Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences

Nick Goldman
Laboratory of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK

Received December 7, 1992; Revised and Accepted March 25, 1993

ABSTRACT

The chaos game representation (CGR) is a scatter plot derived from a DNA sequence, with each point of the plot corresponding to one base of the sequence. If the DNA sequence were a random collection of bases, the CGR would be a uniformly filled square; conversely, any patterns visible in the CGR represent some pattern (information) in the DNA sequence. In this paper, patterns previously observed in a variety of DNA sequences are explained solely in terms of nucleotide, dinucleotide and trinucleotide frequencies.

INTRODUCTION

The chaos game representation (CGR) has been proposed as a novel way of revealing patterns within nucleotide sequences [1]. The CGR consists of a square scatter plot, with each corner of the plot representing one of the bases A, C, G and T (U). One point is plotted for each site of the sequence, the first point plotted halfway between the centre of the square and the corner corresponding to the first nucleotide of the sequence and successive points plotted halfway between the previous point and the corner corresponding to the base of each successive sequence site [1].

If successive bases are chosen at random and with equal probabilities, the square eventually becomes uniformly filled with dots. With unequal probabilities, a pattern of horizontal and vertical bands appears [1, 2]. Jeffrey [1] has investigated the patterns seen in CGRs when DNA sequences are used to provide the sequence of bases A, C, G and T. The CGR for one sequence he studied (human beta globin region, EMBL databank entry HSHBB, 73326 base pairs) is shown in Figure 1. Recent work has looked at ways of classifying and comparing CGRs of DNA sequences [3].

For CGRs to be a useful tool for investigating DNA sequences, it would be necessary to understand the patterns they exhibit and to be able to interpret them in a biologically meaningful way. In this paper, I describe simple features of DNA sequences that give rise to previously-described patterns [1] in CGRs. These are verified by the computer simulation of DNA sequences, following simple rules readily derived from the original sequences. These results indicate that it is unlikely that CGRs can be more useful than simple evaluation of nucleotide, dinucleotide and trinucleotide frequencies.

Biological meaning of the CGR

The point in a CGR corresponding to one base of a sequence is plotted in the quadrant of the square labelled with that base. This is because each quadrant comprises all points that are halfway between one corner and any other point within the square. Conversely, all points plotted within a quadrant must correspond to subsequences of the DNA sequence that end with the base labelling the corner of that quadrant [1]. For example, any base G gives rise to a point in the G (upper-right) quadrant of the square; and every point in that quadrant corresponds to a base G in the DNA sequence. This correspondence between points and subsequences continues recursively to sub-quadrants, sub-sub-quadrants, etc. [4]: the dinucleotide subsequence AG gives rise to a point in the A (lower-left) sub-quadrant of the G quadrant, the trinucleotide TAG gives a point in the T (lower-right) sub-sub-quadrant of the A sub-quadrant of the G quadrant (Figure 2) and so on. By identifying regions of the CGR square in this way, it is possible to identify features of DNA sequences that correspond to patterns of the CGR.

The most obvious feature of the CGR of the human beta globin region sequence (Figure 1) is the repeated (self-similar) pattern of sparse 'double scoop' shaped regions, the largest of which is at the top of the G quadrant. The major part of this region is the upper-left sub-quadrant, corresponding to CG dinucleotides (see Figure 2 and [1]): in other words, a relative rarity of CG dinucleotides is indicated by the sparsely filled CG sub-quadrant.

Jeffrey [1] commented that to understand fully the 'double scoop' pattern, it is necessary to characterize mathematically the rare sequences (oligonucleotides) that produce the shape. This is in fact easily done [2, 4]. The sparse CG sub-quadrant indicates the rarity of CG dinucleotides; this in turn means that there are few trinucleotides containing the dinucleotide CG—the ACG, CCG, GCG, TCG, CGA, CGC, CGG and CGT sub-sub-quadrants will be equally sparsely filled. The first four of these

* To whom correspondence should be addressed

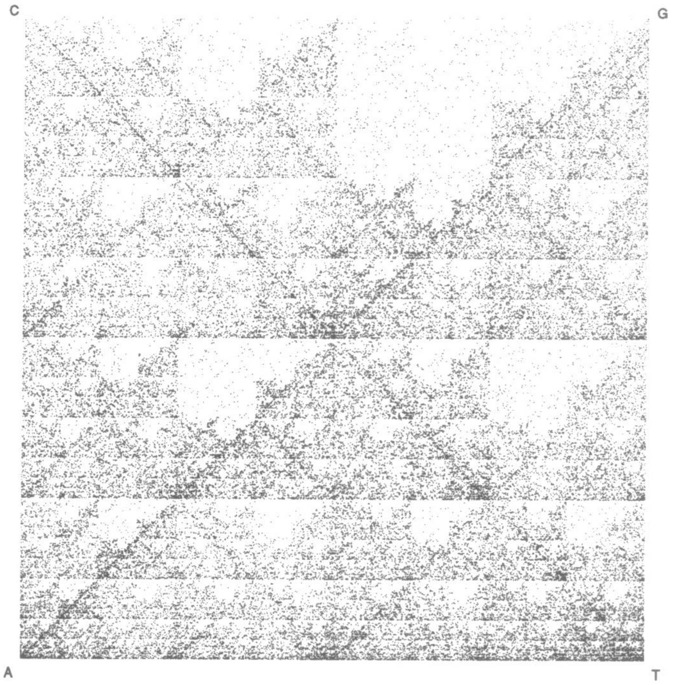

Figure 1. CGR of the human beta globin region sequence (HSHBB; 73326 bp).

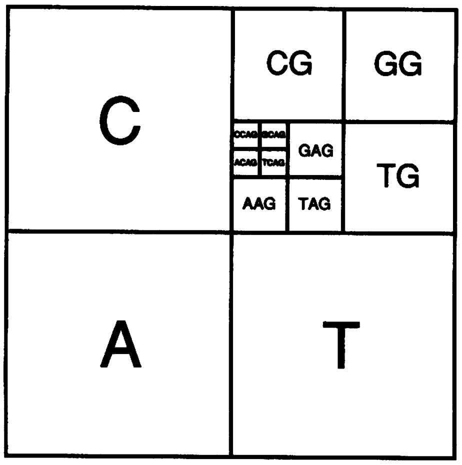

Figure 2. Explanation of the correspondence between oligonucleotides and areas of the CGR of DNA sequences. Each base gives a point in the quadrant labelled with that base; the sub-quadrant is determined by the preceding base, the sub-sub-quadrant (e.g. 'TAG') by the base preceding that, etc. This correspondence continues recursively, but is shown to a maximum of four bases in this figure for clarity.

(XCG) are all within the CG sub-quadrant, and so are already accounted for. The other four of these sub-sub-quadrants (CGX) are distinct from the CG sub-quadrant and can be seen in Figure 1 to be sparsely filled. Similarly, the 16 sub-sub-sub-quadrants corresponding to subsequences CGXY will be sparsely filled, as will the (increasingly small) 64 CGXYZ regions, 256 CGXZYW regions, etc. (where X, Y, Z and W each represent any of the bases A, C, G, T). Figure 3 shows how these regions combine to form precisely the pattern of repeated 'double scoops' evident in the beta globin region CGR.

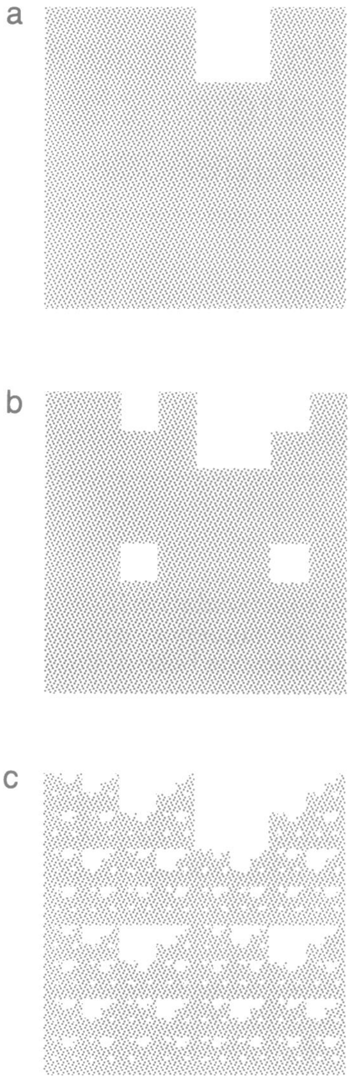

Figure 3. Explanation of the 'double scoop' pattern as a consequence of CG-dinucleotide rarity. **a:** The CGR square with the CG sub-quadrant left unfilled, corresponding to absence of CG dinucleotides. **b:** The CGA, CGC, CGG and CGT sub-sub-quadrants are also unfilled as a consequence of CG rarity. **c:** All oligonucleotides containing CG will be absent, leading to increasingly-many, ever-smaller, regions being blank. This figure shows the CGR square with regions corresponding to all subsequences up to length 5 containing CG unfilled. The pattern of repeats and complex outline of the 'double scoop' appear as in Figure 1.

Also evident in Figure 1 is a division of the square into horizontal and vertical bands. This is explained by the unequal frequencies of the bases A, C, G and T in the beta globin region sequence. As illustrated in Figure 4 of [1] and Figure 1 of [2], if the selection of the next corner (base of the sequence) is at random but with different corners assigned different probabilities of selection, this banding appears as a consequence of the corner regions of each quadrant, sub-quadrant, etc. being more densely filled if their labels are selected with high probability.

A simple model which permits the simulation of these features of DNA sequences is the four state, discrete time Markov Chain [5]. In this model, a 4×4 matrix **P** defines the probabilities with which subsequent bases follow the current base in a DNA sequence. If the base labels A, C, G and T are equated with the numbers 1, 2, 3 and 4, then $\mathbf{P}_{ij}$, the *j*th element of the *i*th row of **P**, defines the probability that base *j* follows base *i*. The row-

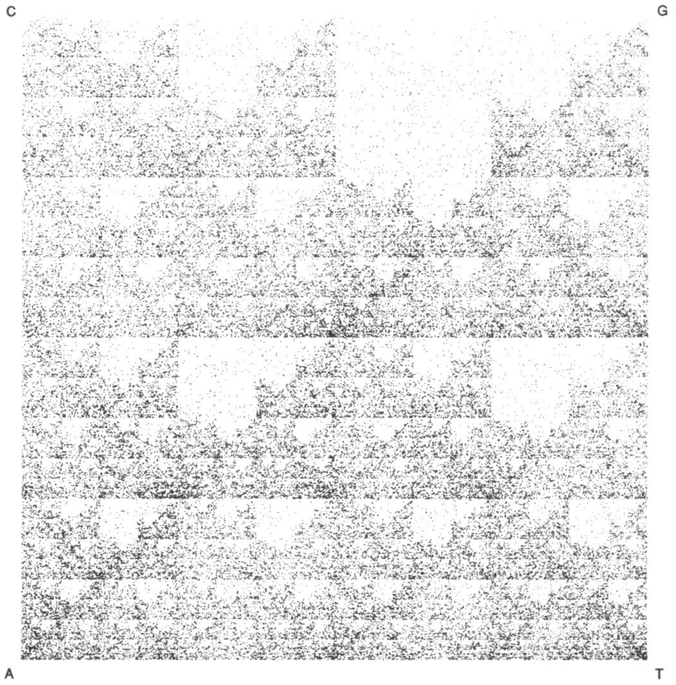

Figure 4. CGR of the first-order Markov Chain simulation of the human beta globin region sequence.

Table 1a. Numbers of dinucleotide occurrences in the human beta globin region sequence

		Second base				
		A	C	G	T	totals
	A	7239	3658	5227	5945	22069
First	C	5166	3293	502	5207	14168
base	G	4580	2839	3676	3694	14789
	T	5087	4379	5383	7444	22293

Table 1b. Probability matrix of the first-order Markov Chain model for the human beta globin region sequence

		Second base			
		A	C	G	T
	A	0.328	0.166	0.237	0.269
First	C	0.365	0.232	0.035	0.367
base	G	0.310	0.192	0.249	0.250
	T	0.228	0.196	0.241	0.334

$P_{XY} = n_{XY}/(n_{XA}+n_{XC}+n_{XG}+n_{XT})$, where the n_{XY} are given in Table 1a.

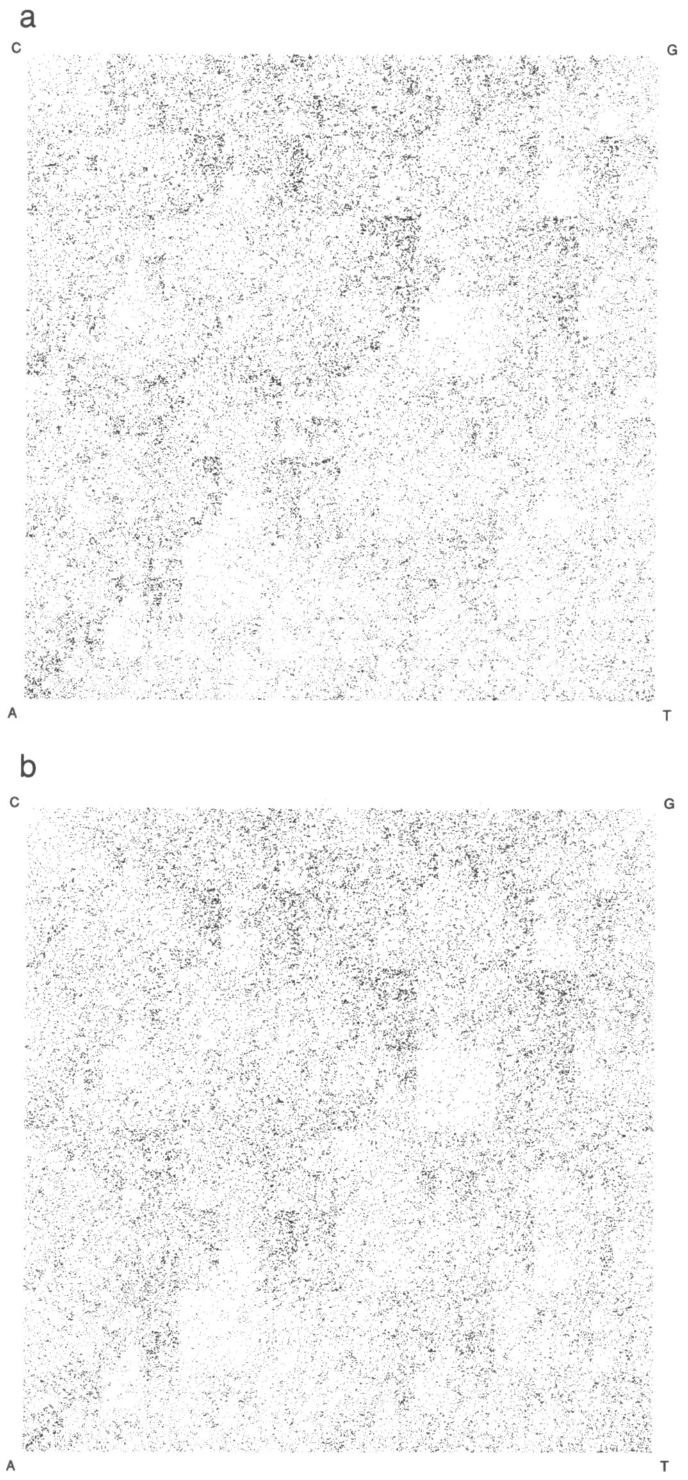

Figure 5. CGRs of the bacteriophage lambda genome. **a:** Original sequence (LAMBDA; 48502 bp). **b:** Second-order Markov Chain simulation.

sums of **P** must equal 1. Using this matrix, a simulated DNA sequence may be obtained by selecting a first base randomly, according the frequencies of the bases in the DNA sequence under study. If this is base i, then the probabilities $\mathbf{P}_{i1}$, $\mathbf{P}_{i2}$, $\mathbf{P}_{i3}$ and $\mathbf{P}_{i4}$ are used to select the next base, and so on until the simulated sequence is of the same length as the original DNA sequence.

This *first-order Markov Chain* model [6], in which successive bases in a sequence depend only on the preceding base, has been successfully used to describe human [7] and other vertebrate DNA sequences [8]. The probabilities in the matrix **P** may be estimated by direct calculation from the sequence's dinucleotide frequencies. If the dinucleotide XY is observed n_{XY} times in the sequence, then probability $\mathbf{P}_{XY}$ is estimated by $n_{XY}/(n_{XA}+n_{XC}+n_{XG}+n_{XT})$. This permits a DNA sequence to be simulated with both individual base frequencies and dinucleotide frequencies matching those of the original sequence. Dinucleotide frequencies (n_{XY}) and the Markov Chain probabilities ($\mathbf{P}_{XY}$) for the beta globin region sequence are given in Table 1.

Table 2. Probability matrix of the second-order Markov Chain model for the bacteriophage lambda genome

		Second base					
		A	C	G	T		
First base	A	0.340	0.260	0.251	0.201	A	Third base
		0.231	0.264	0.291	0.232	C	
		0.202	0.280	0.240	0.299	G	
		0.227	0.196	0.218	0.267	T	
	C	0.217	0.270	0.202	0.113	A	
		0.181	0.165	0.258	0.188	C	
		0.352	0.354	0.309	0.461	G	
		0.250	0.210	0.231	0.238	T	
	G	0.321	0.281	0.267	0.195	A	
		0.201	0.226	0.302	0.211	C	
		0.196	0.257	0.196	0.322	G	
		0.281	0.237	0.234	0.272	T	
	T	0.318	0.320	0.288	0.201	A	
		0.222	0.220	0.278	0.252	C	
		0.100	0.217	0.246	0.219	G	
		0.359	0.243	0.188	0.328	T	

The four values in each cell of the table are for the third bases A, C, G, T respectively.

The CGR of Figure 1 also shows an increased density of points on its A-G and C-T diagonals. These points are caused by runs of sites containing just the bases that label the opposite ends of the diagonals: runs of As and Gs give points near the A-G diagonal, runs of Cs and Ts give points near the C-T diagonal. Less visible in Figure 1, but easily seen if the CGR is drawn with the vertices reordered A, C, T and G, is a similar increased density on the A-T line. It appears that the human beta globin region sequence, as well as having few CG dinucleotides, has an abundance of runs of As and Gs, Cs and Ts, and As and Ts. This is also modelled by the first-order Markov Chain, where an abundance of runs of (for example) As and Gs is indicated by raised probabilities of AA, AG, GG and GA dinucleotides (Table 1b).

A sequence of length 73326 bp was simulated using the first-order Markov Chain model of Table 1b, and its CGR is shown in Figure 4. Notice that this CGR displays all the major features evident in Figure 1, the CGR of the human beta globin region sequence: the repeated 'double scoop' (due to rarity of CG dinucleotides), vertical and horizontal banding (unequal base frequencies) and denser filling of the A-G and C-T diagonals (relative abundance of runs of As and Gs, and Cs and Ts).

Further examples

The first-order Markov Chain model successfully recreates other CGR patterns that have been of interest [1], including those of plants and slime molds, exemplified by the *D. discoideum* myosin heavy chain gene, and of human viruses, exemplified by the human T-cell lymphotropic virus (type III) genome (not shown). I give just one further, more-complex, example here: the CGR for the bacteriophage lambda genome (EMBL entry LAMBDA, 48502 bp), as shown in Figure 5a. An incomplete version of this sequence was studied in [1].

There is little evidence of horizontal or vertical banding in this CGR, but some evidence of sparse regions. The lack of banding suggests approximate equality of the frequencies of the bases A, C, G, T, confirmed by direct calculation from the sequence. Unlike the human beta globin region CGR, where the largest sparse area was one-sixteenth of the square and represented a dinucleotide, the two largest sparse areas of the phage lambda CGR are each 1/64 of the square and represent the trinucleotides CTA and TAG. Other smaller sparse regions appear in the sub-quadrants representing sequences CTAX and TAGX, as expected. In this case, the first-order Markov Chain model will not give the observed patterns, but a more complex *second-order Markov Chain*, in which each base depends on the previous two, does. Second-order Markov Chains have been used to describe both coding and non-coding DNA sequences [9]. $\mathbf{P}_{XYZ}$, the probability that base Z follows the dinucleotide XY, is estimated directly from the DNA sequence trinucleotide frequencies n_{XYZ} using the formula $\mathbf{P}_{XYZ} = n_{XYZ}/(n_{XYA}+n_{XYC}+n_{XYG}+n_{XYT})$. The 64 values $\mathbf{P}_{XYZ}$ for the phage lambda sequence are given in Table 2; the CGR of a sequence of 48502 bp simulated using this model is shown in Figure 5b. This CGR shows all the important features of Figure 5a, the CGR of the original phage lambda sequence.

DISCUSSION

As previously noted [1], a DNA sequence consisting of independent, random bases A, C, G and T (in equal proportions) would exhibit a patternless (uniformly filled) CGR. Any structure to the sequence dictates the patterns observed in the CGR; but complex patterns in the CGR do not necessarily require complex patterns in the sequence. In this paper, I have shown that simple Markov Chain models based solely on dinucleotide and trinucleotide frequencies can account for the complex patterns exhibited in CGRs of DNA sequences. The probabilities defining these models can be calculated directly and easily from the raw DNA sequences, without reference to the CGR, implying that the CGR gives no further insight into the structure of the DNA sequence than is given by the dinucleotide and trinucleotide frequencies. This is in contrast to the method of [2], which relies on trial-and-error comparisons between CGRs.

Figure 3 confirms that the repeated 'double scoop' pattern, reported in the CGRs of almost all vertebrate DNA sequences and in the CGR of human T-cell lymphotropic virus (type III) genome and other human viruses [1], is entirely attributable to a scarcity of CG dinucleotides. The biological reasons for this scarcity are well-understood, being the selective disadvantage of

CG dinucleotides which are prone to methylation and subsequent mutation [10, 11]. All the other patterns previously described in CGRs [1] have in this paper been attributed to simple relationships between single base, dinucleotide and trinucleotide frequencies which are readily calculated from DNA sequences without recourse to chaos game representations. Even better explanations might be derived from a trinucleotide model based on measures of codon usage [12]. Unless more complex patterns are found in CGRs, there is no justification for ascribing their patterns to anything other than the effects described in this paper. Indeed, it must be doubtful whether more-complex patterns of information in DNA sequences will be visible in CGRs, since simpler patterns will tend to obscure more complex ones.

ACKNOWLEDGEMENTS

I thank Axel Kowald for helpful discussions on CGRs and on computer graphics, and Gary Stormo for comments on an earlier version of this paper and for suggesting Figure 2.

REFERENCES

1. Jeffrey,H.J. (1990) *Nucl. Acids Res.*, **18**, 2163–2170.
2. Dutta,C. and Das,J. (1992) *J. Mol. Biol.*, **228**, 715–719.
3. Hill,K.A., Schisler,N.J. and Singh,S.M. (1992) *J. Mol. Evol.*, **35**, 261–269.
4. Jeffrey,H.J. (1992) *Comput. & Graphics*, **16**, 25–33.
5. Feller,W. (1968) An Introduction to Probability Theory and its Applications, 3rd ed., Vol. 1. John Wiley & Sons, New York.
6. Almagor,H. (1983) *J. Theor. Biol.*, **104**, 633–645.
7. Bulmer,M. (1987) *Mol. Biol. Evol.*, **4**, 395–405.
8. Avery,P.J. (1987) *J. Mol. Evol.*, **26**, 335–340.
9. Blaisdell,B.E. (1985) *J. Mol. Evol.*, **21**, 278–288.
10. Josse,J., Kaiser,A.A. and Kornberg,A. (1961) *J. Biol. Chem.*, **236**, 864–875.
11. Bird,A.P. (1980) *Nucl. Acids Res.*, **8**, 1499–1504.
12. Grantham,R., Gautier,C., Gouy,M., Mercier,R. and Pavé,A. (1980) *Nucl. Acids Res.*, **8**, r49–r62.