J. Biosci., Vol. 17, Number 4, December 1992, pp 395–411. © Printed in India.

Genome analysis: A new approach for visualization of sequence organization in genomes

PRADEEP KUMAR BURMA, ALOK RAJ*, JAYANT K DEB* and SAMIR K BRAHMACHARI†
Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India

MS received 22 July 1992; revised 28 September 1992

Abstract. In this article we describe and demonstrate the versatility of a computer program, GENOME MAPPING, that uses interactive graphics and runs on an IRIS workstation. The program helps to visualize as well as analyse global and local patterns of genomic DNA sequences. It was developed keeping in mind the requirements of the human genome sequencing programme, which requires rapid analysis of the data. Using GENOME MAPPING one can discern signature patterns of different kinds of sequences and analyse such patterns for repetitive as well as rare sequence strings. Further, one can visualize the extent of global homology between different genomic sequences. An application of our method to the published yeast mitochondrial genome data shows similar sequence organizations in the entire sequence and in smaller subsequences.

Keywords. Genome mapping; DNA sequence analysis; fractal; chaos game representation; mitochondrial genomes; *Mycobacterium;* nucleotide sequences.

1. Introduction

The advent of the human genome project has led to a manifold increase of information in nucleotide sequence databases. With further progress, a surge of base-sequence information is expected, whose analysis, using present-day methods, may not be possible at the same speed at which it is generated. The task of making the information useful is described as providing the 'informatics' of the project. This is an uphill task, as aptly described by Erickson (1992). Since the information generated will be diverse in nature, analysis would be made easier if the units of information were classified into groups on the basis of their functional and organizational similarities. It has been felt that such analysis would be easier if sequence information could be translated into visual patterns using interactive graphics algorithms.

One of the methods being developed for this purpose is a fractal representation of nucleotide sequences [chaos game representation (CGR)] based on the concept of chaos (Jeffrey 1990). Jeffrey reported patterns generated by different gene sequences of vertebrates as well as invertebrates and raised a series of questions regarding the usefulness of investigating DNA sequences by CGR. Korolev *et al* (1991) used the approach to describe some global patterns of images arising out of genes such as human adenosine deaminase (HUMADAG) and the genome of herpes simplex virus (HSV). Algorithms developed by Solovyev *et al* (1991) can classify sequences on the

*Visiting scholars from Department of Computer Engineering, Birla Institute of Technology. Mesra, Ranchi 835 215.

†Corresponding author.

basis of their possible function in the genome *(e.g.* introns, exons). However, the usefulness of these algorithms can be realized only after megabasepair-long DNA sequences have been classified on the basis of global organizational patterns. Further, the algorithms would have an additional advantage if quantitative information could be obtained about representation of various kinds of subsequences within a sequence of interest.

Visual patterns generated by interactive-graphics algorithms seemed an appropriate approach to analyse long DNA sequences [10 kilobase (kb) to 1 megabase (Mb)]. Keeping this in view, we have addressed the following questions in the present work: (i) Can visual patterns generated by long DNA sequences be used as indicators of homology between two sequences? (ii) Can sequences that are highly repetitive be recognized in such images? (iii) If so, can the repetition frequency be known? (iv) Is it possible to explore the effectiveness of CGR to identify sequences that are rare or absent in a given long sequence? (v) Would images generated by smaller fractions of a long sequence reflect the overall image, or does a fraction of the sequence generate its own characteristic image?

We developed a computer program based on the CGR technique on an IRIS 4-D workstation (Cyber 910 machine). This program, GENOME MAPPING, has novel features that enables one to rapidly analyse megabase-long DNA sequences as a single data file. It not only allows one to visualize global and local patterns in the sequences but also provides quantitative information. In this paper, we describe the versatility of the program. We have used the complete genomic sequences of human, mouse, fin whale and yeast mitochondrial DNA from the EMBL database for illustration of the method. We have also compiled partial genome sequences of *Mycobacterium leprae* and *M. tuberculosis* from the available sequence data, plotted them using GENOME MAPPING, and visualized their organization.

2. Methods

2.1 *Program description*

In GENOME MAPPING the entire sequence is plotted in a square (figure 1) whose corners are marked "A, 'T', 'G' and 'C' following the basic rules of CGR (Jeffrey 1990, 1992), wherein each base in the sequence is a dot occupying one pixel area. The plotting is as follows: the origin is at the centre of the square; the first base is plotted halfway between the origin and the corner corresponding to the base, the second base is plotted midway between the first point and the corner corresponding to the second base, and so on. Figure 2 illustrates the plotting of a five-base sequence, GTCGC. The large square can be subdivided into four quadrants, each quadrant denoting subsequences or strings ending with the base marked at the corner of the quadrant (figure 1). Each of the four quadrants can be further divided into four subquadrants, each subquadrants now describing strings whose last two bases are identical. The step-by-step division of quadrants can be carried out till a quadrant of size one pixel is reached. In other words, if the entire square, with sides of length 1 say, is divided into a number of subquadrants with sides of length 2^{-n}, then each subquadrant will contain points representing all subsequences with the same final n bases. At the present limit of resolution of the

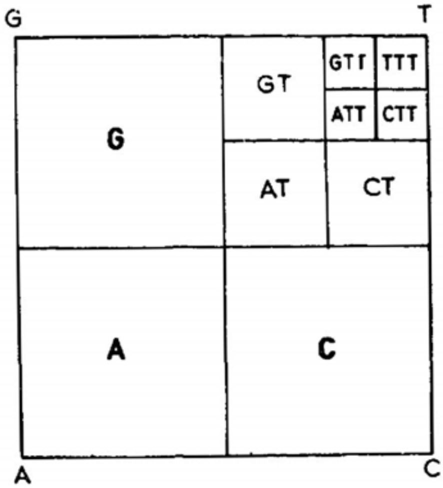


Figure 1. The division of the quadrants in GENOME MAPPING. All bases are plotted within the square, each base (A, T, G or C) going into its respective quadrant. Following the rules of coordinate geometry, the upper right-hand quadrant is referred to as the first quadrant, followed by the upper left-hand quadrant, and so on. Thus in this case the plot is in the TGAC configuration. All the points in a given quadrant (i.e. in one of the T, G, A or C quadrants) of the plot correspond to subsequences with the same last base, *i.e.* all sequence strings ending in T will be plotted in the T quadrant. Further, each subquadrant is defined by a dinucleotide: though all T bases are plotted in the T quadrant, their position in the quadrant will be determined by the base preceding the T, *i.e*., all T bases preceded by A will be plotted in the AT subquadrant, all T bases preceded by C will be plotted in the CT subquadrant, and so on. The figure shows examples of subquadrants defined up to three bases only.

screen, the plot size is 1024×1024 pixels. However, inaccuracies creep in at this resolution in doing quadrant analysis. The next lower plot size is 512×512, or 4^9 pixels, and for our purposes it allows accurate analysis. Here each pixel is defined by a sequence of nine bases. GENOME MAPPING has two versions, *Genfast* and *Gen.* In both versions points are plotted in the same manner as described above, but each version has unique characteristics, which we discuss below.

2.1a *Genfast:* This version of the program plots sequences of a million bases in length in four to five seconds. The characteristics of this version along with examples are described below.

(i) *Configuration:* The sequence to be analysed can be plotted in one of three configurations, CTAG, GTAC and TGAC, as illustrated in figure 3. These configurations have been chosen in such a way that all possible nucleotide strings of the type A_mC_n, A_mT_n, A_mG_n, T_mA_n, T_mC_n, T_mG_n, etc. can be matched along a diagonal. This provides visual impact and allows the user of the program to view the concentration of such strings in a meaningful way. This is illustrated in figure 3 by plots of the 78,521-base-long sequence MISCCG (*Saccharomyces cerevisiae* complete mitochondrial genome) in three different configurations.

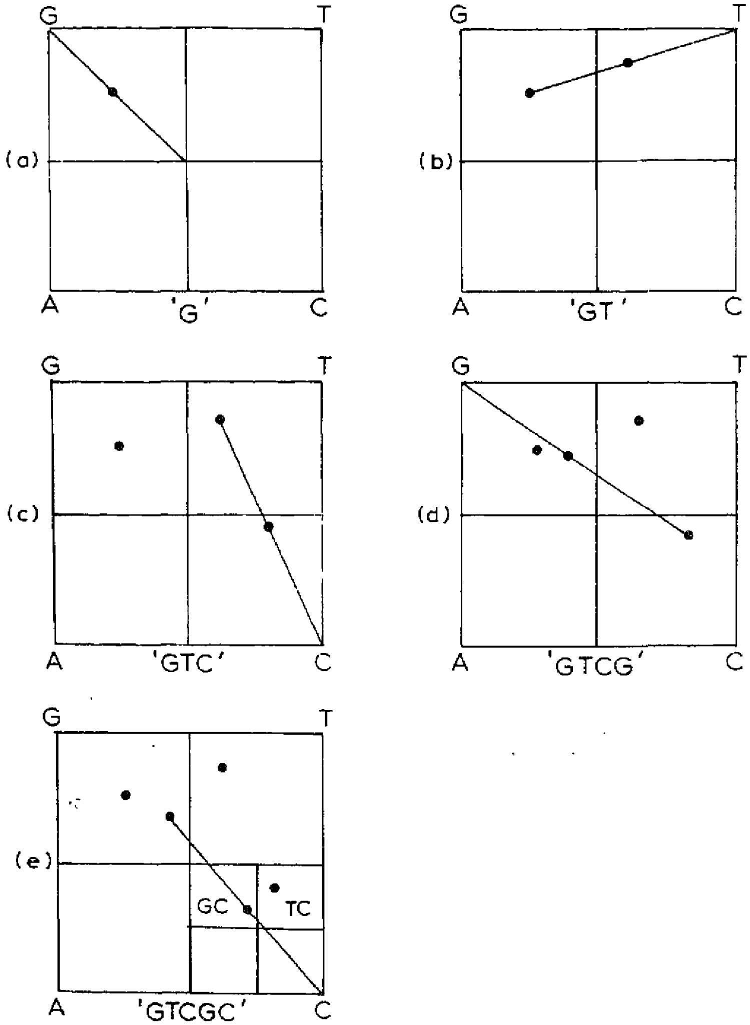
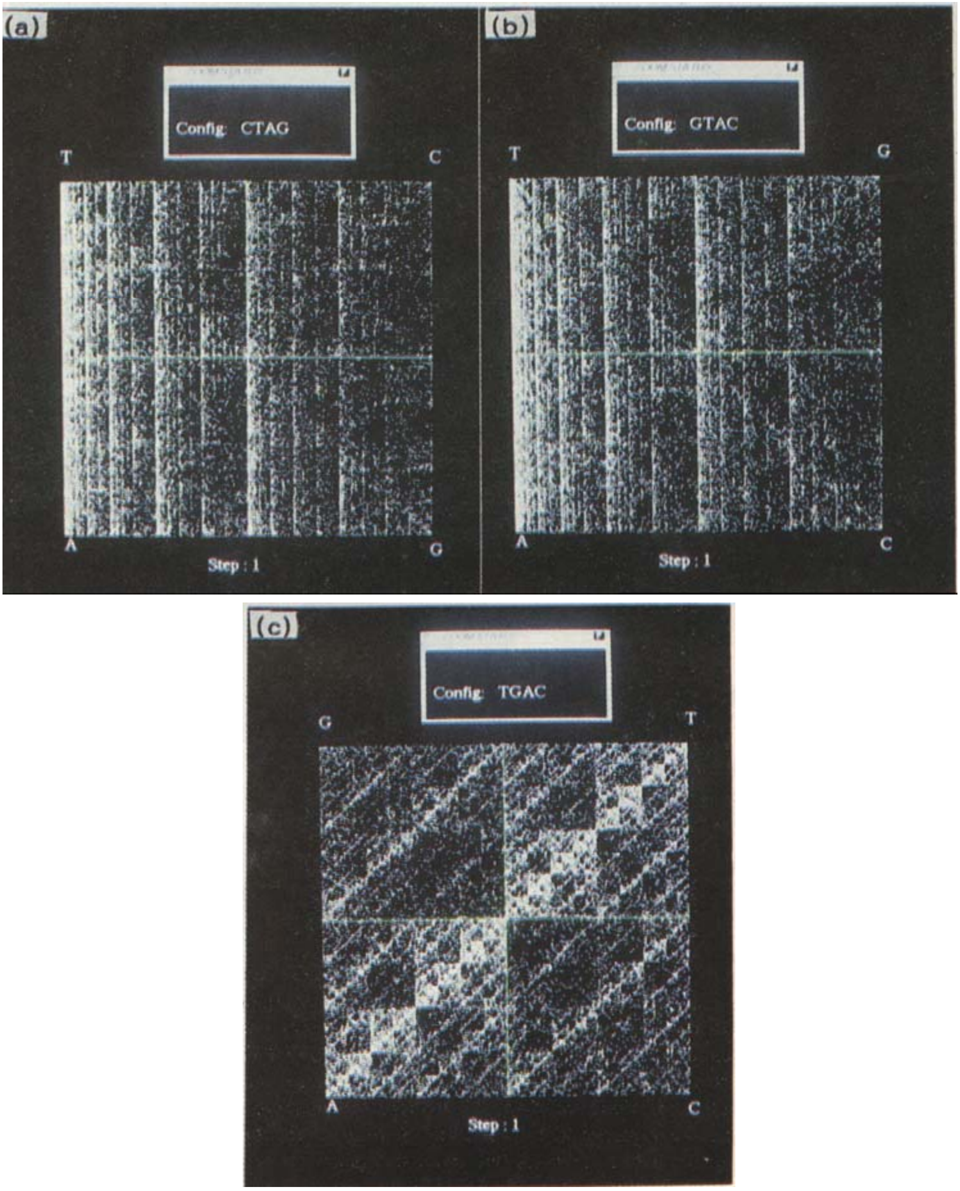


Figure 2. The plotting of a five-base sequence, GTCGC. See text for explanation of the steps (**a** to **e**). Note that each base is plotted in its respective quadrant and that its position within the quadrant is guided by the preceding sequence. For example, of the two points in the C quadrant, one is in the 'TC' subquadrant and the other in the 'GC' subquadrant.

(ii) *Colour code and 'step-up'*: Since the bases in the sequence are plotted as points in their respective quadrants, many points, depending on their string characteristics, fall in the same coordinates. For example, in the case of MISCCG genome, which has approximately 60% of the sequence repetitive, overlapping points would dominate the genome plot. This leads to a significant information loss. To overcome this, a colour code with a 'Step-up' mode (figure 4) was devised to distinguish between unique points and points with overlap in the plot. This is one of the novel features of this program. Figure 4c shows the colour code used. However, by using the basic colour code one cannot distinguish points with overlap of more than seven; this is taken care of by the 'Step-up' function. In the 'Step-up'



Config: CTAG

Step : 1

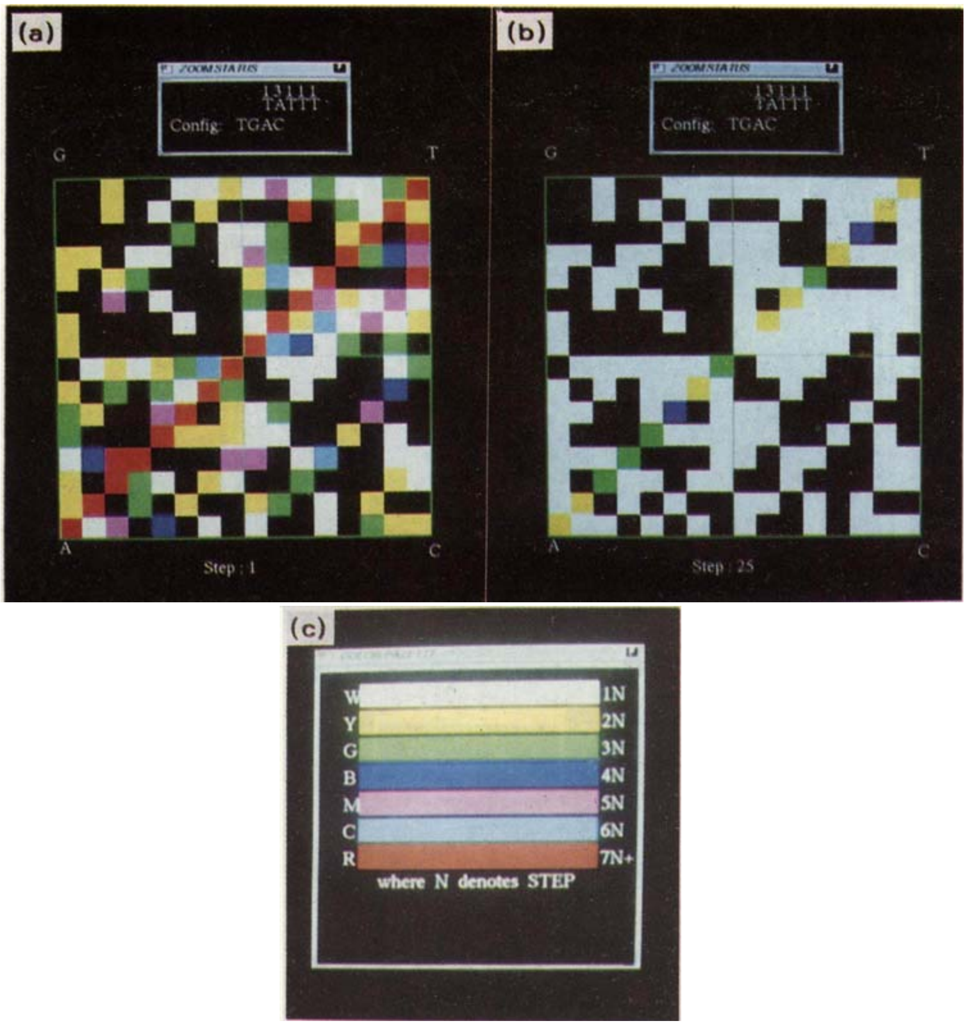
Config: GTAC

Step : 1

Config: TGAC

Step : 1

Figure 3. GENOME MAPPING plots of MISCCG (*Saccharomyees cerevisiae* mitochondrial DNA, 78,521 bases long) in three configurations: **(a)** CTAG, **(b)** GTAC and **(c)** TGAC. The best representation (which generates a pattern characteristic for this genome) is in the TGAC configuration, as this sequence has a high concentration of A_mT_n sequence motifs/repeats, which is best viewed as a diagonal.



Config: TGAC

Step : 1

Config: TGAC

Step : 25

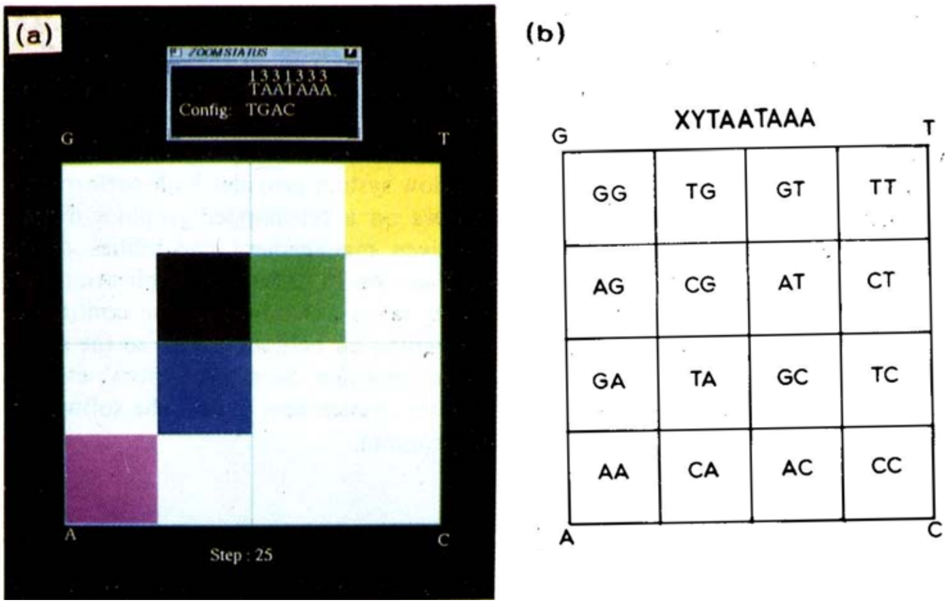
where N denotes STEP

Figure 4. (**a,b**). GENOME MAPPING plots of MISCCG sequence in the TATTT subquadrant in (**a**) 'Step-up' 1 and (**b**) 'Step-up' 25, to exemplify the use of colour code and 'Step-up'. The plot in a signifies that in MISCCG strings ending with TTATTT and ATATTT are represented more than strings ending with GTATTT and CTATTT, and that many such strings are present more than once in the sequence (all coloured points). Further, points (red) along the diagonal are represented more than seven times in this genome. The same sequence when plotted in 'Step-up' 25 (**b**) shows coloured points only along the diagonal, which indicates that these strings are represented more than 25 times in the genome (yellow 26-50, green 51-75, and blue 76-100). (c) The colour code as explained in text, where *N* stands for the step-up mode. For example, in b, *N* = 25.

mode one selects steps ranging from 5 to 2000. Selecting a 'Step-up' of *N* = 5 means that the colours of the basic colour code indicated change by multiples of five; white indicates points arising owing to overlap of 1 to 5 points, yellow 6 to 10, and so on. Figure 4 also illustrates a result of use of 'Stepup'.

As already stated the 'Step-up' mode combined with the colour code can be used to search for highly repeated sequences. In MISCCG, cyan is the highest colour level observed at step 100 and there are only two such points in the whole plot arising due to the occurrence of $(AT)_5$ sequence motifs. Once such points have been identified the question arises as to what sequence strings give rise to these points. The 'Zoom' function provides an answer to this question.

(iii) *'Zoom' function:* Two zoom functions, 'Zoom' and 'Zoom quadrant' have been incorporated, in 'Zoom', an area of 40 × 40 pixels (defined by a string of six) around the cursor is magnified 6·25-fold and displayed on the side screen along with the string defining the area. This function helps the user to get an idea of the general pattern of string representations in the sequence. The 'Zoom quadrant' function helps the user to find the sequence of points up to a string length of nine. This function is basically the converse of the idea of quadrants and subquadrants mentioned earlier. Figure 5a shows the plot of MISCCG sequence after the seventh zoom. In this plot all points (16 in number) end with the string definition TAATAAA. Each of these points can then be defined by strings of nine (figure 5b). The 'Zoom' facility along with the colour code and 'Step-up' mode immediately shows the approximate number of occurrences of these strings in the sequence being analysed.



GG	TG	GT	TT
AG	CG	AT	CT
GA	TA	GC	TC
AA	CA	AC	CC

Figure 5. (a). GENOME MAPPING plot of MISCCG after seventh zoom in the 'Zoom quadrant' function. As indicated in the top of the figure the quadrant zoomed first was the A quadrant. The second zoom was of the A subquadrant of the A quadrant, *i.e.* the square defined by the dinucleotide AA. Thus, after seven zooms, all the points in the plot are defined by the string TAATAAA. This can further be zoomed twice, at the present limit of resolution of the screen. **(b)** Definition of each small quadrant of the plot in a. As mentioned all points in this plot represent strings ending with TAATAAA, and the different small squares define the two bases (Y and X) preceding this string. For example, the yellow square in **a** represents sequences of the type TTTAATAAA, while the blue square is defined by the string TATAATAAA.

2.1b *Gen*

In *Gen* points are plotted in a similar manner, but the position of each base in the sequence along the string is stored in the computer memory. Therefore, the time taken for plotting is longer than with *Genfast,* approximately 90 min for a million-base-long sequence.

A *Gen* map can be plotted in any one of the three configurations mentioned and in any one of the 'Step-ups' described. It also has the same colour code as *Genfast.*

Once the plot is completed, the 'Zoom' functions can be used to analyse points of interest. In this plot the 'Zoom' not only shows strings (three to six bases long) defining the zoomed area but also displays the number of occurrences of a particular string and its position in the loaded sequence. The colour code also shows points with overlap. By analysing the occurrence of the repeated sequences represented by such points, one can tell whether the repeats are tandem or dispersed in nature. While *Genfast* can be used for rapid analyses of sequences, *Gen* is suitable for more detailed analyses.

2.2 *Software description*

The GENOME MAPPING software was implemented on a Silicon Graphics IRIS 4-D series workstation running NEWS (Network Extensible Window System). The minimal computer configuration required is an IRIS 4-D workstation with 8-bit colour plane. The software was written in the C programming language in a UNIX environment. The programming interface was provided by the GL graphics library, which contains both low-level and high-level routines and types written in C for interacting with the window system. The window system provides high-performance graphics to a hierarchy of resizeable windows on a bit-mapped-graphics display device. The software exploits the mouse-driven management capabilities of the window system, based on the ability to 'steal' events. In order that application may respond in real time to every event, the user takes advantage of the continuous polling system rather than queuing, without getting far behind relative to the actual position of the mouse. All the described functions like 'Step-up', 'Zoom', etc. are carried out using the mouse. Figure 6 shows the system flow chart. The software is extendible with few changes in the existing program.

3. Results and discussion

The human genome contains 50,000 to 100,000 genes of which only about 5000 have been sequenced, most of them partially. To date, sequences of approximately 70 million base-pairs of the expected three billion have been generated. At present, the total length of accumulated sequences is doubling every two years. The vast amount of sequence information thus generated will have no meaning until it can be assessed, visualized, tested and studied. As mentioned earlier, this software was developed keeping the human genome sequencing programme in view, so that the immense amount of sequence data expected to be generated could be analysed speedily. The main utility of the program in the present form is, besides its facility for classifying sequences into different groups on the basis of their local

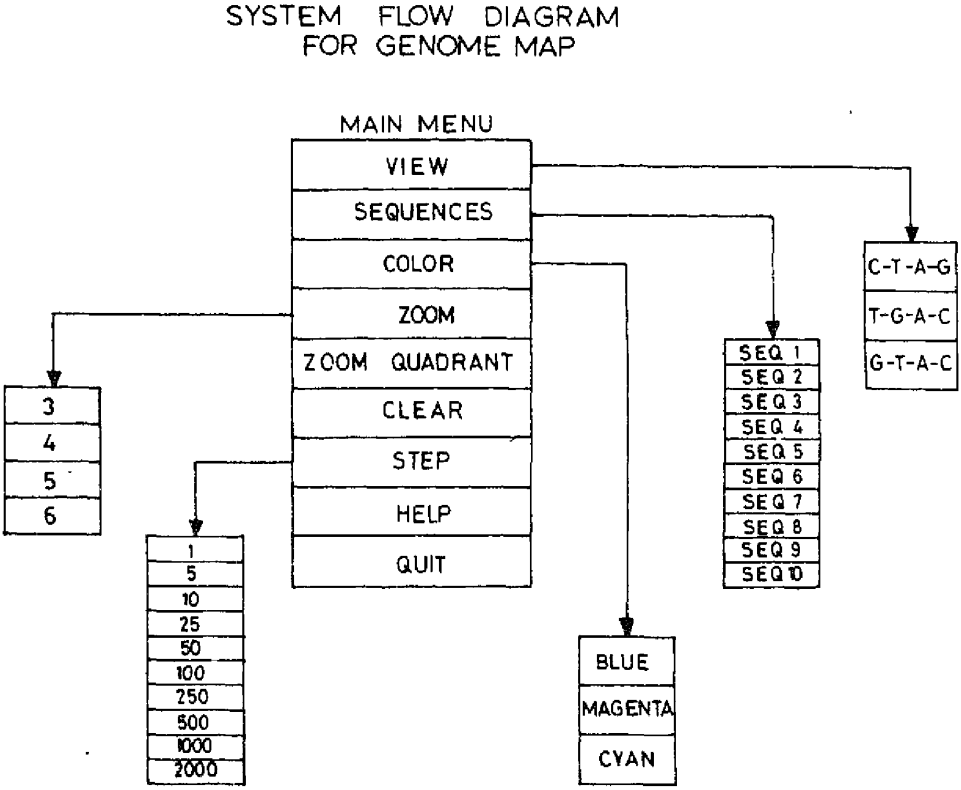


Figure 6. System flow diagram for GENOME MAPPING. TO plot a sequence the configuration of the plot is chosen first with the view menu, followed by the 'Step'. The sequence to be plotted is chosen from the sequences stored in the memory of the program (at a time ten different sequences can be stored). Zooming can be done using the 'Zoom' (arbitrary zooming) or 'Zoom quadrant' menus. In the *Gen* version of the program, while using the 'Zoom' menu one can also select the string size (3 to 6).

and global patterns, is in identification of highly repetitive sequences, as well as sequences that are rare or absent.

In the preceding section, the program was described in detail. In this section we demonstrate its utility in investigating long DNA sequences. In this paper we do not seek to analyse new sequence data but to highlight the versatility of the method using available sequence data that have previously been analysed using other approaches.

3.1 *Genome plots as indicators of homology*

The entire nucleotide sequences of several mammalian mitochondrial DNAs are known. These genomes are similarly organized. Nevertheless, the similarity ranges from 20% to 90% in different regions and large changes in extent of homology occur even in respect of the same coding region (Singer and Berg 1991). For example, in the mouse mitochondrial genome there are regions that show 70–75% homology with the human mitochondrial genome as well as regions that show 20–

30% homology (Bibb *et al* 1981). We used GENOME MAPPING to visualize the global homology between three representative mammalian mitochondrial genomes. Figure 7 shows GENOME MAPPING plots of complete sequences of *Homo sapiens* (MIHSCG, 16,569 bases; figure 7a), *Mus musculus* (MIMM, 16,295 bases; figure 7b) and *Balaenoptera physalus* (MIBPCG, 16,398 bases; figure 7c) mitochondrial DNA taken from the EMBL database. The striking similarity between the three plots in figure 7 provides a visual impression of the overall (giobal) sequence homology between the genomes. There are algorithms available to examine sequence homology between a query sequence and a target sequence, which would give a quantitative analysis of the regions of homology (Wilbur and Lipman 1983). However, these algorithms become computer-intensive with increase in size of either the query or the target sequence. The aim of using the GENOME MAPPING plots is to compare homology between very large sequences in a qualitative fashion. A comparison of complete genome patterns can be obtained by this method. Depending upon global homology a decision can be taken whether computer-intensive but quantitative algorithms are required for estimates of sequence homology.

3.2 *Global sequence organization gives rise to signature patterns in* GENOME MAPPING *plots*

The pattern of dense and sparse triangles in the plots of mammalian mitochondrial genomes (figure 7) arises because the G quadrant as well as the upper diagonal halves of the adjacent quadrants (i.e. the T and A quadrants) are sparsely populated with points (see figure 7d). This global pattern is reflected in progressively smaller subquadrants. For example, the C subquadrant can be diagonally divided into a dense and a sparse region. The sparse region arises owing to rarity of sequences ending with GC, as well as sequences represented in the upper diagonal halves of the subquadrants TC and AC, *i.e.* sequences of the type GTC, GAC, GTAC, etc. On the other hand, the yeast mitochondrial genome (see figure 3c) exhibits a pattern different from that of the mammalian mitochondrial genomes. The high density of dots in the AT diagonal of the plot of the former reflects the presence of A_mT_n repeat motifs in the genome. Looking at the same plot in other configurations, one realizes that whereas T_mC_n sequences are also represented (figure 3b), G_mC_n are represented to a much lesser extent (figure 3c). The presence of dot-rich diagonals along the AT line in the G and C quadrants (figure 3c) also signifies that sequences of the kind $\mathrm{A_mT_nG}$ and $\mathrm{A_mT_nC}$ are highly prevalent in the yeast mitochondrial genome. The organization of the repeats, meaning whether the repetitive strings are randomly dispersed in the genome or are present in clusters, can be found out using the *Gen* version of the program, as this version indicates the positions of occurrence of a query string.

Thus, related sequences are likely to exhibit signature patterns. To prove that the patterns generated by genomic sequences are not an artefact of the methods of analysis, we plotted a 20,000-base-long random sequence (generated using random number generator program). This sequence did not lead to any distinct pattern. However, as we limited the randomization by giving weights to particular bases, patterns began to appear (figure 8). The implication is that a pattern usually

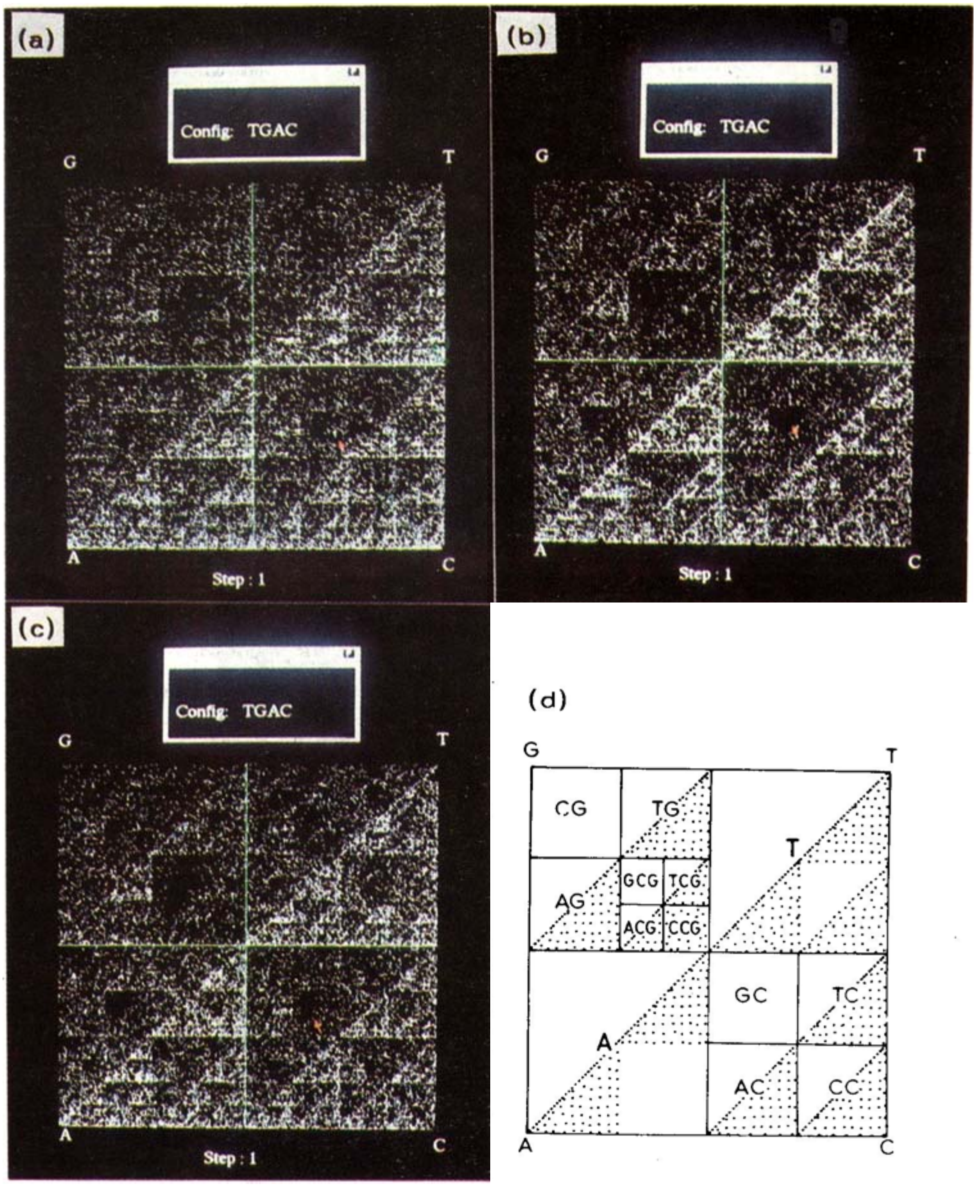
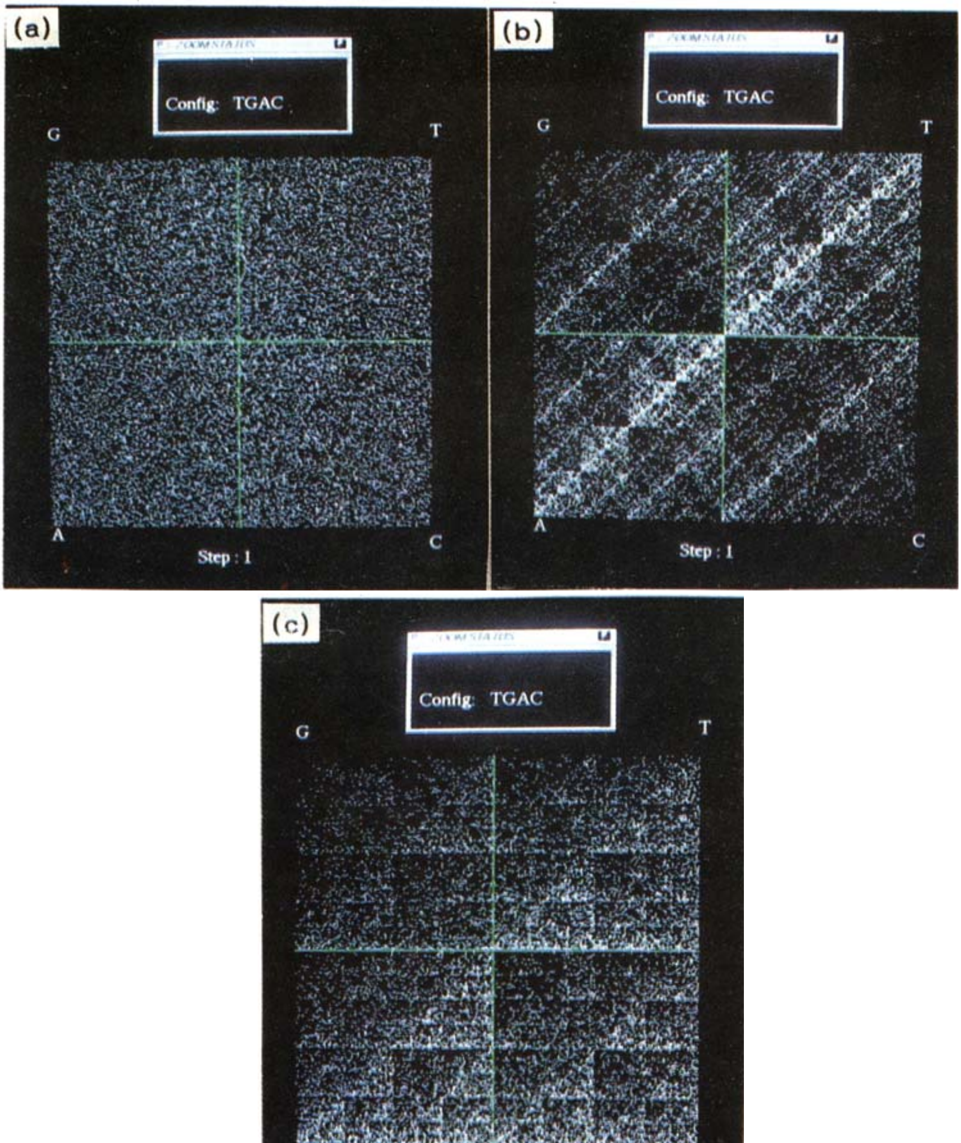


Figure 7. (a-c). GENOME MAPPING plots of mitochondrial DNA sequences of **(a)** man, *Homo sapiens* (MIHSCG, 16,569 bases); **(b)** mouse, *Mus musculus* (MIMM, 16,295 bases); and **(c)** fin whale, *Balaonoptera physalus* (MIBPCG, 16,398 bases). Note the similarity of pattern exhibited by the three mammalian mitochondrial genomes. The red arrow in these plots points to an example of 'holes' that have arisen owing to the rare occurrence of strings ending with CGC, of which GCGC is practically absent. **(d)** Diagram showing the distribution of points in the various subquadrants of the plots in **a, b, c.** Dotted regions represent particular strings that occur with higher frequency in the DNA sequence.



(a) Config: TGAC

Step : 1

(b) Config: TGAC

Step : 1

(c) Config: TGAC

Step : 1

Figure 8. Plots of a 20,000-base-long generated random sequence in which **(a)** all the four bases are present in equal percentage; **(b)** A, T, G and C are in 30, 30, 20 and 20% respectively; and **(c)** A and C comprise 31 % each, T 25% and G 13% of the generated sequence, as found in the MIHSCG sequence. Note that, whereas a random sequence with equal weights to all four bases does not lead to any distinct pattern **(a),** limiting the randomization by skewing the base composition leads to generation of certain patterns **(b** and **c)** with some similarity to the plots of real genomic sequences. However, absence of particular subsequences and presence of certain distinct repeat motifs are not seen in the plots of randomly generated sequences.

emerges if the sequence being plotted is skewed in base composition and is rich in repetitive sequences or shows rare occurrence of certain sequence motifs. Once a series of such patterns are generated from a database, any new sequence can easily be classified after matching with the library of patterns.

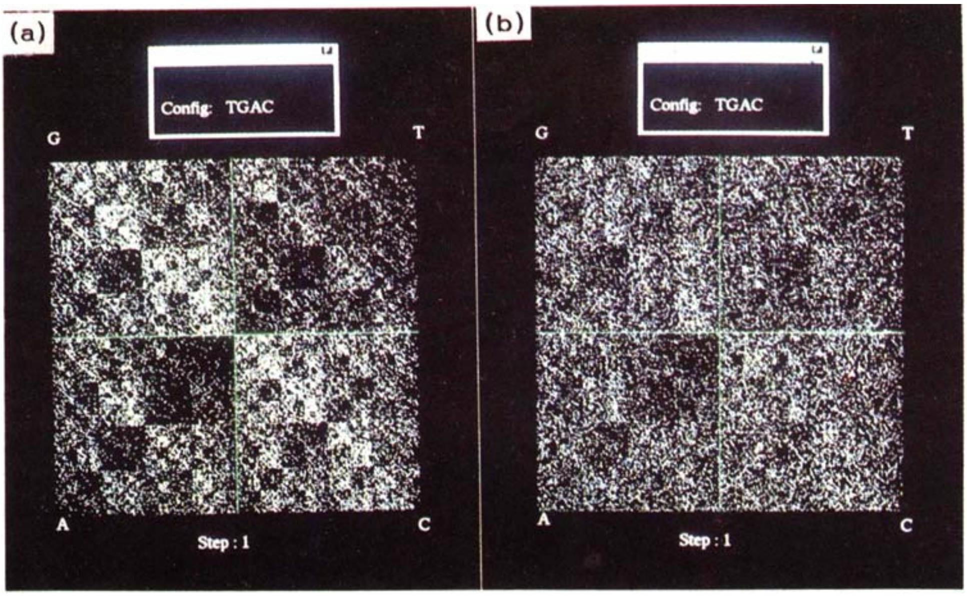
3.3 *Sequences absent or under-represented are viewed as 'holes' in the plot*

A striking feature of the plots of the mammalian mitochondrial genomes (figure 7) is the complete absence of points in certain regions, which we have termed 'holes'. These 'holes' represent strings or subsequences that are absent in the plotted sequence. In the example, the hole in the G quadrant is mainly due to rare occurrence of sequence strings of the type NGCG, where N is any of the four bases (*i.e.* CGCG, TGCG, AGCG and GGCG). This implies that in the C quadrant the GC subquadrant would be sparsely populated with points as this subquadrant represents a subset of the NGCG type of string. This is also clear in figure 7. Further, of the four possible smaller subquadrants in GC (defined by triplets), sequences ending with CGC are rare and those ending with GCGC are practically absent. These plots thus allow identification of strings of sequences that are rare or absent. To our knowledge there is at present no other algorithm available if one wants to search for absence of certain sequences in a given long sequence (although algorithms to search for repetitive sequences are available).

3.4 *GENOME MAPPING plots of Mycobacterium sequences*

As we mentioned at the beginning of this section, this software will be of great help in analysing long stretches of new sequence data. As an example, we have compiled the available sequence data of *Mycobacterium tuberculosis* and *M. leprae* (figure 9). The EMBL database has 23 entries for *M. tuberculosis* and 22 for *M. leprae.* All sequence entries for each species were linearly arranged in a single file, 31,153 bases in the case of *M. tuberculosis* and 26,698 bases in the case of *M. leprae.* While compiling the available sequences into a single file, we took care to see that sequences of the same genes were not duplicated. The plots in figure 9 clearly highlight the differences in sequence organization between the two genomes. *M. leprae* has a GC content of 55·8% while *M. tuberculosis* has a GC content of 63% (Imaeda *et al* 1982). This difference is reflected in the respective genome plots (figure 9). In the plot of the *M. tuberculosis* sequence, the A and T quadrants are more sparsely populated than in the plot of the *M. leprae* sequence. Further, owing to the higher GC content of the *M. tuberculosis* genome, its plot (figure 9a) shows a very distinct pattern, whereas the *M, leprae* plot (figure 9b) is rather diffuse. Further, sequences of the kind A_m T_n are under-represented in both genomes (note the blank diagonals along AT in figure 9).

At this point one might wonder whether these plots are a faithful representation of the whole *Mycobacterium* genome because less than one per cent of the entire sequence has been plotted in each case. In other words, will these patterns be retained when more of the genome is sequenced and plotted on these maps?



Config: TGAC

Step : 1

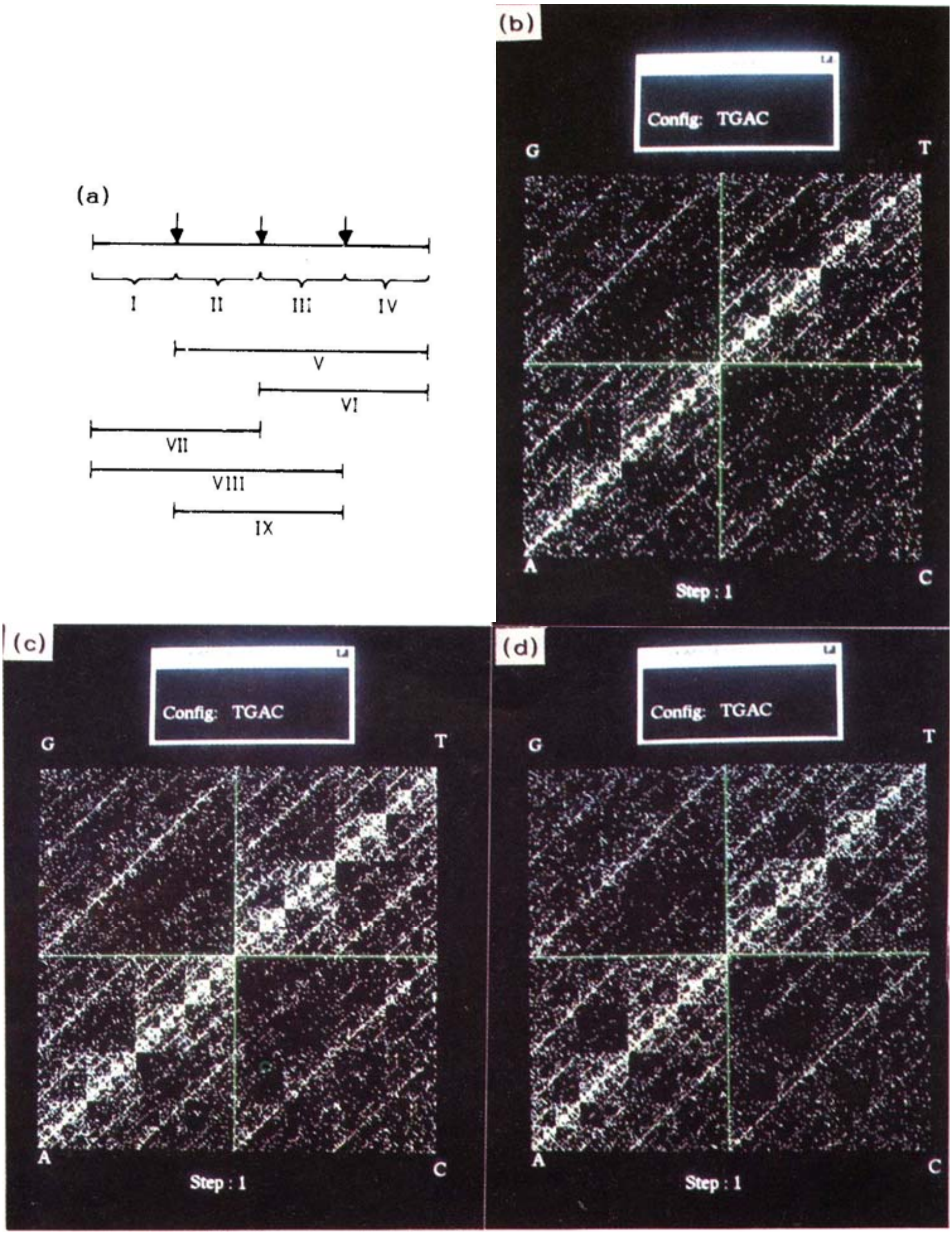
Config: TGAC

Step : 1

Figure 9. GENOME MAPPING plots of sequences of **(a)** *M. tuberculosis* (31,153 bases) and **(b)** *M. leprae* (26,698 bases) compiled from the existing EMBL database.

3.5 *Do patterns exhibited by a small percentage of the genome reflect the global or overall pattern of the genome?*

To address this problem we subdivided the yeast mitochondrial genome (MISCCG) arbitrarily into four smaller units ranging in size from 17,000 to 20,000 bases, and plotted them on GENOME MAPPING plots (figure 10). Interestingly, we found that the basic pattern exhibited by the complete genome is also seen in plots of these units. Figure 10b shows the plot of the last 18,493 bases of MISCCG (fragment IV in figure 10a), and figure 10c the plot of fragments III and IV combined (36,472 bases). We also compiled bits of sequences picked up randomly from MISCCG genome into one sequence and plotted it (figure 10d). Here too the basic pattern of the MISCCG plot is present. This plot is similar in exercise to that of the compiled sequences of the *Mycobacterium* genome in figure 9. From this exercise we conclude that once a pattern has set in even from a small percentage of the genome, it will be preserved upon plotting more and more sequences of that genome. In other words, the sequence organization of the yeast mitochondrial genome is similar at several vastly different length scales. This feature, suggestive of (fractal-like) long-range order in sequence, is becoming increasingly apparent from published analyses of a number of genomic DNA sequences (Peng *et al* 1992). At this juncture we can only speculate that the patterns seen in the plots of the partial *Mycobacterium* genomes may indeed be a true representation of the whole genome. We point out here that sequence similarities at different scales are exactly what one would expect if at least some of genomic evolution were to occur through gene duplication (Ohno 1970).



(b) Config: TGAC
Step : 1

(c) Config: TGAC
Step : 1

(d) Config: TGAC
Step : 1

Figure 10. **(a)** Diagram showing subdivision of MISCCG into four parts: I, 20,872; II, 20,951; III, 17,979; and IV, 18,493 bases. The plots of these smaller fragments exhibited the basic pattern of the whole MISCCG. **(b)** Plot of sequence IV. **(c)** Plot of sequence VI (subdivisions III + IV). As is clear the plots of sequences by addition of these smaller fragments e.g., sequence V to IX only reinforces the pattern. **(d)** Plot of a 20,243-base-long sequence generated by randomly taking bits of sequences from MISCCG and compiling them into a single sequence. All these plots reflect the global pattern of MISCCG (see figure 3).

4. Future prospects

4.1 *Searching for new repeats*

With appropriate modifications to the present version of the program, the plots can be used to identify unknown repetitive sequence motifs in a given genomic sequence by masking known repeats. This could be achieved through pixel-by-pixel subtraction of one plot from another.

4.2 *Gene/sequence duplications*

GENOME MAPPING can be used as a good indicator of gene duplication. If a long stretch of sequence is duplicated it would give rise to a large number of yellow points (in step 1) in the plot, as a large number of nonanucleotide sequences, *i.e.* sequence strings with the last nine bases set by the positions of the points (pixels) that represent them, would be present at least twice. The probability of a large number of nonanucleotide sequences occurring twice in a random sequence is extremely small. A large number could, however, occur if a sequence (say a gene) is duplicated *in toto*.

4.3 *Analysis of chromosomal regions using* GENOME MAPPING *plots*

All eukaryotic chromosomes show specific banding patterns with various banding techniques (*e.g.* G-banding). These patterns represent the basic organization of the band-positive or band-negative regions in the chromosomes. Similarly, the polytene chromosomes of the fruit fly *Drosophila* show a band and interband organization. We are attempting to identify GENOME MAPPING patterns of band and interband regions of polytene chromosomes to understand sequence organization in these regions.

5. Conclusions

Our approach of graphical representation of genomic sequences based on CGR is likely to have a number of varied applications. As we have demonstrated in this paper, GENOME MAPPING plots can be used to look at global homology between two given sequences, as well as for identification of sequence motifs that are either repetitive or rare within genomes. In the case of repetitive sequences, using the *Gen* version of the program one can tell whether they are clustered or dispersed. Further, patterns generated by genomic sequences including partial sequences, such as those of the *Mycobacterium* genomes described here, show interesting characteristics. We are investigating examples of other applications as mentioned.

Acknowledgements

The GENOME MAPPING program was developed on an IRIS workstation, at the Supercomputer Education and Research Centre, Indian Institute of Science, Bangalore. We thank Dr N Balakrishnan of SERC for providing computational facilities, 'Distributed Information Centre', IISc, for help in compilation of DNA sequences from the database, and Dr Dhananjay Bhattacharya for the random number generator software. We also thank Drs V Nanjundiah, N V Joshi and Vani Brahmachari for discussion and critical comments on the manuscript. We gratefully acknowledge financial support from the Department of Biotechnology, New Delhi, through a project on Genome Analysis to one of us (SKB).

References

Bibb M J, Van Etten R A, Wright C T, Walberg M W and Clayton D A 1981 Sequence and gene organisation of mouse mitochondrial DNA; *Cell* **26** 167–180

Erickson D 1992 Hacking the genome; *Sci Am.* **266** 98–105

Jeffrey H J 1990 Chaos game representation of gene structure; *Nucleic Acids Res.* **18** 2163–2170

Jeffrey H J 1992 Chaos game visualisation of sequences; *Comput. and Graphics* **16** 25–33

Imaeda T, Kirchheimer W F and Barksdale L 1982 DNA isolated from *Mycobacterium leprae:* Genome size, base ratio and homology with other related bacteria as determined by optical DNA-DNA reassociation; *J. Bacteriol.* **150** 414–417

Korolev S K, Lim H A and Tumanyan V G 1991 Interactive graphics program of chaos game representation of DNA sequences; *Report FSUSCRI-91-02* (Supercomputer Computations Research Institute, Tallahassee, Florida)

Ohno S 1970 *Evolution by gene duplication* (Berlin: Springer)

Peng C K, Buldyrev S V, Goldberger A L, Havlin S, Sciortino F, Simons N and Stanley H E 1992 Long range correlations in nucleotide sequences; *Nature* (*London*) **356** 168–170

Singer M and Berg P 1991 *Genes and genomes* (London: Blackwell)

Solovyev V V, Korolev S V and Lim H A 1991 A new approach for the classification of functional regions of DNA sequences based on fractal representation; *Report FSU-SCRI-91-40* (Supercomputer Computations Research Institute, Tallahassee, Florida)

Wilbur W J and Lipman D J 1983 Rapid similarity searches of nucleic acid and protein data banks; *Proc. Natl. Acad. Sci. USA* **80** 726–730