BIOINFORMATICS

Vol. 17 no. 5 2001
Pages 429–437

Analysis of genomic sequences by Chaos Game Representation

Jonas S. Almeida[1, 2, 3], João A. Carriço[1], António Maretzek[1], Peter A. Noble[2] and Madilyn Fletcher[2]

[1]ITQB/Universidade Nova Lisboa, PO Box 127, 2780 Oeiras, Portugal, [2]Belle W. Baruch Institute for Marine Biology and Coastal Research, Marine Science Program and Department of Biological Sciences, University of South Carolina, Columbia, SC 29208, USA and [3]Department of Biometry and Epidemiology, Medical University of South Carolina, PO Box 250551, Charleston, SC 29425, USA

Received on September 20, 2000; revised and accepted on January 5, 2001

ABSTRACT

Motivation: Chaos Game Representation (CGR) is an iterative mapping technique that processes sequences of units, such as nucleotides in a DNA sequence or amino acids in a protein, in order to find the coordinates for their position in a continuous space. This distribution of positions has two properties: it is unique, and the source sequence can be recovered from the coordinates such that distance between positions measures similarity between the corresponding sequences. The possibility of using the latter property to identify succession schemes have been entirely overlooked in previous studies which raises the possibility that CGR may be upgraded from a mere representation technique to a sequence modeling tool.

Results: The distribution of positions in the CGR plane were shown to be a generalization of Markov chain probability tables that accommodates non-integer orders. Therefore, Markov models are particular cases of CGR models rather than the reverse, as currently accepted. In addition, the CGR generalization has both practical (computational efficiency) and fundamental (scale independence) advantages. These results are illustrated by using *Escherichia coli* K-12 as a test data-set, in particular, the genes *thrA*, *thrB* and *thrC* of the threonine operon.

Availability: A web page interface has been created where analyses of arbitrary sequences by CGR can be performed on-line (http://www.itqb.unl.pt:1111/biomat/resources/resources.htm). The test data-set used in this report is also included.

Contact: almeidaj@musc.edu; mfletcher@biol.sc.edu

INTRODUCTION

Genomic sequences are in a constant state of variation due to processes such as transposition, transformation, translocation, recombination and excision (Karlin *et al.*, 1998; Casjens, 1998). Consequently, different strains of the same species may exhibit considerable differences in global sequence due to extensive reordering of coding regions. The resulting fluidity of microbial genomes has to be taken into account when characterizing the global and local properties of sequences. Consequently, the identification of homologous regions has to follow a bottom-up approach, by first matching individual oligonucleotides, which can then be used as markers to align larger sequences (Pearson and Miller, 1992) as implemented by BLAST and FASTA (Altschul *et al.*, 1990; Altschul and Koonin, 1998; Pearson, 1996). In contrast, a top-down approach targets the statistics of nucleotide succession, which are not affected by shuffling of individual sequences. This is achieved by identifying Markov chain models (MCMs), which take the form of probability tables describing the succession scheme (Rabiner and Juang, 1986), and is further extended by hidden Markov models (HMM) for the identification of succession of alternative patterns of nucleotide succession (Krogh, 1998).

Chaos Game Representation (CGR) was proposed as a scale-independent representation for genomic sequences by Jeffrey in 1990. The technique, formally an iterative map, can be traced further back to the foundations of statistical mechanics, in particular to Chaos theory (Bar-Yam, 1997). The original proposition has been considerably expanded and generalized for sequences of arbitrary symbols (Tino, 1999), and therefore including other biological sequences such as proteins (Basu *et al.*, 1997; Pleißner *et al.*, 1997). However, the possibility that the CGR format can be used for representing the nucleotide sequence as well as identifying the resulting sequence scheme has never been fully explored. Three years after the original proposition, the translation of CGR quadrant frequencies into oligonucleotide frequencies was demonstrated and interpreted as an indication that ‘it is unlikely

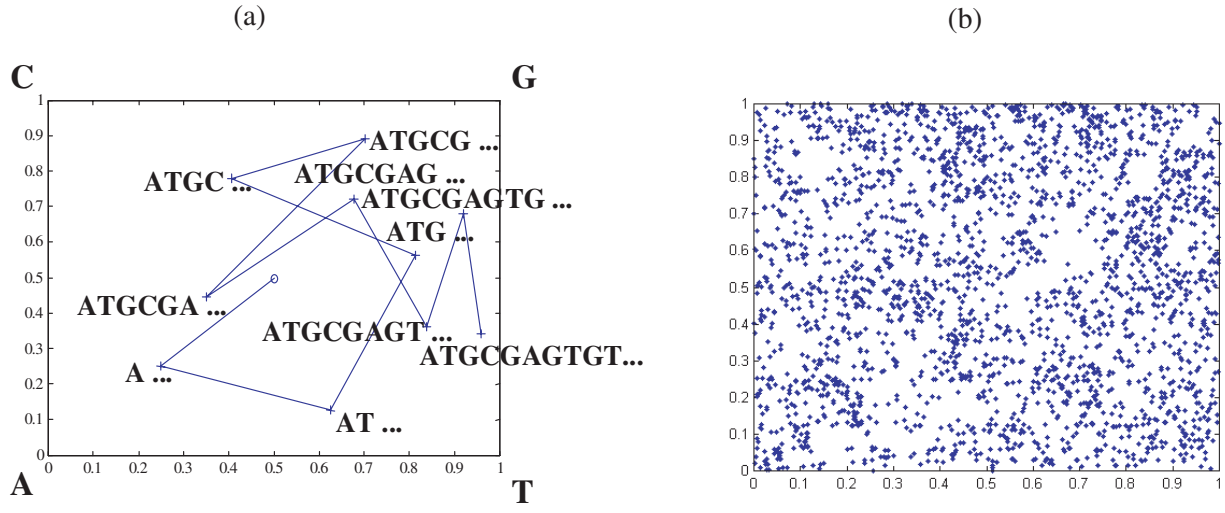


Fig. 1. (a) Chaos Game Representation (CGR) of the first 10 nucleotides of *E.coli* gene *thrA*: `ATGCGAGTGT`. The coordinates for each nucleotide are calculated iteratively using (0.5, 0.5) as an arbitrary starting position (equation 1). The pointer is moved half the distance to the next nucleotide to determine the next position (equation 1). (b) CGR of the full *thrA* sequence, totaling 2463 pairs of bases.

that CGRs can be more useful than simple evaluation of nucleotide, dinucleotide and trinucleotide frequencies' (Goldman, 1993). The subsequent development of HMM further relegated CGR to the status of a representation technique. The results reported below are all the more relevant as the limitations of conventional HMM for functional genomics have been increasingly noted (Baldi and Brunak, 1998). This report shows that CGR is in fact a generalized, scale-independent, Markov probability table. The CGR space is a continuous reference system where all possible sequences of any length have a unique position. Consequently, all possible nucleotide succession schemes will be encoded in the continuous space, and a new generation of scale-independent Markov models accommodating non-integer orders is made possible.

SYSTEM AND METHODS

Computation

The algorithms described in this manuscript were coded using MATLAB^{TM} 5.3 language, licensed by The MathWorks Inc. (http://www.mathworks.com). An internet interface was also developed to make them freely accessible through user-friendly web-pages (see Availability).

Test data

The genome of *E.coli* K-12 MG1655 was analyzed to illustrate CGR properties. The sequence was obtained from the University of Winsconsin *E.coli* Genome Project (http:/www.genetics.wisc.edu). The test sequences corresponding to the threonine genes *thrA* (positions 337–2799), *thrB* (positions 2801–3733) and *thrC* (3734–5020) correspond to the second, third and fourth open reading frames (ORFs), respectively.

ALGORITHM AND IMPLEMENTATION

CGR of *E.coli* thrA

The properties of the CGR procedure (see also Jeffrey, 1990; Burma *et al.*, 1992; Deschavanne *et al.*, 1999) are hereby illustrated by analyzing the sequence of *E.coli* threonine gene *thrA*. The CGR space generated by genomic sequences is planar and is confined by the four possible nucleotides as vertices of a binary square (Figure 1). The CGR positions, CGR_i, of each nucleotide, g_i, of a sequence, g of length n_G, is calculated by moving a pointer to half the distance between the previous position and the current binary representation (equation 1). The binary CGR vertices were assigned to the four nucleotides as $A = (0, 0)$; $C = (0, 1)$; $G = (1, 1)$; $T = (1, 0)$. The procedure is illustrated in Figure 1.

$$CGR_i = CGR_{i-1} + 0.5 \cdot (CGR_{i-1} - g_i)$$
$$\text{with} \quad i = 1, \ldots, n_G \quad \text{and} \quad CGR_o = (0.5, 0.5). \quad (1)$$

Inversely, the nucleotide sequence can also be recovered from the CGR coordinate, where the number of bases resolved is a function of the resolution of the CGR coordinates (Goldman, 1993). This is illustrated in Figure 2 for the 8th position in *E.coli*'s *thrA* gene recovered with a resolution of 4 bits for each coordinate, which implies that a sequence of 4 bases can be resolved (underlined in the figure). It is interesting to note that the position in the CGR

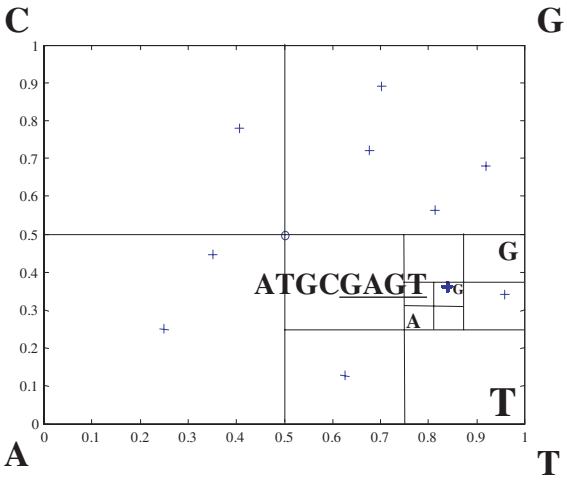


Fig. 2. Resolution of 8th CGR coordinate for *thrA* in order to recover the DNA sequence. If each CGR coordinate has a precision of 4 bits, the corresponding sequence will be recovered within 4 bases (underlined).

space is only partially dependent on the position in the genomic sequence, as the 'memory' of its coordinates is limited to the precision with which they are recorded. As a consequence, the trajectory of identical sequences, independent of their location in different genomes, will converge in the CGR space (see also Discussion).

The frequency of alternative oligonucleotide combinations, the 'genomic signature', can be determined by dividing the CGR space with a grid of appropriate size and counting occurrence in each quadrant. In order to obtain the frequency matrix of oligonucleotide length n_C, a $2^nC \times 2^nC$ grid must be used. In Figure 3, the trinucleotide frequency matrix ($n_C = 3$) is obtained for the same *thrA* sequence plotted in Figure 2.

The frequency matrices extracted from CGR (FCGR) can now be reordered in the more useful MCM format (Goldman, 1993; Almagor, 1983; Avery, 1987) as illustrated for trinucleotide frequency (Figure 4). The length considered for the preceding sequence determines the order of MCM, $n_M = n_C - 1$.

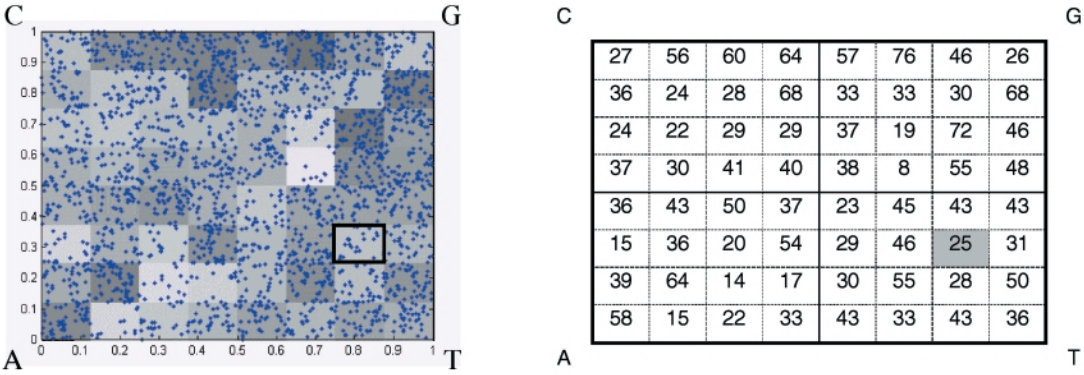


Fig. 3. Determination of trinucleotide frequency matrix for *thrA*. The CGR coordinates for the 2463 base pairs are plotted with the relative frequencies for each 8×8 quadrant represented as a grayscale (left). The distribution of counts is listed in the table (right). The quadrant highlighted in the plot and in the table is defined by the vertices `...AAGT`, `...CAGT`, `...GAGT` and `...TAGT`.

First Base	Second base A	Second base C	Second base G	Second base T	Third Base
A	58	15	20	22	A
	37	36	28	41	C
	38	33	30	55	G
	43	29	25	43	T
C	39	36	50	14	A
	24	27	60	29	C
	37	57	46	72	G
	30	23	43	28	T
G	64	43	37	17	A
	22	56	64	29	C
	19	76	26	46	G
	55	45	43	50	T
T	15	36	54	33	A
	30	24	68	40	C
	8	33	68	48	G
	33	46	31	36	T

First Base	Second base A	Second base C	Second base G	Second base T	Third Base
A	0.33	0.13	0.19	0.14	A
	0.21	0.32	0.27	0.25	C
	0.22	0.29	0.29	0.34	G
	0.24	0.26	0.24	0.27	T
C	0.30	0.25	0.25	0.10	A
	0.18	0.19	0.30	0.20	C
	0.28	0.40	0.23	0.50	G
	0.23	0.16	0.22	0.20	T
G	0.40	0.20	0.22	0.12	A
	0.14	0.25	0.38	0.20	C
	0.12	0.35	0.15	0.32	G
	0.34	0.20	0.25	0.35	T
T	0.17	0.26	0.24	0.21	A
	0.35	0.17	0.31	0.25	C
	0.09	0.24	0.31	0.31	G
	0.38	0.33	0.14	0.23	T

Fig. 4. Reordering of CGR quadrant frequencies, FCGR, left, obtained for *thrA* (Figure 3), to determine the corresponding probability matrix for the second-order Markov Chain model, right, order $n_M = n_C - 1 = 2$. The shaded cell counts `...AGT` positions (like the 8th position, Figure 1).

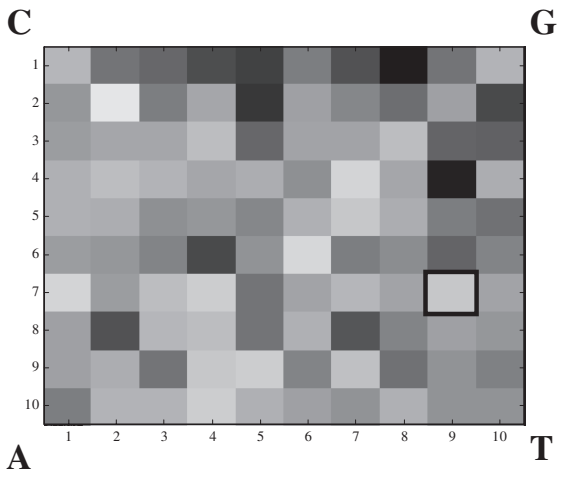


Fig. 5. $\mathrm{FCGR}_{ThrA,3.32}$: frequency 10×10 table for CGR of *thrA*, $k = 100 \Rightarrow n = 3.32$. The gray scale represents frequencies between 0 (white) and the maximum frequency in any quadrant (black). The 8th position of *E.coli*'s *thrA* will now fall in the framed quadrant, delimited by ...(agga)aggaagt; ...(cgta)cgtacgt; ...(ggaa)ggaaggt; ...(tgca)tgcatgt.

The conversion of FCGR to MCM is only straightforward if the number of quadrants k satisfies the condition in equation (2), necessary to produce an integer order.

$$k = 2^{2n_C}, n_C \text{ is an integer } \geqslant 1. \tag{2}$$

However, unlike MCM, FCGR is not constrained to represent sequences with an integer number of bases. This fundamental characteristic of CGR is illustrated for *E.coli thrA* in Figure 5 where the frequency of oligonucleotides with a fractionary length has been computed by dividing the CGR plane with a 10×10 grid ($k = 100$ violates condition in equation (2)). The non-integer portion of a fractionary sequence quantifies a restriction or redundancy in the oligonucleotide resolution as shown below. The resolved length, n_C, can be determined by removing the condition in equation (2):

$$n_C = \log_2(k)/2. \tag{3}$$

Therefore, a $10 \times 10 = 100$ quadrant FCGR ($k = 100$) tracks the frequency of oligonucleotides with 3.32 resolved bases. The number of quadrants itself can also be non-integer (e.g. by using uneven quadrants, more below), which implies that the resolved length, n_C, can indeed be any positive real number. The identity of the fractionary sequence can be determined by calculating the delimiting sequences at the vertices of the corresponding quadrant. This is accomplished by subdividing the CGR plane consecutively as illustrated in Figure 2. For $k = 100$, the 8th base of *thrA* (Figure 2) would now fall in a quadrant, highlighted in Figure 5, delimited by:

...(cgta)cgtacgt (ggaa)ggaaggt

...(agga)aggaagt ... (tgca)tgcatgt

The tetromers within brackets are repeated *ad infinitum* and the underlined trinucleotides correspond to the integer portion of the resolved length, $n = 3.32$. The definition of areas in the CGR space and the mapping of its frequency back to the sequence enables the determination of the frequency of redundant of fractionary sequences. This property of inverse frequency mapping was missed in earlier studies and the implications will be discussed below (see Discussion). By reordering the quadrants as illustrated in Figures 3 and 4 it is observed that non-integer precision in FCGR is equivalent to an incomplete higher order MCM probability table (not shown here), which is equivalent to complete MCM of non-integer order.

Global distance

The use of FCGR to access the dissimilarity between genetic sequences can be illustrated by comparing *thrA* with the remaining two *E.coli* genes involved in the metabolism of threalose, i.e. *thrA*, *thrB* and *thrC*. These genes correspond to the second, third and fourth ORFs of *E.coli* genome and code for homoserine dehydrogenase, homoserine kinase and threonine synthase, respectively (Cami *et al.*, 1993; Malumbres *et al.*, 1995). The calculation of the global distance between sequences at any scale was based on the determination of a weighted Pearson correlation coefficient, rw between the FCGRs (equation 4). The modification of the standard definition consisted on weighting the variance with the compounded frequency, nw, to determine the correlation between the two sets of FCGR quadrants x and y, each containing the occurrence of the same k oligomeric sequences.

$$\begin{aligned}
nw &= \sum_{i=1}^{k} x_i \cdot y_i \\
\bar{x}w &= \frac{\sum_{i=1}^{k} x_i^2 \cdot y_i}{nw}, \bar{y}w = \frac{\sum_{i=1}^{k} y_i^2 \cdot x_i}{nw} \\
sx &= \frac{\sum_{i=1}^{k} (x_i - \bar{x}w)^2 \cdot x_i \cdot y_i}{nw}, \\
sy &= \frac{\sum_{i=1}^{k} (y_i - \bar{y}w)^2 \cdot x_i \cdot y_i}{nw} \\
rw_{x,y} &= \frac{\sum_{i=1}^{k} \frac{x_i - \bar{x}w}{\sqrt{sx}} \cdot \frac{y_i - \bar{y}w}{\sqrt{sy}} \cdot x_i \cdot y_i}{nw}.
\end{aligned} \tag{4}$$

The advantage of using the weighted correlation coefficient defined above over using the standard definition, is

C ***thrA*** G

143	220	199	170
113	139	102	221
130	160	143	142
177	86	162	175

A total: 2482 T

C ***thrB*** G

50	94	80	85
40	47	52	77
52	59	49	56
53	31	51	77

A total: 953 T

C ***thrC*** G

64	119	105	79
66	67	65	115
74	93	73	73
104	34	70	86

d	*thrA*	*thrB*	*thrC*
thrA	0.00	0.17	0.07
thrB	0.17	0.00	0.25
thrC	0.07	0.25	0.00

Fig. 6. Dinucleotide counts ($n_C = 2$) for the FCGR of the three selected ORFs of *E.coli*, *thrA*, *B* and *C* and corresponding distance matrix (lower right table).

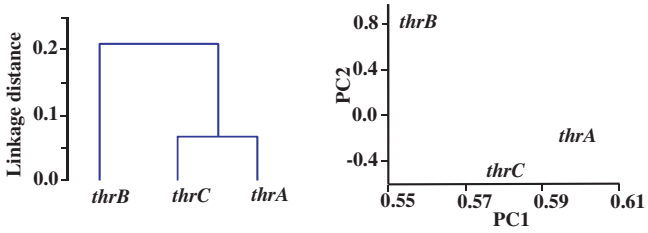
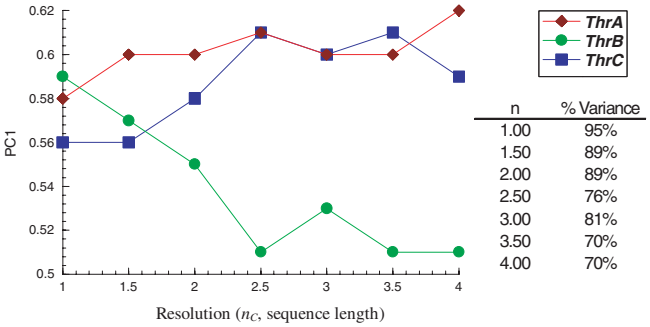


Fig. 7. Cluster analysis by unweighted pair-group average (left) and principal component extraction (right) applied to the cross-correlation (d) matrix between the three *trh* genes (Figure 6). The first component (PC1) represents 89% and the second (PC2) represents 9% of total variance.



n	% Variance
1.00	95%
1.50	89%
2.00	89%
2.50	76%
3.00	81%
3.50	70%
4.00	70%

Fig. 8. First principal component analysis of the distance matrix, d, of *thrA*, *thrB* and *thrC* with different resolutions, n_C. The table on the right lists the fraction of the variance represented by the first principal component.

that the importance of each quadrant is proportional to its frequency. In addition to giving equal weight to each occurring sequence, this procedure also enables the use of partial or uneven quadrants, and therefore the FCGR for any positive value of n_C can be obtained without biasing the correlation coefficient. The distance, d, between two DNA sequences is defined in equation (5), and has values between 0 and 2. Values above 1 would correspond to negative correlation coefficients and a null value would correspond to exact similarity.

$$d = 1 - rw. \tag{5}$$

The value of d is specific for the resolution of the frequency decompositions (FCGR) being compared. The sequences can have different lengths, n_G, but the frequency decomposition for the two sequences must have the same resolution, n_C, for d to be calculated. The ORFs *thrA*, *thrB* and *thrC* were used to illustrate the properties and applications of distance measures. The FCGR decomposition at a resolution of $n_C = 2$ (dinucleotide) for the three *thr* ORFs is presented in Figure 6, with the cross tabulation of distances in the lower right table. The distances can now be subjected to several multivariate statistical analyses, such as cluster and factor analysis (Figure 7).

Moreover, the multivariate statistical representations can be performed with arbitrary resolved lengths (n_C), screening a continuous range of scales. For example, in Figure 8 the first principal component for the correlation matrixes between three sequences was extracted with 0.5 nucleotide length increments. This plot shows that the three *thr* genes have similar succession schemes up to dinucleotide length. For longer sequences the signature of *thrB* (homoserine kinase) is observed to diverge from that of *thrA* (homoserine dehydrogenase) and *thrC* (homoserine kinase) pointing to a higher evolutionary radiation for *thrB* (Cami *et al.*, 1993; Malumbres *et al.*, 1995).

CGR has been used before to derive measures of global similarity based on ratios between quadrant frequencies (Solovyev, 1993), who also proposed unweighted correlation metrics, and distances between CGR positions (Deschavanne *et al.*, 1999). The derivation of a measure of global similarity based on a weighted Pearson correlation coefficient, above, introduces a metric more convenient for multivariate statistical analysis.

Local similarity

The identification of local homology between genomic sequences is particularly important for functional genomic studies. The development of scale-independent procedures to identify local similarity is particularly relevant since it may occur at different locations and at different or even multiple scales. Although position in the CGR plane is determined by sequence, distance between positions is only dependent on similarity between sequences. This property is a consequence of the fact that CGR position is determined by moving the pointer half the distance to the next nucleotide in the sequence (equation 1). Therefore, two sequences with the same last nucleotide cannot be further than 0.5 from each other in any coordinate

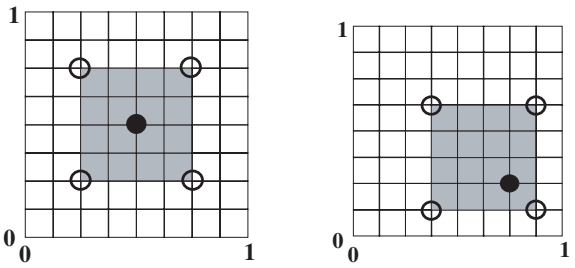


Fig. 9. Example of two sequences (filled circles) and the four possible positions for the sequence that ends in the next nucleotide (hollow circles). In both cases, the four possibilities are away from each other 0.5 in horizontal and/or vertical directions.

(Figure 9). Accordingly, two sequences that have the same last two nucleotides will not be further than 0.25 from each other, and a sequence of three similar last units will be no more than 0.125 apart. The existence of similar nucleotides in other positions upstream will shorten that distance (if all nucleotides are similar, the sequences being compared are equal and the distances between CGR positions will be null). This observation can be generalized in order to measure similarity as length of the similar sequence, n_H, as a function of the maximum absolute difference between either CGR coordinate (equation 6).

$$n_H = -\log_2(\max|\Delta CGR_{1\ldots n_G}|). \qquad (6)$$

The properties of n_H as a measure of similarity are illustrated in Figure 10.

This measure of similarity is independent of the particular nucleotides used for each vertices of the CGR plane. This can be deduced by noting that all vertices of a square, a cube, or any regular quadrangular shape are equally further apart if distance is measured as the maximum distance along any individual axis. As a consequence, the Heaviside function for homology in CGR domains, will define squares (or cubes, etc.) instead of circles (or spheres, etc.). As demonstrated in Figure 10, the measure of similarity using CGR can be the basis of a new set of algorithms to align sequences with considerable advantages over the conventional scoring methods. The investigation of local similarity between the sequences, as highlighted in Figure 10 for position 216 of *thrA*, can be quickly scanned by following a procedure akin to wavelet decomposition (Altaiski *et al.*, 1996). The maximum similar length, n_H, observed within doubling window size is recorded (Figure 11a). If the average n_H value is used instead, the scale relevance of similar length is described (Figure 11b). This representation highlights the fact that although the maximum similar length to *trhA 216* is indeed observed at *trhB 220*, for lower resolutions (similar lengths averaged over longer segments, Figure 11b), similar length is higher for *trhA* ∼*830*.

The procedure used to screen for local similarities in Figure 11 can also be used to evaluate global matching. In that case the comparison would be made between average or median CGR positions. Another possibility would be to use the average (or median) minimum distance for each CGR positions. These two approaches are similar to using minimum or average distance measures in cluster analysis.

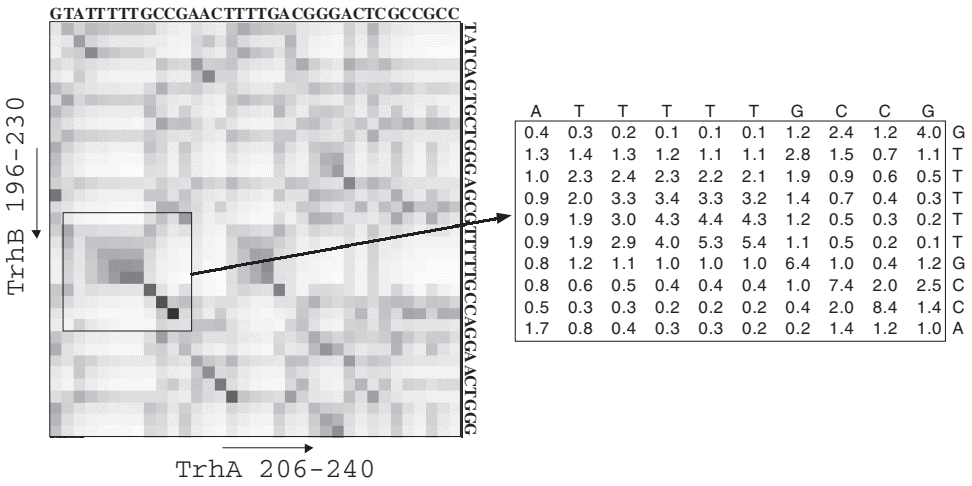


Fig. 10. Map of similarity between two sequences of *thrA* and *thrB* genes of *E.coli* as measured by n_H (equation 6). The grayscale map on the left ranges between 0 (white) and 8.4 (black), the maximum length observed for n_H between the two sequences. The numerical values for the cross-tabulation of n_H for the longest homologous sequence (framed region zooming in a similarity of eight consecutive nucleotides) are listed on the right. The maximum value of n_H corresponds to positions *thrA 216* and *thrB 220*.

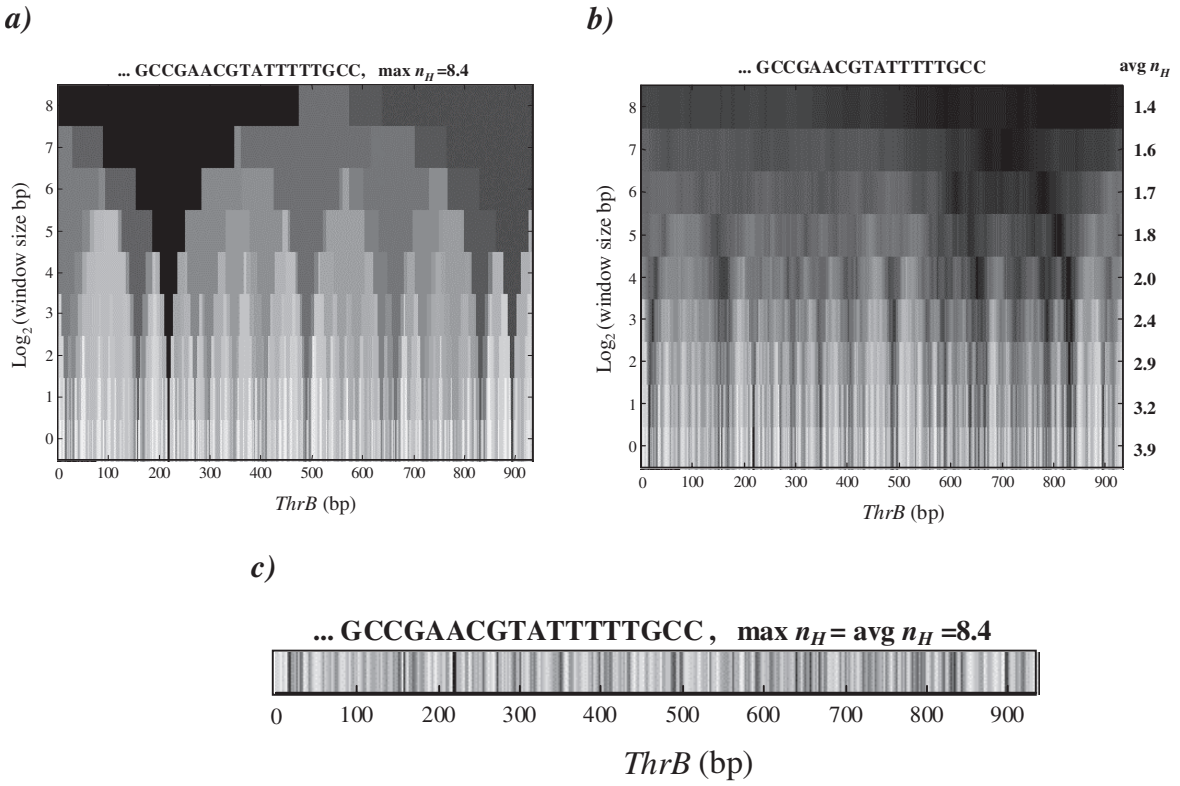


Fig. 11. Similar lengths, n_H, of different positions in *thrB* to position 216 of *trhA*, determined at different scales. The grayscale is always normalized to the maximum value, in the plot title. The scale is controlled by the width of a sliding window, represented in the *y*-axis as the logarithm of the number of nucleotides in either direction included under the window, in addition to the nucleotide in the position being analyzed. For example, for $\log_2(\text{window}) = 3$, the width of the siding window is $1 + 2{\cdot}2^3 = 17$. The n_H values for each position of *thrB* are represented in (c). The plot in (a) presents the maximum similar lengths in any position in the sliding window, determined for different window width lengths. The same procedure was followed to obtain (b) except that the average value was used instead.

CGR also offers new possibilities to resolve scale dependencies for information content in sequences. The convenience of using CGR for entropic studies of genomic sequences has been noted before (Román-Roldán *et al.*, 1994; Oliver *et al.*, 1993) albeit only for integer length resolutions. In order to use units consistent with the previous representations, the Shannon information number (Khinchin, 1957) was defined in sequence length units (equation 7). Each nucleotide corresponds to a maximum 2 bit information content and, therefore, logarithm base $2^2 = 4$ was used instead of the conventional logarithm base 2. The values for S_4 reported below can be compared with the literature values, reported in *bits*, by multiplying them by 2.

$$S_4 = \sum_{i=1}^{k} p_i \cdot \log_4 p_i$$
$$p_i = \frac{f(\text{quadrant}_i)}{\sum_{i=1}^{k} f(\text{quadrant}_i)}. \qquad (7)$$

The maximum possible value of S_4 would be obtained for a uniform random nucleotide sequence and is equal to the length of the oligonucleotide whose signature is being considered (e.g. for trinucleotide signatures $\max S_4 = 3$). It has been noted by different authors that natural sequences can be so compacted that their statistical properties are indistinguishable from random sequences (Oliver *et al.*, 1993; Hairiri *et al.*, 1990). This perplexing characteristic can now be verified with FCGR resolved in a continuous scale. In the example below (Figure 12), the value of S_4 for the entire *E.coli* genome and for the open reading frame of *thrA* are plotted together for fractional increases in the resolved length (n_C).

DISCUSSION

The relevance of accessing the frequency of non-integer genomic sequences may not be apparent at first given that all sequences are physically made of integer number of nucleotides. However, the loss of resolution by redundant reading is in fact equivalent to the sequence resolved

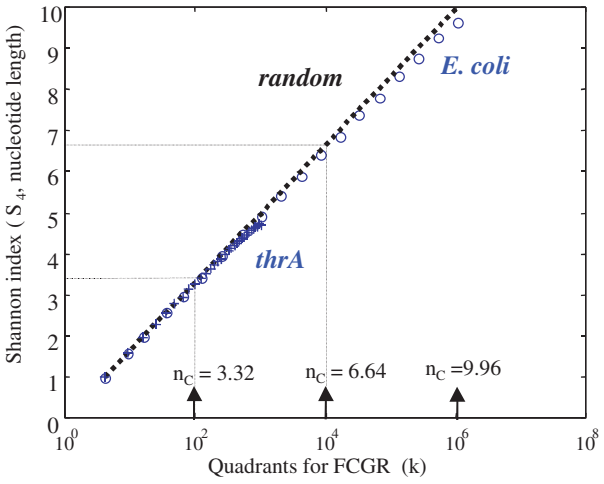


Fig. 12. Entropic content, S_4, of *E.coli* genome (circles), and that of *thrA* ORF (crosses). A random sequence would have an entropic content equal to the resolved length, n (dashed line). Recall that n_C is related to k by equation (3).

as being composed of a non-integer number of units (Figure 5). For example, if for a particular position A and G can be always interchanged, as can C and T, then the equivalent sequence would only encode information equivalent to 1/2 unit (using equation (3): $k = 2, n_C = \log_2^{(2)}/_2 = 1/2$). This is particularly relevant, as redundancy is a fundamental attribute of genome structures (Jeffrey, 1990). The degenerated translation of trinucleotides (Hsiung, 1997) is illustrative: 61 sense trinucleotide codons encode for 20 amino acids, and the remaining three nonsense codons signal for the end of transcription. The resolved length, n_C, is only 2.2 (equation 5). Synonym sense codons differ in the third nucleotide (the nonsense codons may also differ in the second position), which sets the scale for this particular redundancy. The amino acid sequence itself is also redundant in the sense that protein function is not affected by some amino acid substitutions. Redundancy in proteins is a function of position and a consequence of similar physicochemical properties between some amino acids. In conclusion, the property that is the object of natural selection, i.e. function, depends on a redundant sequence, that is to say a sequence resolved with a non-integer resolution. Consequently, functional correlations are bound to be maximal for non-integer resolutions (n_C), which identify the relevant scale dependency.

The fact that CGR simultaneously represents sequences at different scales also allows for new measures of homology between sequences to be developed. The scale independence of CGR of genetic sequences can be used to investigate local and global homology.

CONCLUSIONS

The CGR of sequences is a method to ordinate the entire domain of possibilities in a continuous two-dimensional space. This enables determination of global distance (equation 3) and local similarity (equation 4) for a continuous range of resolved lengths, as illustrated in Figures 6–11 using *E.coli*'s threonine operon. Consequently, the CGR transformation makes DNA sequences amenable to an entire new set of statistical analysis tools. For example, the average sequence can now be calculated by recovering the sequence of the position with the average CGR coordinates. Similarly, the standard deviation can now be determined and then converted to similar sequence length units using equation (6). For example, the *thrA* gene analyzed above has an average sequence of ATGGGGTTCGCATCTGCTAGCACGAGAAGAGACACTGTCGGACGCGGAAAAG, the median sequence of TGGTATGCTTGACGGGGGAAAAG, which has a similar length, n_H, to the average sequence of 7.7 nucleotides, which highlights a sequence distribution skewness of 0.13 unit length. Finally, the standard deviation and standard errors of *thrA* (2463 bp) sequence have similar lengths, n_H, of 4.8 and 1.8 nucleotides, respectively. This simple example illustrates the fact that CGR is a formalism that bridges between sequences of discrete units and numeric coordinates in a continuous space. Consequently, basic statistic measures and techniques can now be applied to sequences and a wide range of new tools can now be devised for statistical analysis. The utilization of mutivariate statistical methods, such as cluster analysis and factor extraction was demonstrated in Figures 7 and 8. Non-hierarchical clustering methods such as Kohonen mapping and K-nearest means (Kohonen, 1997), as well as advanced classification methods such as neural networks (Bishop, 1995), can be easily applied to the FCGR of biological sequences. The authors also expect the properties of CGR to be relevant for computational and fundamental advances in the Bayesian modeling in general. The current limitations of Hidden Markov Chains to cope with simultaneous dependencies at multiple scales may be overcome by identifying multiple alternative states with CGR. This may be particularly effective when coupled with the machine learning approaches increasingly advocated (Baldi and Brunak, 1998) to extend the capabilities of Markov models.

ACKNOWLEDGEMENTS

This research was supported by the NSF grant no. MCB-9802342. J.S.A. also acknowledges the financial support by grant FMRH/BSAB/60/98 Fundação para a Ciência e Tecnologia /Min. Ciência e Tecnologia, Lisbon, Portugal (FCT/MCT) and access to National Computational Science Alliance (NCSA) supercomputers provided by grant

BCS99004N: Statistical Mechanic Studies of Biological Systems. J.A.C. was supported by grant from contract PRAXIS/P/SAU/14051/1998, FCT/MCT.

REFERENCES

Almagor,H. (1983) *J. Theor. Biol.*, **104**, 633–645.

Altaiski,M., Mornev,O. and Polozov,R. (1996) *Genet. Anal. Biomol. Eng.*, **12**, 165–168.

Altschul,S.F. and Koonin,E.V. (1998) *Trends Biochem. Sci.*, **23**, 444–447.

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.

Avery,P.J. (1987) *J. Mol. Evol.*, **26**, 335–340.

Baldi,P. and Brunak,S. (1998) *Bioinformatics, the Machine Learning Approach*. MIT Press, Cambridge, MA.

Bar-Yam,Y. (1997) *Dynamics of Complex Systems*. Addison-Wesley, Reading, MA.

Basu,S., Pam,A., Dutta,C. and Das,J. (1997) *J. Mol. Graph. Model.*, **15**, 279–289.

Bishop,C.M. (1995) *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford.

Burma,P.K., Raj,A., Deb,J.K. and Brahmachari,S.K. (1992) Genome analysis—a new approach for visualization of sequence organization in genomes. *J. Biosci.*, **17**, 395–411.

Cami,B., Clepet,C. and Patte,J.C. (1993) *Biochimie*, **75**, 487–495.

Casjens,S. (1998) *Annu. Rev. Genet.*, **32**, 339–377.

Deschavanne,P.J., Giron,A., Vilain,J., Fagot,G. and Fertil,B. (1999) *Mol. Biol. Evol.*, **16**, 1391–1399.

Fiser,A., Tusnády,G.E. and Simon,I. (1994) Chaos game representation of protein structures. *J. Mol. Graphics*, **12**, 302–304.

Goldman,N. (1993) *Nucleic Acids Res.*, **21**, 2487–2491.

Hairiri,A., Weber,B. and Olmsted,J. (1990) *J. Theor. Biol.*, **147**, 235–254.

Hsiung Li,W. (1997) *Molecular Evolution*. Sinauer, Sunderland, MA, pp. 21–23.

Jeffrey,H.J. (1990) *Nucleic Acids Res.*, **18**, 2163–2170.

Karlin,S., Campbell,A.M. and Mrázek,J. (1998) *Annu. Rev. Genet.*, **32**, 185–225.

Khinchin,A.I. (1957) *Mathematical Foundations of Information Theory*. Dover, New York.

Kohonen,T. (1997) *Self-Organizing Maps*. Springer, Berlin.

Krogh,A. (1998) *Computational Biology: Pattern Analysis and Machine Learning Methods*. Chapter 4, Salzberg,S., Searls,D. and Kasif,S. (eds), Elsevier, Amsterdam.

Malumbres,M., Mateos,L.M., Guerrrero,C. and Martin,J.F. (1995) *Folia Microbiol.*, **40**, 595–606.

Oliver,J.L., Bernaola-Galvan,P., Guerrero-Garcia,J. and Roman-Roldan,R. (1993) *J. Theor. Biol.*, **160**, 457–470.

Pearson,W.R. (1996) *Meth. Enzymol.*, **266**, 227–258.

Pearson,W.R. and Miller,W. (1992) *Meth. Enzymol.*, **210**, 575–601.

Pleißner,K.P., Wernisch,L., Osvald,H. and Fleck,E. (1997) *Electrophoresis*, **18**, 1709–2713.

Rabiner,L.R. and Juang,B.H. (1986) *IEEE ASSP Magazine*, **3**, 4–16.

Román-Roldán,R., Bernaola-Galván,P. and Oliver,J.L. (1994) *Pattern Recognition Lett.*, **15**, 567–573.

Solovyev,V.V. (1993) Fractal graphical representation and analysis of DNA and protein sequences. *Biosystems*, **30**, 137–160.

Tino,P. (1999) Spatial representation of symbolic sequences through iterative function systems. *IEEE Trans. Syst. Man Cybernet.*, **29**, 386–393.