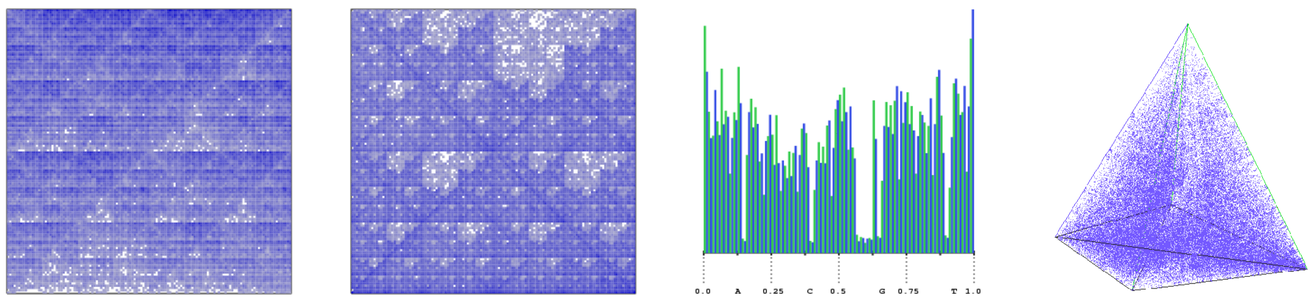
Test on the Structure of Biological Sequences via Chaos Game Representation

Peggy Cénac

INRIA Domaine du Voluceau, B.P. 105, 78 153 Le Chesnay Cedex (France)
Peggy.Cenac@inria.fr



1 Chaos Game Representation

The CGR is both a graphical representation method of sequences and a storage tool. This iterative mapping technique was apparently for the first time applied to genomic sequences by Jeffrey [2]. From a given sequence –e.g. nucleotides in a DNA sequence or amino acids in a protein– one can define trajectories in a bounded set conserving all its statistical properties. **Each point of the CGR contains the whole history** of the sequence. One of the central goals of Cénac *et al.* [1] is to figure out whether the CGR provides more information than the classical methods based on word-counting.

Definitions

➢ $U = u_1 \dots u_n$ a sequence of letters in a d-letter alphabet $\mathcal{A}$.

➢ The **Chaos Game Representation** of U, on a bounded Borel set $S \subset \mathbb{R}^q$ is a sequence $\{X_0, \dots, X_n\}$ defined by

$$\begin{cases} X_0 \in S \\ X_{i+1} = \theta\Big(X_i + \ell_{u_{i+1}}\Big) \stackrel{\text{def}}{=} T_{u_{i+1}}(X_i), \end{cases}$$

for $0 < \theta < 1$.

➢ Jeffrey's definition for DNA sequences, $\mathcal{A} = \{A, C, G, T\}$:

$$\begin{cases} S = [0,1[^2, \quad \theta = \frac{1}{2}, \quad X_0 = (\frac{1}{2}, \frac{1}{2}) \\ \ell_A = (0,0), \quad \ell_C = (0,1), \quad \ell_G = (1,1), \quad \ell_T = (1,0). \end{cases}$$

➢ Counting points in $Sw \stackrel{\text{def}}{=} \sum_{k=1}^{i} \theta^{i-k+1} \ell_{u_k} + \theta^i S \Leftrightarrow$ counting occurrences of the word w.

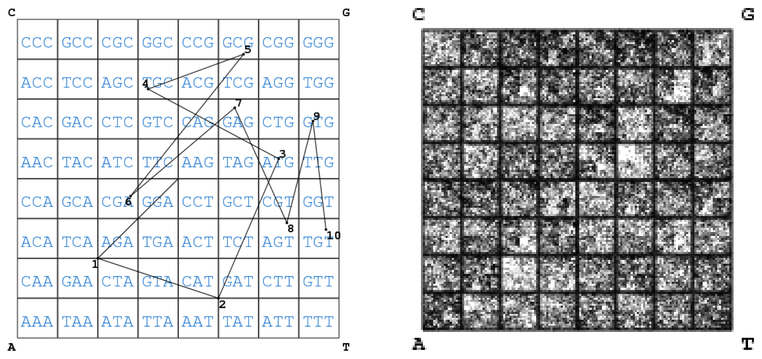


Figure 1 : On the left, CGR of the 10 first nucleotides ATGCGAGTGT of the *E. Coli* threonine gene. Point number 3 corresponds to the first 3-letter word *ATG*. It is located in the corresponding quadrant. The second 3-letter word *TGC* corresponds to point 4 and so on. On the right, CGR of a sequence of length 80000 of *E. Coli*.

Stochastic properties of the CGR

➢ U is supposed to be a stationary ergodic sequence.

➢ $(X_n)_{n\geq 0}$ is a Markov chain of order 1, and **converges almost surely to a random vector X with distribution π**.

➢ When U is i.i.d. and uniformly distributed, π is the Lebesgue measure on S. Whenever S is not uniformly distributed, π is continuous, singular with respect to the Lebesgue measure. The law of large number holds, and the empirical measures converge.

Characterization of Structure

➢ For any word $w = u_1 \dots u_i$ and for any set $B \subset S$, $Bw \stackrel{\text{def}}{=} T_{u_i} \circ \dots \circ T_{u_1}(B)$.

Proposition 1.1. *The stationary random sequence U is*

- *an i.i.d. sequence if and only if*

$$\pi(Bu) = \pi(B)\pi(Su), \quad \forall u \in \mathcal{A}, \ \forall B \subset S.$$

- *a Markov chain of order m if and only if $\forall B \subset S$, $\forall w \in \mathcal{A}^m$, $\forall u \in \mathcal{A}$,*

$$\frac{\pi(Bwu)}{\pi(Bw)} = \frac{\pi(Swu)}{\pi(Sw)}, \quad \forall B \subset S, \quad \forall w \in \mathcal{A}^m, \quad \forall u \in \mathcal{A}.$$

In particular the ratio $\frac{\pi(Bwu)}{\pi(Bw)}$ does not depend on B.

➢ Characterization of independence and of Markov chains

➤ construction of a test.

➤ genomic signature (see Cénac *et al.* [1])

2 Testing the structure of a sequence

- H_0 : "$U = u_1 \dots u_N$ *is an i.i.d. sequence*"
- H_m : "U *is a Markov chain of order* m"
- H : "U *is a stationary ergodic sequence*

➢ *Construction*

Denoting $q_\alpha(d)$ the $(1-\alpha)$-quantile of the chi-square distribution $\chi^2(d)$, and $\hat{\pi}_n(E) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\{E\}}(X_j)$ the empirical measure of π, for any partition $\{B_1, \dots, B_K\}$ of S, with $K > 1$, the following sets are reject region with asymptotic level α, of a test of H_0 against $H \setminus H_0$ and respectively of H_m against $H \setminus H_m$

$$\left\{ \sum_{\substack{1\leq i\leq K \\ v\in\mathcal{A}}} \frac{n\Big(\hat{\pi}_n(Bv) - \hat{\pi}_n(B)\hat{\pi}_n(Sv)\Big)^2}{\hat{\pi}_n(B)\hat{\pi}_n(Sv)} > q_\alpha\Big[(d-1)(K-1)\Big] \right\},$$

$$\left\{ \sum_{\substack{wu\in\mathcal{A}^m\times\mathcal{A} \\ 1\leq i\leq k}} \frac{n\Big(\hat{\pi}_n(Sw)\hat{\pi}_n(Bwu) - \hat{\pi}_n(Swu)\hat{\pi}_n(Bw)\Big)^2}{\hat{\pi}_n(Sw)\hat{\pi}_n(Swu)\hat{\pi}_n(Bw)} > q_\alpha\Big[d^m(d-1)(K-1)\Big] \right\},$$

➢ *Consistence*

Next, assume that $H \setminus H_0$ (resp. $H \setminus H_m$) holds, and let $B \subset S$, $w \in \mathcal{A}^d$ and $v \in \mathcal{A}$ be such that

- $$\pi(Bv) \neq \pi(B)\pi(Sv).$$

- respectively
$$\pi(Sw)\pi(Bwu) \neq \pi(Bw)\pi(Swu).$$

If B is one of the set forming the partition, then the test is asymptotically consistent.

For this test, Reinert *et al.* [3] propose to make use of Pearson statistics. In the special case when the partition is $\{Su, u \in \mathcal{A}\}$, the tests are asymptotically the same.

3 Numerical experiments

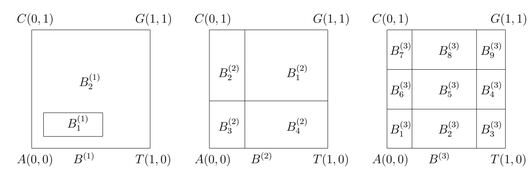


Figure 2 : The 4 different partitions of the square $[0,1[^2$ arbitrary chosen for the test.

We did generate 1000 sequences of length n of Markov chains of order m with random transition matrix, for various values of n.

order	n	Pearson	$B^{(1)}$	$B^{(2)}$	$B^{(3)}$
0	100	4.2%	3.6%	3.4%	2.5%
	500	6.1%	4.8%	4.8%	3.8%
	1000	5.0%	5.6%	6.2%	5.3%
	10000	6.5%	4.8%	5.1%	7.7%
1	100	86.4%	12.9%	51.1%	28.9%
	500	100%	54.2%	98.7%	94.5%
	1000	100%	70.9%	99.9%	99.0%
	10000	100%	97.6%	100%	100%
5	1000	8.6%	6.8%	8.6 %	8.4 %
	10000	54.6%	28.7%	55.6%	85.3%
	80000	99.4%	84.5%	99.6%	100%
2 mixed	500	5.8%	16.5%	49.9%	76.8%
	1000	7.0%	26.9%	73.7%	95.1%
	10000	7.3%	73.2%	99.8%	100%
5 mixed	80000	5.8%	29.7%	76.7%	85.8%

order	n	Pearson	$B^{(2)}$	$B^{(3)}$
1	500	5.2%	5.4%	3.6%
	1000	4.1%	5.2%	5.4%
	10000	5.3%	6.4%	6.0%
2	100	60.8%	25.8%	1.3%
	500	100%	98.4%	91.6%
	1000	100%	99.9%	99.6%
	10000	100%	100%	100%
4	500	22.8%	20.2%	30.3%
	1000	51.0%	44.3%	69.7%
	5000	99.9%	99.6%	100%
	10000	100%	100%	100%
3 mixed	500	5.4%	29.6%	48.5%
	1000	8.1%	55.6%	79.3%
	5000	8.3%	96.4%	99.8%
	10000	6.3%	99.0%	100%

Table 1 : The left (resp. right) hand side of the following Table shows the fraction of cases when H_0 (resp. H_1) is rejected.

➢ The **choice of the partition is crucial**.

➢ The reject of long dependence Markov chains increases with the size of the partition.

➢ In the special case of markov chains of order-m given by the aggregation of m indepedent markov chains of order 1, the CGR-based test behaves pretty well. This illustrates **the strength of the CGR**: it does not impose any constraint on the input sequence besides stationarity.

References

[1] P. Cénac, G. Fayolle, and J.M. Lasgouttes. Dynamical systems in the analysis of biological sequences. Technical Report 5351, INRIA, october 2004.

[2] H.J. Jeffrey. Chaos Game Representation of gene structure. *Nucleic Acid. Res*, 18:2163–2170, 1990.

[3] G. Reinert, S. Schbath, and M.S. Waterman. Probabilistic and statistical properties of words: An overview. *Journal of Computational Biology*, 7(1/2):1–46, 2000.