the axes so as to minimize lost information, such as AG and CT on the ACGT-axes system; one would normally choose dyads with low frequencies.) Thus, when considering eukaryotic genome sequences with their strong A, T dominance, the ACGT-axes system would be the natural choice. This axes system also has the added advantage that the purines and pyrimidines are plotted on horizontal and vertical axes, respectively, thus providing a convenient frame for relating to DNA-walk analysis[5], and also, by suppressing transition types of evolutionary changes (A, G and C, T), leads to easier identification of significant evolutionary developments[9].

Application to structure analysis

As an application of this technique to highlight nucleotide abundances, we consider the rat embryonic skeletal muscle myosin heavy-chain (MHC) gene sequence (RNMHCG in EMBL database). The plot of this sequence is shown in Figure 1 *a*. The map can be analysed in several different segments marked I to V on the map. The first part, I, covering the first 8816 bases of the sequence appears like an almost random pattern, followed by a segment (base numbers 8817–10976) proceeding horizontally to the left, implying that it is relatively A-rich. The next segment (10977–14400) proceeds closely along a diagonal towards the bottom



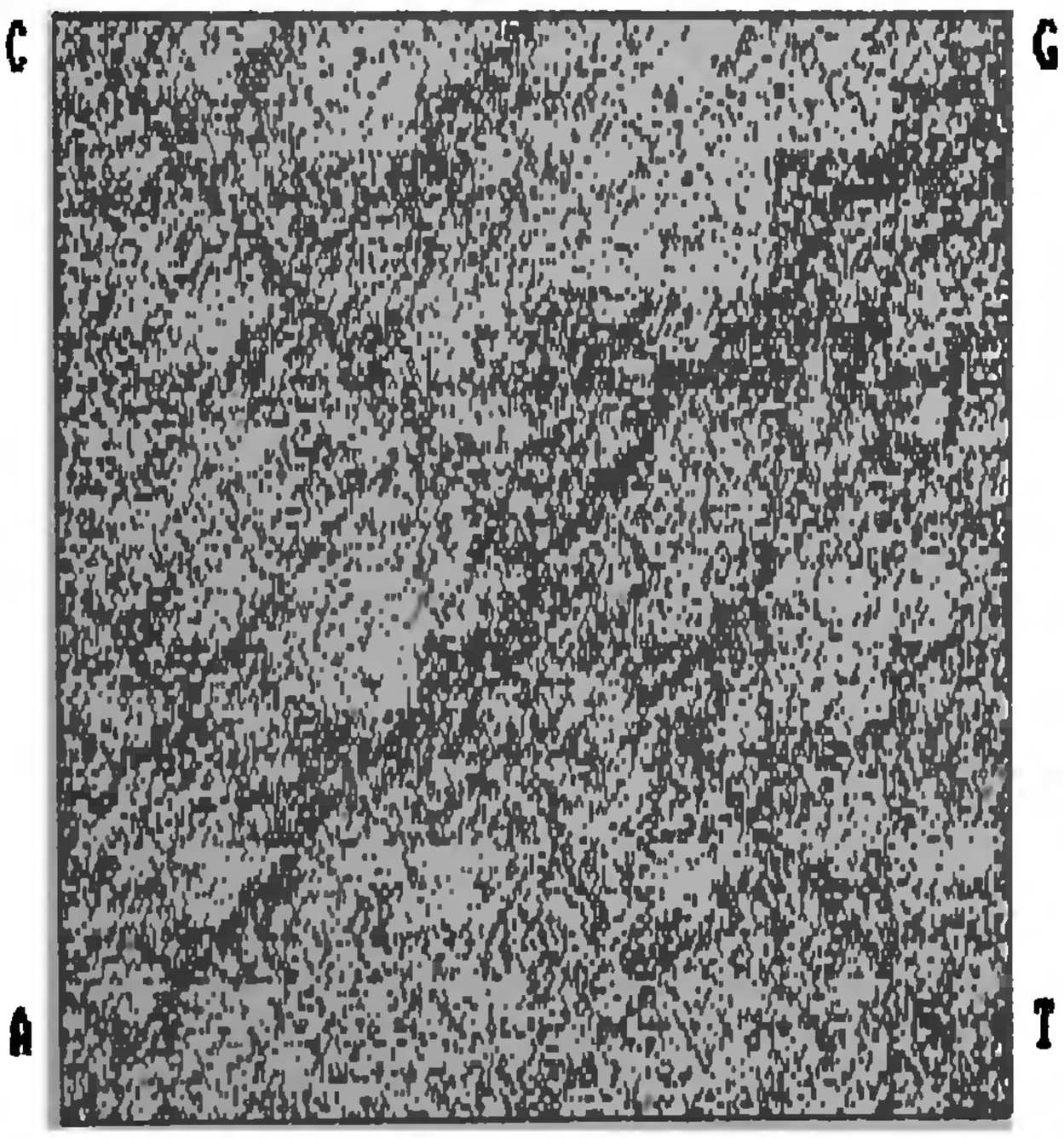


Figure 1. *a*, Map of the rat embryonic skeletal myosin heavy-chain gene complete sequence (25759 bases) plotted on the ACGT-axes system (G – A along *x*-axis, C – T along *y*-axis; the cross shows the direction of progression of the map for each base type). For clarity, only every 30th base position has been plotted. The regions marked in roman numerals cover the following base positions: I: 1–8816; II: 8817–10976; III: 10977–14400; IV: 14401–24720; V: 24721–25759. *b*, CGR diagram of the rat embryonic MHC gene. Compared to the chicken MHC CGR[3], this diagram shows a wider scatter of points, general blurring of the 'double-scoop' pattern and a noticeable scarcity of points in the lower right triangle of the A-quadrant.

left corner, indicating that it must be AT-rich, while the next two segments (14401–24720 and 24721-end) trace out a path going generally eastwards (with several small C- and T-rich segments breaking the progression) and then down, implying that the local sequences must have more G than A and more T than C, respectively. Counting the nucleotide abundances in each of these segments, we find that indeed in the second segment the A's occur about 25% times more often than G or T and about 50% times more often compared to C (ratio A:C:G:T is 1:0.7:0.8:0.8). In the third segment the A and T occur more than 10% oftener than C or G, in the fourth segment the G's occur 10% more often than A's, and in the fifth segment the T's occur 55% more often while A's and G's are almost equal, leading to the vertical nature of the plot in that region. The first segment (I) has the total number of purines and pyrimidines almost equal, leading to null displacement and an almost random pattern.

It is instructive to compare this map with the chaos generator representation (CGR) of Jeffrey[1]. A CGR diagram of this gene (Figure 1 *b*) shows a relatively more scattered distribution of points, with the double-scoop pattern and fractal-type repetitions much less distinct than for the vertebrates discussed by Jeffrey. This is due mainly to the very large differences in local abundances in different parts of the sequence, which is not the case for genes like chicken myosin heavy-chain, the human β-globin cluster, etc., which produce a neat double-scoop pattern on the CGR (see below).

Evolutionary significance

Segmentwise analyses of this type may be useful in studying the evolutionary history of a gene or a class of genes, especially in intron-rich sequences. Such clear distinctions between different regions in one gene may imply stages in evolutionary history where different segments probably evolved separately to form the current structure of the gene. The complex structure of the rat embryonic skeletal MHC gene was, in fact, noted by Strehler *et al.*[10] as being indicative of different evolutionary pathways for the different segments of the gene. From the conservation of intron locations between mammalian and nematode MHC genes they hypothesized that these may have evolved from a highly split ancestral gene in which intron deletions and insertions may have taken place. The evolution of the rod-encoding segment with its dispersed repetitive 28-residue coding unit is considered to have evolved by exon duplication in which introns were later insertion events with wide variety in sizes and very little sequence homology. Thus, insertions, deletions and fast divergence of intron sequences in various segments of the rat embryonic MHC gene would have led to the current structure and it is therefore not surprising that the map of the complete gene with its strong intron domination shows such a complex shape in our sequence

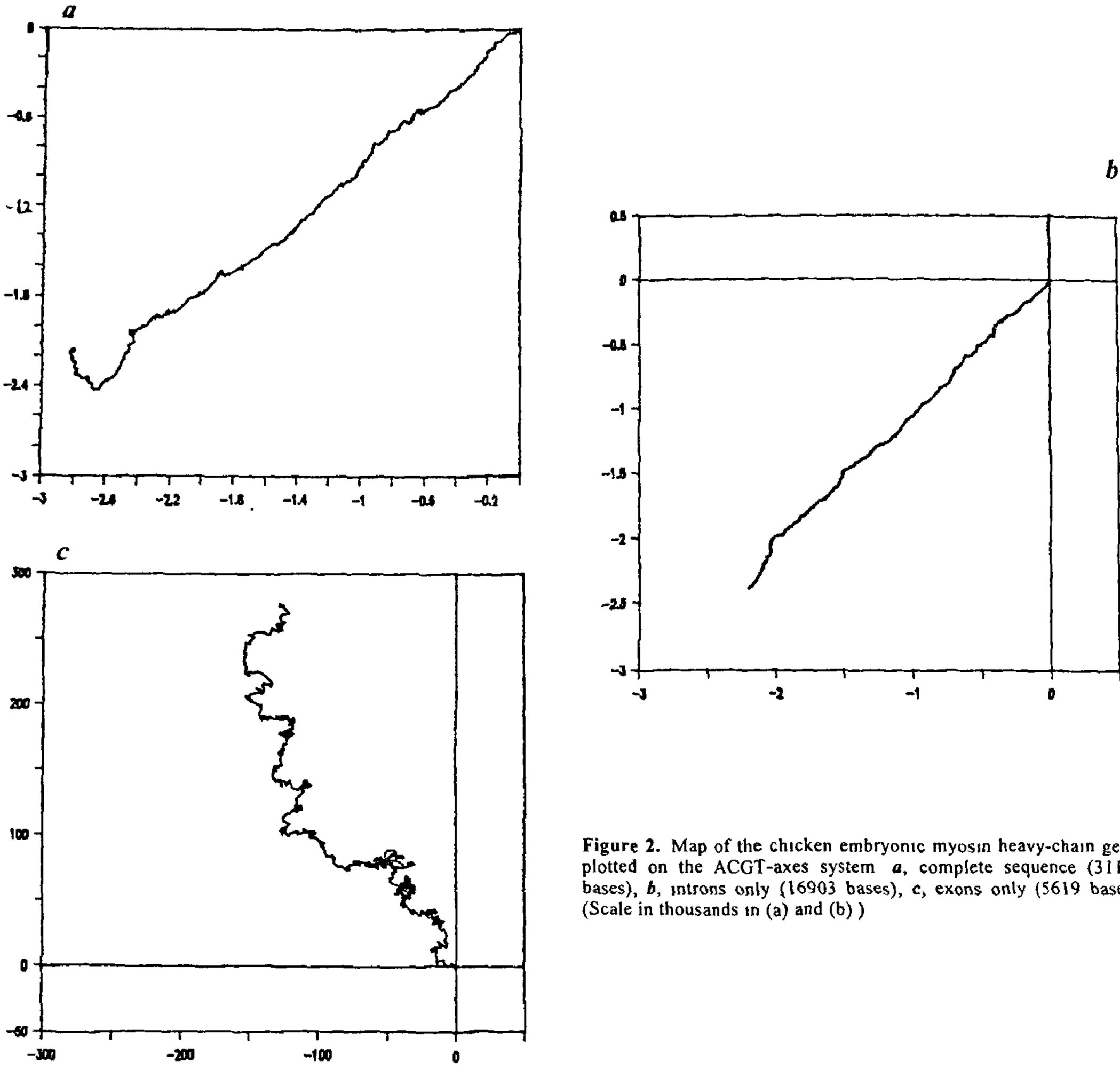


Figure 2. Map of the chicken embryonic myosin heavy-chain gene, plotted on the ACGT-axes system *a*, complete sequence (31111 bases), *b*, introns only (16903 bases), *c*, exons only (5619 bases) (Scale in thousands in (a) and (b))

plot (where regions I–III cover almost the entire globular head segment of the gene and region IV the rod-encoding part).

The assumption of gene evolution by gene duplication[11] can be expected to lead, depending upon the extent of such duplication, to a more uniform distribution of the nucleotide abundances and thus a more rich CGR pattern, and possibly a less convoluted map in our representation. In the case of the chicken myosin heavy-chain gene (GGMYHE) the CGR shows the familiar double-scoop pattern of eukaryote vertebrates; the pattern is found to be similar even at different length scales of the genome, which can be interpreted as being indicative of at least a part of the genomic evolution having occurred through gene duplication[4]. In our representation also we find on a global scale an almost uniform pattern: the sequence plot shows an approximately constant slope for over 70% of the sequence (Figure 2 *a*). This implies that this stretch must be constructed out of one or more repeating segments with a A_m/T_n dominating motif and the relative nucleotide abundances over this long stretch must be so organized as to conserve the ratio $(C-T)/(G-A)$; we have seen this to be true by taking sample sizes of 20% and above of the sequence at a time in this region.

That this feature is due to intron dominance becomes clear when we plot the intron and exon segments of chicken MHC separately. In the case of introns the previous approximate linearity with a constant slope factor of $(C-T)/(G-A)$ is clearly retained (Figure 2 *b*), whereas the exon segment being predominantly

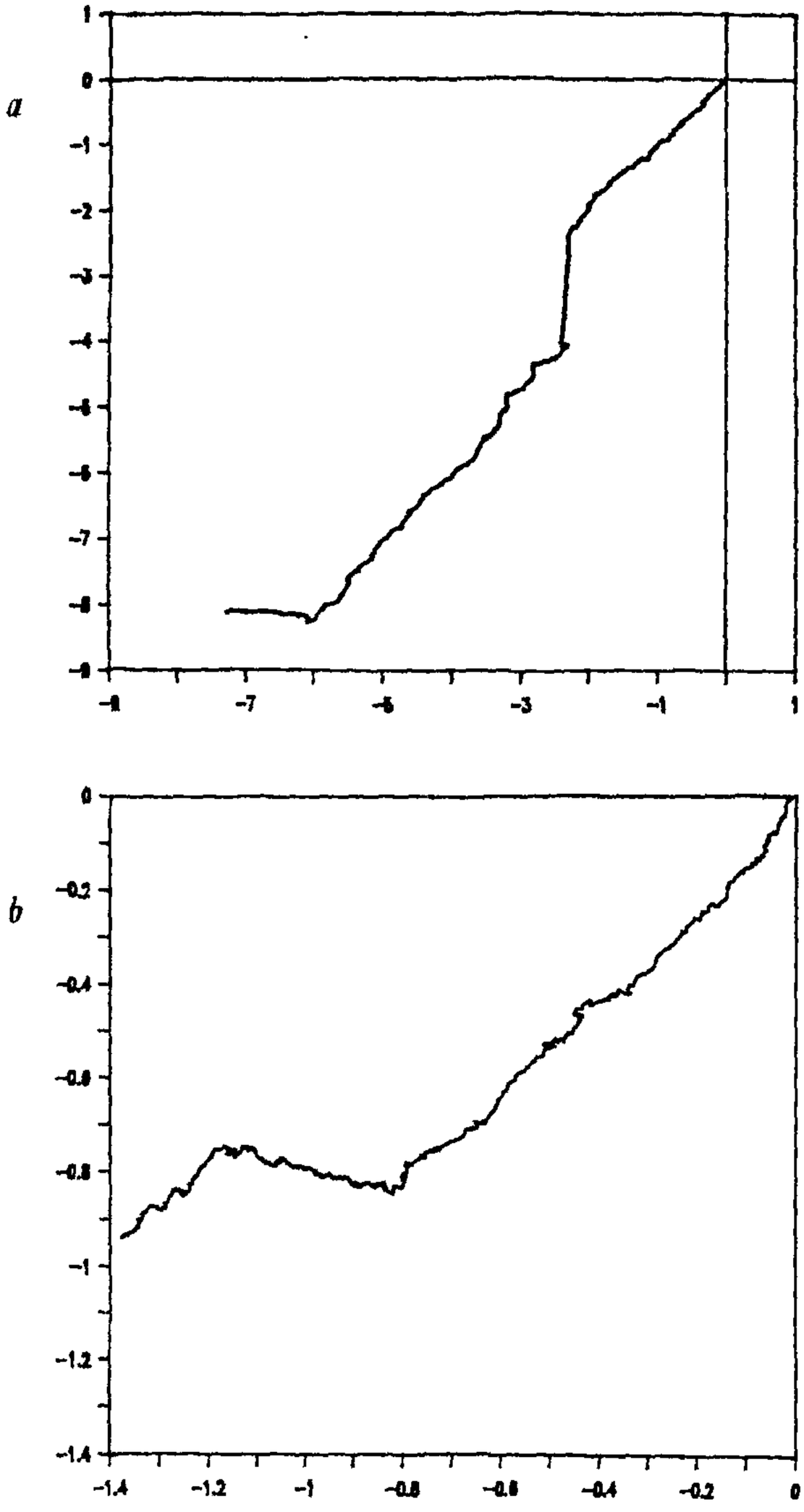


Figure 3. Maps of *a*, the human β-globin complex region on chromosome 11 (73326 bases) and *b*, *C. elegans* myo-1 myosin heavy-chain gene (12241 bases), complete sequences including 5′ and 3′ flanking regions. (Scale in thousands, axes ACGT)

A/G-rich takes an upward curvilinear path (Figure 2 *c*) quite distinct in nature from the intron map. This is also seen in the case of the *C. elegans* myosin heavy-chain gene and the human β-globin gene region from chromosome 11: the exon segments are again found to be rich in structure whereas the intron segments produce maps, on the scale drawn, with distinctly large relatively uniform stretches; the maps of the complete genes with strong intron dominances are shown in Figure 3. The degree and extent of the uniformity will depend on the size of the repeating unit and the extent of repetition relevant to the scale of the plot; the smaller the unit and the larger the number of repetitions, the more linear will be the map of the segment on the plot, and it will map in directions dictated by the base combinations of the repeating motif, as can be seen from the plots of these two genes.

Considering the intron segment of the chicken myosin heavy-chain gene, we find indeed that there are a large number of repeating small structural units (see Table 3). A dinucleotide analysis shows that the four A and T combinations are in overwhelming majority, comprising 38.4% of the total dinucleotides. Taking oligonucleotides of e.g. 5 bases length in all combinations such as ATATA, TTTTA, CGGCC, TGGGT, etc., 38.5% of all repetitive segments are found to be A, T oligonucleotides with negligible contribution from homopolymers like AAAAA, etc. Increasing the oligonucleotide length to 8 again shows A, T combinations to be the dominant type, with oligonucleotides with alternating A and T comprising almost 60% of the total tested; this alternating AT motif is, of course, the dominating dinucleotide and is found also clustered in long combinations such as 5 times for $(AT)_5$ and once for $(AT)_6$. Thus, given this A, T dominance with a large number of repeats of short oligonucleotides, it is not surprising that the intron map on the ACGT-axes system turns out to be nearly linear in the third quadrant at almost 45° angle. The exon sequence, on the other hand, does not show such large A, T repeats (Table 3) and thus does not form a linear structure on the ACGT-axes, but, as we show later, can be expected to do so on the ACTG-axes system by virtue of the dominating A, G repeats.

Part of the reason for differences in the intron–exon maps lies in the compositional differences of the bases in the two cases. A comparison of the A, C, G, T abundances will show that the introns of most of these genes are predominantly AT-rich, giving an AT-rich character to the entire sequence; the exons, on the other hand, are seen to have comparatively more equal distribution of the four bases. Table 1 shows the statistics of the compositional variances for a sample from a wide class of genes. It can be seen that within one standard deviation, the exon composition is uniform across the four bases for the entire spectrum, whereas in the case of introns there is a wide difference. Exon maps, therefore, would generally tend to cluster to null to small displacement, whereas intron maps would tend to be more open and cover larger distances, as can easily be seen e.g. by considering the end points on an ACGT map of any of the genes listed in Table 1.

Repetitive sequences maps

Sequences that are highly repetitive, whether due to gene duplication or otherwise, will also lead to almost

Table 1. Sequence composition differences – AT-rich sequences

	EMBL Code	Total sequence length	Percentage A	C	G	T	$\frac{C+G-A-T}{A+C+G+T}$%
EXONS	AGHAPSE	428	26 17	22 90	25 47	25 47	–3 27
	AGHBD	444	19 37	25 00	31 76	23 87	13 51
	CEMYO1	5814	30 75	23 53	23 72	22 00	–5 50
	CEMYO2	5841	29 99	24 69	24 72	20 60	–1 18
	DMHSP82	2300	26 17	26 30	27 65	19 87	7 91
	GGMYHE	5619	31 00	22 58	28 72	17 69	2 62
	GGTUBA4A	681	24 23	28 19	21 88	25 70	0 15
	GMHSP	678	28 02	18 44	23 30	30 24	–16 52
	HSHBB	2220	22 07	24 91	29 28	23 74	8 38
	Average		26 69	23 54	26 83	22 93	
	STD		3 76	2 34	2 83	3 85	
INTRONS	AGHAPSE	1451	26 74	19 57	21 09	32 60	–18 68
	AGHBD	1013	29 81	16 78	16 98	36 43	–32 48
	CEMYO1	1648	34 65	13 47	14 44	37 44	–44 17
	CEMYO2	958	30 58	13 88	15 76	39 77	–40 71
	DMHSP82	1129	30 47	17 45	17 36	34 72	–30 38
	GGMYHE	16903	31 79	17 67	18 81	31 74	–27 05
	GGTUBA4A	183	24 59	18 03	15 85	41 53	–32 24
	GMHSP	388	31 70	12 37	15 46	40 46	–44 33
	HSHBB	41850	28 84	18 43	19 78	32 94	–23 57
	Average		30 57	16 20	17 46	35 76	
	STD		2 21	2 46	2 09	3 11	
TOTAL	AGHAPSE	1879	26 61	20 33	22 09	30 97	–15 17
	AGHBD	1457	26 63	19 29	21 48	32 60	–18 46
	CEMYO1	7462	31 61	21 31	21 67	25 41	–14 04
	CEMYO2	6799	30 08	23 17	23 46	23 30	–6 75
	DMHSP82	3429	27 59	23 39	24 26	24 76	–4 70
	GGMYHE	22520	31 59	18 89	21 28	28 23	–19 64
	GGTUBA4A	864	24 31	26 04	20 60	29 05	–6 71
	GMHSP	1066	29 36	16 23	20 45	33 96	–26 64
	HSHBB	44070	28 50	18 76	20 26	32 48	–21 96
	Average		29 00	20 17	21 87	28 96	
	STD		2 00	2 34	1 22	3 84	

Table of base composition of exons, introns and total gene and averages of a sample of gene sequences from the EMBL genetic sequence data library The genes are identified by their EMBL codes representing the following

AGHAPSE	*A geoffreyi* (spider monkey) η-globin pseudogene
AGHBD	*A geoffreyi* (spider monkey) δ-globin gene
CEMYO1	*C elegans* (nematode) myo-1 gene for myosin heavy chain
CEMYO2	*C elegans* (nematode) myo-2 gene for myosin heavy chain
DMHSP82	*Drosophila melanogaster* heat shock protein 82
GGMYHE	Chicken embryonic nonmuscle myosin heavy-chain gene
GGTUBA4A	Chicken α-4-tubulin gene – exon 4
GMHSP	Soybean heat shock protein (Gmhsp26-A) gene
HSHBB	Human β-globin cluster on chromosome 11

straight maps in our representation; the presence of long linear segments in maps of this type was in fact also noted by Gates[8] in viral and other sequences as also the rich structures in some exons. We have found that the map (Figure 4) of the *Leishmania tarentolae* kinetoplast maxicircle DNA fragment (2.76 kb) reported by Muhich *et al.*[12] is an almost straight line from beginning to end, due primarily to the fact that this fragment consists almost entirely of repeated segments. These repeats can be grouped into six families, some of which are present throughout the remainder of the 12 kb divergent region of the maxicircle, and are oriented in a head-to-tail fashion with the simplest repeats clustered into large arrays. The repeats are again dominated by oligomers of the type (A_mT_n), e.g. AATAATAT, AAATT, etc., and therefore produce an almost linear plot on our ACGT-axes map (like the chicken MHC intron case, but now at a different angle related to the *m/n* ratio). Similarly, the *Crithidia fasciculata* mitochondrial maxicircle DNA containing the gene for the cytochrome oxidase subunit III (coxIII), the 5′ flank of the human β-globin region on chromosome 11, the chicken β-4-tubulin gene, the yeast

Table 2. Constant-slope region for selected gene sequences

Sequence	Code	Length	Translation region	Linear region	Percentage of sequence
D. discoidium MHC gene	DDMYHC	6681	70–6420	1740–6420	70.05
Chicken embryo MHC gene	GGMYHE	31111	2236–24757	1–21500	69 11
Chicken β-4-tubulin gene	GGTUB4B	3153	380–2782	360–2760	76 12
Human β-globin region on chromosone 11, 5′ flank	HSHBB1	3365		1080–2760	53.28
Human α-cardiac MHC	HSMHCAG1	2366	1253–1540	1–1678	70 92
C. fasciculata mitochondrial DNA	MICFCOX3	3353	1102–3353	1380–3353	58 84
L. tarentolae maxicircle DNA fragment	MILTMXC1	2759	1–2759	1–2759	100 00
Yeast heat shock protein	SCHSP90	2734	333–2462	1200–2734	56.11
V. cholera toxin	VCCTX	2020	515–2020	480–2020	76.24

A sample analysis of 9 sequences selected across a wide variety of phylogenetic types. The table lists the sequence name, the EMBL ID code, the length of the sequences in terms of the number of nucleotides and the translation region, where noted separately. The next column lists the region in terms of base numbers over which the map of the sequence in the ACGT-axes system appears to be predominantly linear, and this is indicated in terms of percentage of the sequence length as given in the last column. All data for the sequences are taken from EMBL sequence library, Release 31. The sequences are arranged alphabetically by their codes.

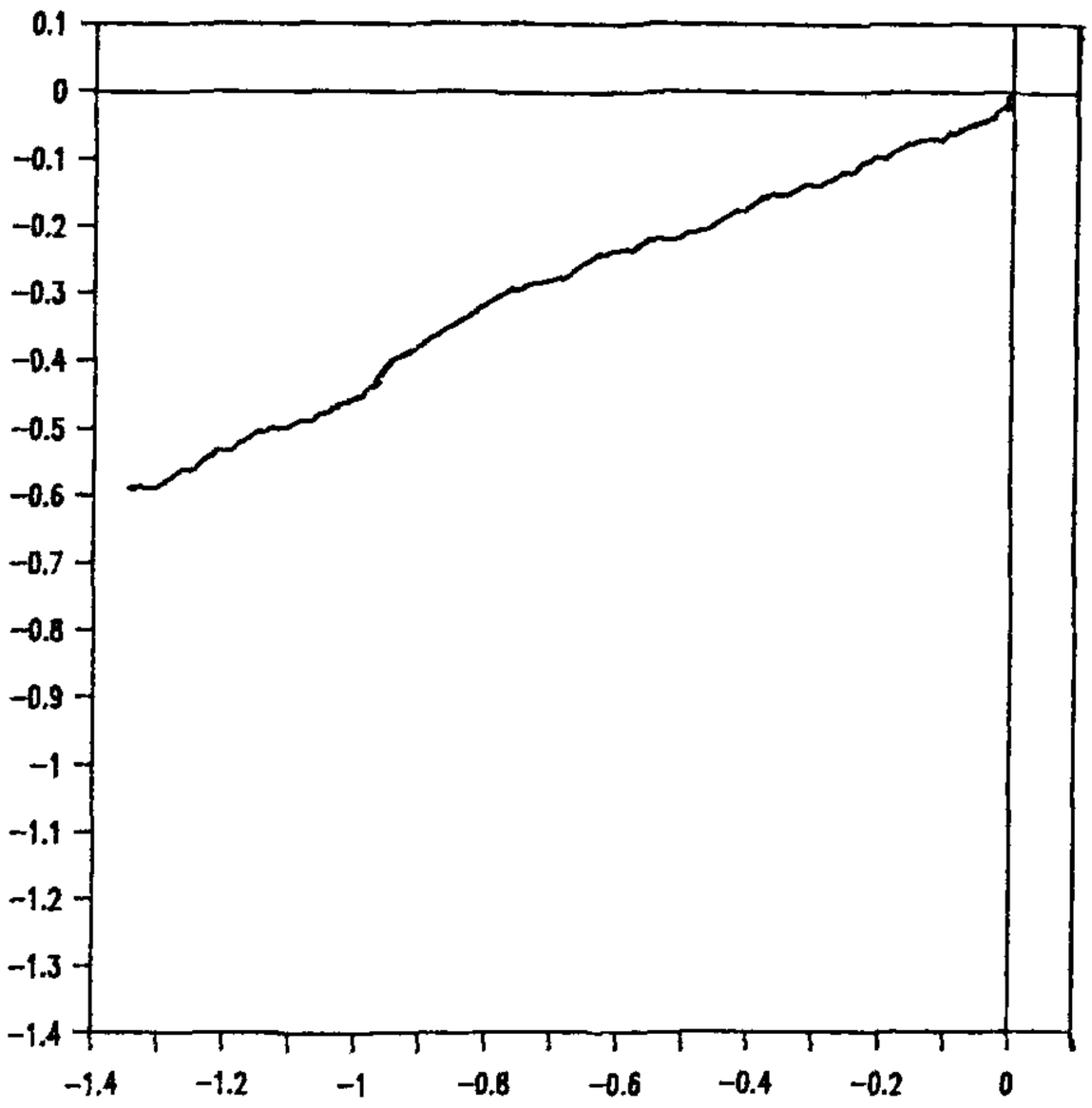


Figure 4. Map of the *L tarentolae* kinetoplast maxicircle DNA fragment (2759 bases) (Scale in thousands, axes: ACGT)

heat shock protein hsp90, etc., all show approximately linear-like segments for considerable lengths of their sequences (Table 2), implying the presence of long-range repetitive sequences and therefore a significant level of sequence homogeneity in this context. In fact, fairly long uniform stretches are noticeable in a wide variety of sequences, implying that repetitive sequences may be quite widespread in nature.

However, in those cases where the compositional variances of a gene or its segment are low, the relevant maps may turn out to be rich in structure as in the case of the exons, or in cases such as the rat myosin heavy-chain gene discussed earlier. As a rule of thumb, we have found that in such cases, say on the ACGT-axes system, the ratio $(G+C-A-T)/(A+C+G+T)$ is less than 5%, but the onset of rich structural map can take place at slightly higher ratios also. If uniformity in base distribution and a corresponding linear-like map in our graphical representation is to be interpreted in terms of gene duplication[11], then in cases where the map shows rich structure and significant variations along the path, one could conclude that the gene structure may have evolved by gene deletions and insertions, but the contribution to the total by way of gene duplication may not have been significant.

It is instructive at this stage to compare the intron and exon segments of the rat myosin heavy-chain gene with that of the chicken myosin heavy-chain gene, where the intronic, and therefore the overall, base compositions are markedly different. Whereas in the case of the chicken MHC the intron base composition is highly skewed towards the A, T bases and the exon base composition is almost uniform, the rat MHC exons show an A, C, G, T mix of 28.9%, 24.4%, 29.9%, 16.9%, respectively, while in the case of introns this is almost uniform at 25.3%, 23.0%, 25.3% and 26.4%. Thus, a map of only the exons of the rat MHC gene on the ACGT axes can be expected to show a relatively more open curve compared to a map of the introns alone, which would generate a complex structure (Figure 5 *c*, *d*).

However, in view of the strong A, G dominance of the exon compositions of the MHC genes, the natural choice of axes should be the ACTG-axes system, and this immediately reveals explicitly the latent structure in all these exons; the importance of the choice of axes system is seen by the maps in Figure 5 for the rat MHC. Figure 5 *a* plotted on the ACTG-axes system shows the exon sequence starting with a small congestion pattern

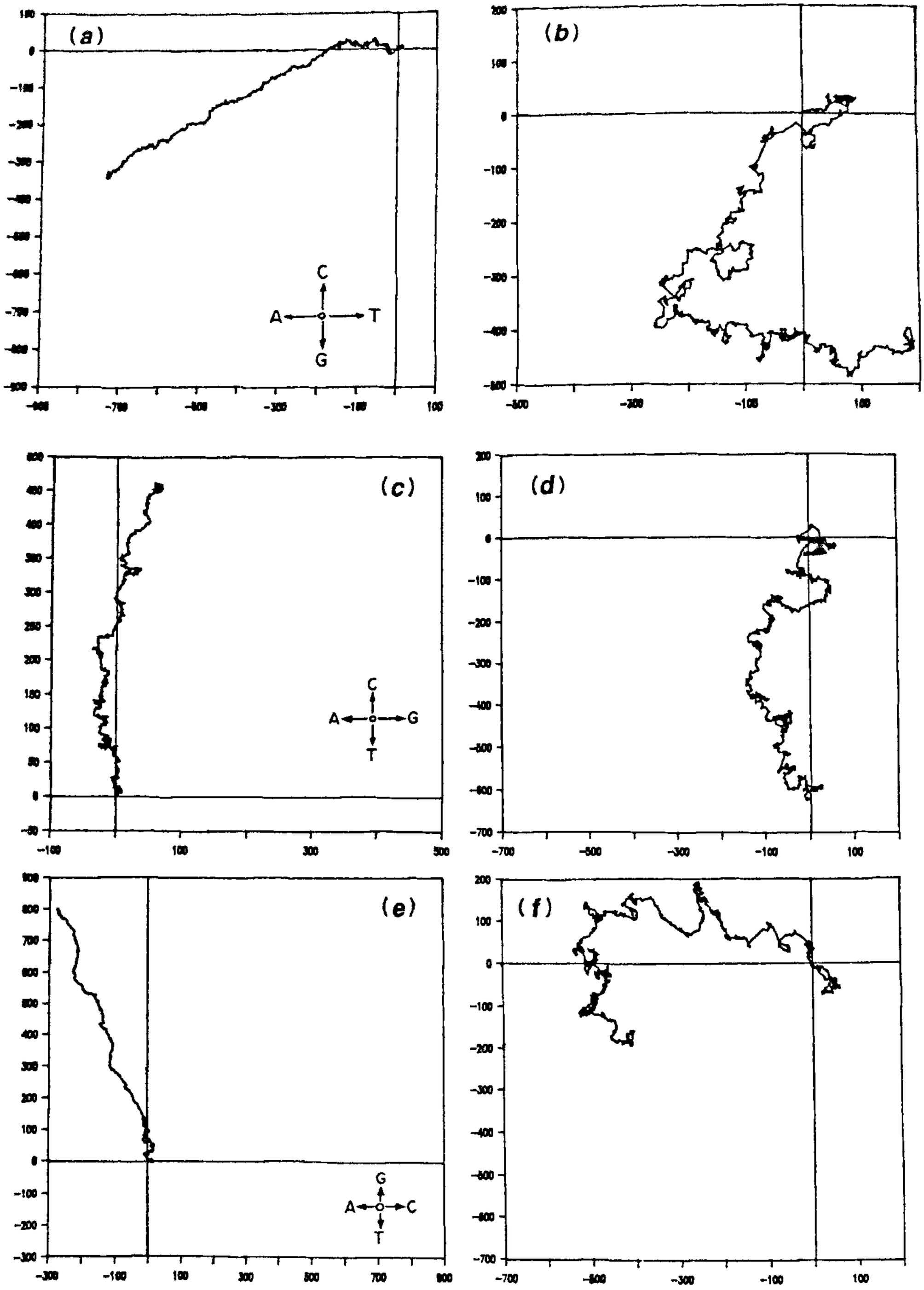


Figure 5. Map of the rat MHC gene segments plotted on various axes systems *a*, exons only (6035 bases), ACTG-axes, *b*, introns only (17750 bases), ACTG-axes, *c*, exons and *d*, introns on ACGT-axes, *e*, exons and *f*, introns on AGCT-axes systems

Table 3. Frequency table for repetitive units in chicken MHC sequence

Chicken MHC intron sequence (16903 bases)

Dinucleotide frequencies

AA: 1793	(10 6%)	AC: 877	(5.2%)	AG: 1155	(6 8%)	AT. 1547	(9 2%)
CA: 1224	(7 2%)	CC. 550	(3 3%)	CG: 89	(0 5%)	CT. 1123	(6 6%)
GA: 1017	(6 0%)	GC: 580	(3 4%)	GG: 693	(4 1%)	GT: 888	(5 3%)
TA 1338	(7 9%)	TC. 979	(5 8%)	TG: 1241	(7 3%)	TT: 1806	(10 7%)

Other frequencies

	$m = 0, n = 5$	$m = 1,$	$n = 4$	$m = 3,$	$n = 2$	Total	$(m + n = 5)$	$m = 4, n = 4$
$A_mT_n + T_mA_n$	193	631	(36 2%)	927	(35 6%)	1751	(38 5%)	39
$A_mG_n + G_mA_n$		330	(18 9%)	527	(20.3%)	857	(18 8%)	3
$A_mC_n + C_mA_n$		187	(10 7%)	299	(11.5%)	486	(10.7%)	1
$T_mG_n + G_mT_n$		229	(13 2%)	335	(12 9%)	564	(12 4%)	3
$T_mC_n + C_mT_n$		347	(19 9%)	498	(19.1%)	845	(18 6%)	3
$G_mC_n + C_mG_n$	10	19	(1.1%)	16	(0 6%)	45	(1 0%)	–
	203	1743		2602		4548		

Chicken MHC exon sequence (5619 bases)

Dinucleotide frequencies

AA 528	(9 4%)	AC: 294	(5 2%)	AG. 652	(11 6%)	AT 268	(4 8%)
CA. 493	(8 8%)	CC 314	(5 6%)	CG. 117	(2 1%)	CT: 345	(6 1%)
GA: 622	(11 1%)	GC: 411	(7.3%)	GG: 383	(6 8%)	GT. 197	(3 5%)
TA: 98	(1 7%)	TC. 250	(4 4%)	TG: 462	(8 2%)	TT: 184	(3.3%)

Pentanucleotide frequencies

	$m = 0, n = 5$	$m = 1,$	$n = 4$	$m = 3,$	$n = 2$	Total	$(m + n = 5)$
$A_mT_n + T_mA_n$	6	27	(8 3%)	37	(4.3%)	70	(5 8%)
$A_mG_n + G_mA_n$		150	(46.3%)	490	(56 3%)	640	(53 2%)
$A_mC_n + C_mA_n$		64	(19 8%)	149	(17 1%)	213	(17 7%)
$T_mG_n + G_mT_n$		27	(8 3%)	72	(8.3%)	99	(8 2%)
$T_mC_n + C_mT_n$		36	(11 1%)	87	(10 0%)	123	(10 2%)
$G_mC_n + C_mG_n$	4	20	(6 2%)	35	(4 0%)	59	(4.9%)
	10	324		870		1204	

Frequency table for repeated motifs in chicken myosin heavy-chain gene intron and exon sequences. For the intron sequence, we show the complete dinucleotide and pentanucleotide frequencies For units of 8 bases, we consider only some combinations such as ATATATAT, AATTAATT, AAATATTT, AAAATTTT, TTAATTAA, etc., since large homopolymeric units are found to be in low frequencies; thus, the comparison is largely confined to small unit repeats for different base combinations For the exon sequence, only the dinucleotide and pentanucleotide frequencies are shown. The dominance of A, T combinations in introns and A, G combinations in exon sequences is clearly evident

growing into a comparatively smooth structure for most of its length, while the intron sequence, shown in Figure 5 *b*, has a rich convoluted structure reminiscent of the total gene. The exon sequence shows a fairly uniform linear structure for over 3500 bases from exon 22 (AA810) onwards, which comprises the rod-encoding region. (The numbers of A, C, G, T are fairly constant counting for every 600 bases from base number 2700 to 5100 on the exon: 186 ± 9, 136 ± 5, 198 ± 10, 80 ± 6, respectively.) This would be unexpected from a first glance at the complete map of this gene but can be explained if we consider the hypothesis of Strehler *et al* [10] that there has been considerable duplication of exons coding for the rod-like structure of the myosin, whereas the intron segments in contrast have evolved primarily by insertions and deletions. Similar maps are produced by the exon segments of the chicken embryonic MHC, the *D. discoideum* nonmuscle MHC and the several nematode MHC genes (Figure 6), irrespective of the nature of the map of the total gene: all the exon maps on the ACTG-axes have a complex structure for approximately the first 2500 bases, following which the balance of around 3500 bases produce a relatively more uniform linear structure compared to e.g. the structure for the rat MHC introns. This is only to be expected since these exons have a high level of homology and all the MHC genes are known to have a high degree of 28-residue and 196-residue repeats in the rod-encoding segment, first critically analysed for the nematode gene by McLachlan[13]. These repeats are dominated by oligomers of the type A_mG_n (see e.g. Table 3), thus showing up more prominently on the map in the ACTG-axes system. The CGR plots of the rod-encoding parts of the rat and chicken exons show, in fact, strong congestion patterns along and parallel to the AG diagonal, and a dinucleotide analysis shows that for all these exon segments, the AG/GA are the dominant dinucleotides.

We note in passing that complete sequences of most of these MHC genes, the genes for the human β-globin

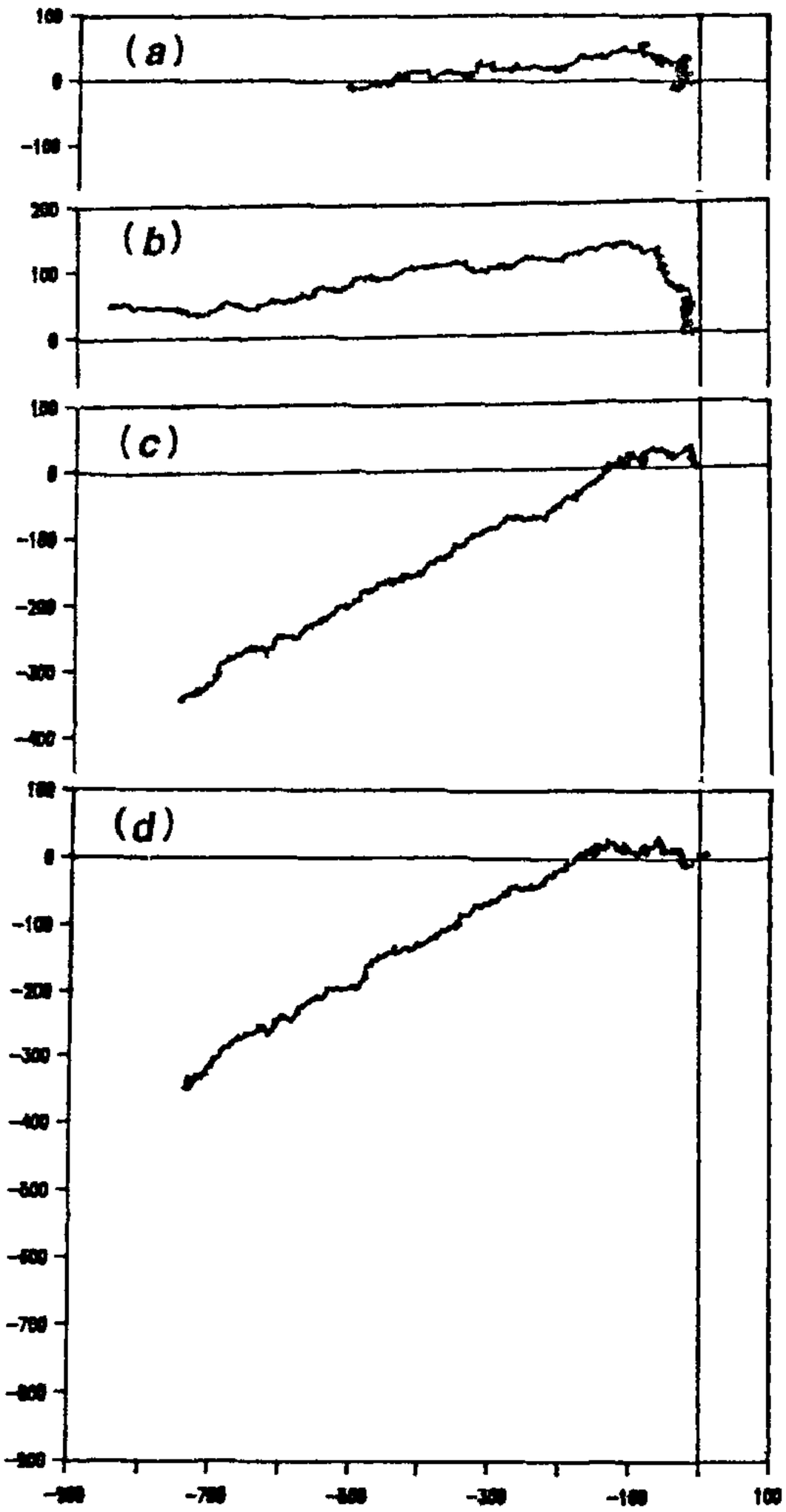


Figure 6. Maps of the exon segments of several MHC genes plotted on the ACTG-axes system: *a*, *C. elegans* myo-1 myosin heavy chain, *b*, *D. discoideum* myosin heavy chain, *c*, chicken embryonic myosin heavy chain, *d*, rat embryonic skeletal myosin heavy chain.

complex, and other vertebrate genes produce fairly similar patterns in a CGR diagram (e.g. the double-scoop depletion regions for vertebrate sequences in varying degrees of clarity), whereas the maps in our graphical system show directly visually characteristic differences as also large-scale variations in base composition in different segments of a gene, as in the case of the rat MHC or the globin genes[7]. The CGR technique thus will be useful for noting the presence (through congestion patterns) or absence (through depletion zones) of a few base motifs such as CG, CGG, etc., in a global context within major classes of animals, whereas the present mapping technique can provide estimates of relative richness of one or more nucleotides in specific sequence segments.

Long-range correlations

The class of maps indicating long uniform stretches in gene sequences can also be construed as providing a possible clue to the long-range correlations within DNA reported by Peng *et al.*[5] in intron-rich sequences but not in the exons or intronless sequences, although the latter assertion has been subsequently disputed by others[14–16] We have seen that introns are generally more well-ordered over the sequence for the MHC genes such as for chicken and nematode and their large presence in these genes imposes such an order on the global map. This raises the possibility that the long-range correlations noticed by Peng *et al.*[5] could have arisen due to the presence[17] of large repetitive segments by possibly the gene duplication phenomenon, which generates linear maps in our graphical representation.

There is also the possibility that the type of repeats may be of importance in this context. We note that the MHC exons have purines-only repeats like A_mG_n and the introns have predominantly purine–pyrimidine repeats like A_mT_n as shown for the chicken MHCs, and compare with Peng *et al.*'s observation of long-range correlations in only total genes comprising approximately over 70% introns, but not for the coding segments. If we consider large-scale repetitions as the basis for these correlations, then we may conclude that, since the purine/pyrimidine attribute forms a basic unit for their DNA-walk analysis, repetitions arising out of purine–pyrimidine combinations will provide these long-range correlations, but purines-only repeats as in the chicken and other MHC coding segments will not contribute to such correlations. However, the rat MHC remains a noticeable puzzle to this explanation, implying perhaps that long-range correlation effects in gene sequences may have other origins as well[18].

Summary and conclusions

In summary, we conclude that the graphical technique discussed in this paper represents a useful method for analysing gene sequences from few tens to megabases in length. The maps generated in this method identify directly visually regions of different base composition and will be useful in providing an insight into the evolutionary history of the gene segments. This representation is also seen to be complementary to the chaos generator diagrams of Jeffrey[1], but carry more information on the local and global scale; for instance, in the case of the myosin heavy-chain sequences for different vertebrates, the maps in the present graphical

representation exhibit large differences arising out of different base abundances, but the CGRs of these genes appear to have almost similar pattern. We have seen that a detailed analysis of the maps produced in our system can reveal the presence of gene duplication and repeats on a global scale, and this we have found to be fairly widespread in nature; the contraindications can be interpreted as gene evolution through gene insertions and deletions. We have also highlighted through this technique the significant differences in the base compositions of introns and exons, especially in intron-rich sequences, and noted that these tally in part with the observation of Peng *et al.*[5] on long-range correlation effects in gene sequences. Depending upon the sequence composition, different choice of axes could be made to generate the maps, as in the extension to CGR proposed by Burma *et al.*[4], and we have shown how crucial it may be in revealing the deep underlying structures. Nevertheless, the method discussed here is simple to implement and produces more and important information about the distribution of nucleotides along a DNA sequence on a local as well as a global scale that is not readily accessible by other methods currently in use.

1. Jeffrey, H J., *Nucl. Acids Res.*, 1990, **18**, 2163–2170
2. Nussinov, R , *Comput. Appl. Biosci* , 1991a, **7**, 287–293.
3 Nussinov, R , *Comput Appl Biosci* , 1991b, **7**, 295–299
4. Burma, P. K., Raj, A., Deb, J. K. and Brahmachari, S , *J. Biosci* , 1992, **17**, 395–411
5. Peng, C-K., Buldyrev, S V., Goldberger, A L., Havlin, S., Sciortino, F., Simons, M and Stanley, H. E , *Nature*, 1992, **356**, 168
6. Voss, R., *Phys Rev. Lett.*, 1992, **68**, 3805–3808.
7 Nandy, A , *Curr Sci* , 1994, **66**(4), 309–314
8. Gates, M A , *J. Theor. Biol.*, 1986, **119**, 319–328
9. Nandy, A , *Curr Sci* ,1994, **66**, 821.
10. Strehler, E E , Strehler-Page, M A., Perriard, J. C., Periasamy, M. and Nadal-Ginard, B., *J. Mol. Biol* , 1986, **190**, 291–317.
11 Ohno, S., *Evolution by Gene Duplication*, Springer, Berlin, 1970.
12. Muhich, M L., Neckelmann, N and Simpson, L., *Nucl Acids Res.*, 1985, **13**, 3241–3260.
13. McLachlan, A. D., *J Mol. Biol* , 1983, **169**, 15–30.
14 Nee, S , *Nature*, 1992, **357**, 450.
15. Prabhu, V V. and Clavier, J-M., *Nature*, 1992, **359**, 782.
16. Chatzidimitriou-Dreismann, C. A. and Larhammar, D., *Nature*, 1993, **361**, 212–213
17 Maddox, J , *Nature*, 1992, **358**, 103.
18 Karlin, S. and Brendel, V , *Science*, 1993, **259**, 677–680.

ACKNOWLEDGEMENTS. We wish to thank Prof. D. Balasubramanian, Prof M. A. Vishwamitra and Dr M. W. Pandit for encouragement and helpful comments during the early stages of this work. Thanks are also due to the DBT Bioinformatics Cell at the Centre for Cellular and Molecular Biology, Hyderabad, for access to the databases The authors also gratefully acknowledge the extremely helpful comments and suggestions of the referees

Received 24 January 1994, revised accepted 20 August 1994