Available online at www.sciencedirect.com
SCIENCE DIRECT®

Physics Letters A 332 (2004) 250–255

PHYSICS LETTERS A

www.elsevier.com/locate/pla

Recurrence plot analysis of DNA sequences

Zuo-Bing Wu

State Key Laboratory of Nonlinear Mechanics, Institute of Mechanics, Chinese Academy of Sciences, Beijing 100080, China

Received 3 June 2004; received in revised form 9 September 2004; accepted 14 September 2004

Available online 7 October 2004

Communicated by C.R. Doering

Abstract

Recurrence plot technique of DNA sequences is established on metric representation and employed to analyze correlation structure of nucleotide strings. It is found that, in the transference of nucleotide strings, a human DNA fragment has a major correlation distance, but a yeast chromosome's correlation distance has a constant increasing.

Keywords: DNA sequences; Recurrence plot; Correlation structure

1. Introduction

The heredity and variation information of all organisms is embodied in DNA sequences. To understand the one-dimensional symbolic sequence made of four letters A, C, G, and T, some statistical and geometrical methods are developed in bioinformatics. Chaos game representation (CGR) [1], which generates a two-dimensional square from a one-dimensional sequence, provides a technique to visualize the composition of DNA sequences. The characteristics of CGR image is described as genomic signature [2]. A visualization scheme of the string composition of DNA sequences is proposed and used to trace evolutionary relatedness of species [3,4]. Recently, a one-to-one metric representation (MR) [5,6] is proposed to make an ordering of subsequences in a plane and determine suppression of certain nucleotide strings in DNA sequences. By using the MR method, self-similarity limits of genomic signatures are determined as optimal string lengths for generating the genomic signatures [6].

For a chaotic system, the dynamical trajectory is always attracted in a finite set. To depict the finite set, i.e., measure its self-similarity, information dimension is designed [7]. At the same time, to describe the ergodicity of trajectory, i.e., reflect natural time correlation information of dynamical system, a recurrence plot technique is presented [8]. The methods can be used to diagnose unknown dynamical information from an experimental time series. Especially, by using the tool of recurrent plot, dynamical assessment of physiological systems is illustrated [9], nonrandom

E-mail address: wuzb@lnm.imech.ac.cn (Z.-B. Wu).

0375-9601/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.physleta.2004.09.061

dynamical components in synaptic noise of central neurons are evidenced [10] and structure–function relationships of proteins are quantified [11].

In DNA sequences, correlation structure of nucleotide strings, as an important part of the DNA architecture, is covered in the transference of nucleotide strings. In addition, DNA transposable elements, which are found in all organisms, have ability to move from place to place and make many copies within the genome via the transposition [12,13]. In general, the correlation structure of nucleotide strings can provide an understanding of the elements transposition in an extensive region. In this Letter, using the MR method, we extend recurrence plot technique to analyze correlation structure of nucleotide strings and determine transference of nucleotide strings.

2. Method

A DNA sequence is a one-dimensional symbolic sequence $s_1 s_2 \ldots s_i \ldots s_N$ $(s_i \in \{A, C, G, T\})$. In a two-dimensional MR, we make a correspondence of points (α, β) and subsequences $\Sigma_m = s_1 s_2 \ldots s_m$ $(1 \leqslant m \leqslant N)$. Subsequences with the same ending k-nucleotide string, which are labeled by $\Re_k$, correspond to points (α, β) in the zone encoded by the k-nucleotide string. For a given k-nucleotide string, we have a set $\Re_k$ and a correspondent zone size $\epsilon_k = 3^{-k}$. Taking a subsequence $\Sigma_i \in \Re_k$, we calculate

$$\Theta\big(\epsilon_k - |\Sigma_i - \Sigma_j|\big) = \Theta\Big(\epsilon_k - \sqrt{(\alpha_i - \alpha_j)^2 + (\beta_i - \beta_j)^2}\Big), \tag{1}$$

where Θ is the Heaviside function [$\Theta(x) = 1$, if $x > 0$; $\Theta(x) = 0$, if $x \leqslant 0$] and Σ_j is a subsequence $(k \leqslant j)$. When $\Theta(\epsilon_k - |\Sigma_i - \Sigma_j|) = 1$, i.e., $\Sigma_j \in \Re_k$, we plot a point (i, j) in a plane. If Σ_j is taken from the beginning of one-dimensional symbolic sequence and shifted forward, we plot all correspondent points in the plane and obtain some points at the position i. When the position i is moved from the beginning of one-dimensional symbolic sequence, Σ_i belongs to another k-nucleotide string set. Repeating the above process, we obtain a recurrence plot of the DNA sequence. In the recurrent plot, there exists a mirror symmetry with respect to the diagonal $i = j$. A point in the recurrence plot means that two subsequences Σ_i, $\Sigma_j \in \Re_k$ have a distance $|j - i|$ in the DNA sequence. Along with the increase of k, each zone size in the MR plane decreases. For a given Σ_i, neighboring points (encoded by Σ_j) of its correspondent point in the zone decrease, so that the points (i, j) in the recurrence plot plane decrease.

To quantify the correlation structure in recurrence plot plane, we define a correlation intensity at a given correlation distance l

$$\Xi(l) = \sum_{i=1}^{N-l} \Theta\big(\epsilon_k - |\Sigma_i - \Sigma_{i+l}|\big). \tag{2}$$

The quantity displays the transference of all k-nucleotide strings with the correlation distance l in the DNA sequence.

3. Results

3.1. Correlation structure of HUMHBB

To display properties of recurrence plots, we take HUMHBB (human β-region, chromosome 11) with 73308 bases as an example and draw the recurrence plots for $k = 7,\ 9,\ 11,\ 13$ and 15. Along with the increase of k, a point density in the recurrence plot plane decreases monotonically. The recurrence plot with a high density is easier to investigate global properties than that with a low density, but to find local properties such as the transference of long nucleotide strings, latter is better. In Fig. 1(a), a recurrence plot of HUMHBB for $k = 9$ is plotted. Beside a high and a low densities appear locally in the recurrence plot plane, most of regions have a similar distribution density, i.e., the transference of 9-nucleotide strings in HUMHBB is well-distributed. The high and low densities display that the 9-nucleotide strings at the position close to one third of sequence length have a high frequency in the themselves positions and have a low frequency near the ending position of sequence.

To analyze local properties in the recurrence plot plane, we move to the coarse-grained recurrent plot for $k = 15$. In Fig. 1(b), there exist some short and long lines, which parallel to the diagonal. The diagonal parallel lines describe that some long nucleotide strings $(> k)$ are transfered in the sequence. Especially, at the same i position, several short parallel

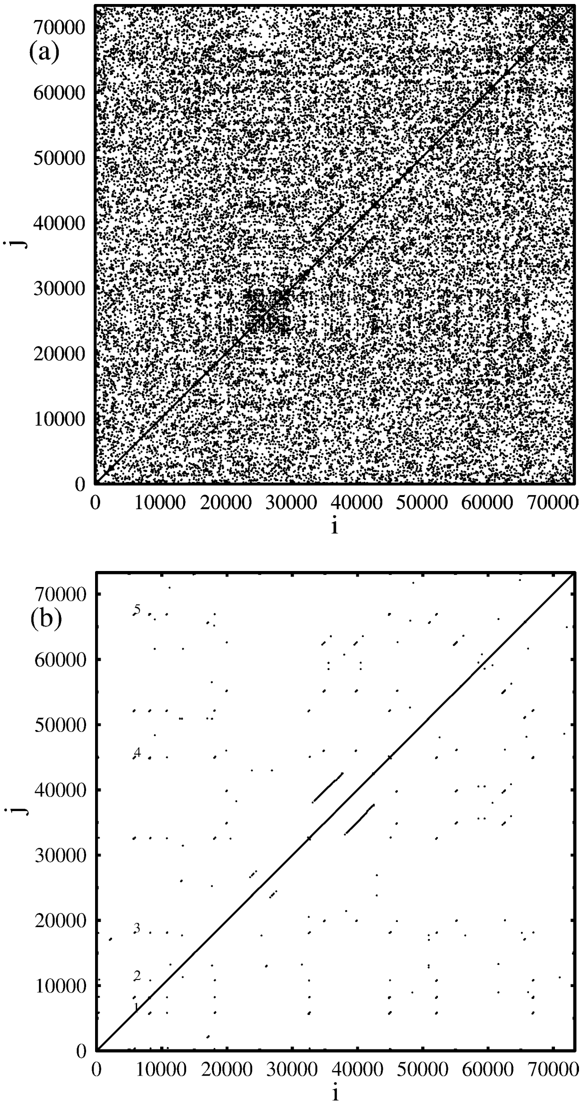

Fig. 1. Recurrence plots of HUMHBB for (a) 9-nucleotide strings; (b) 15-nucleotide strings.

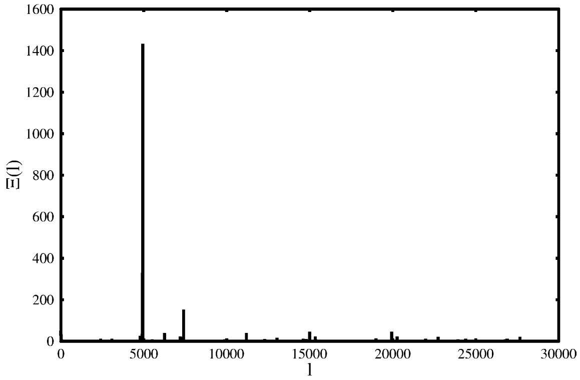

Fig. 2. A plot of correlation intensity $\Xi(l)$ versus correlation distance l calculated from Fig. 1(b).

lines with different correlation distances appear. They correspond to that a long nucleotide string is copied many times. For example, five short parallel lines near $i = 5800$ correspond to that a 21-nucleotide string $g^2ag^2ctgag^2cag^2aga^2tc$ repeats four times, i.e., the first one exists at position 5797 to 5817 in the diagonal labeled by 1, the second one exists at position 10783 to 10803 labeled by 2, the third one exists at position 18077 to 18097 labeled by 3, and the fourth one exists at position 66936 to 66956 labeled by 5, as well as its subsequence, a 20-nucleotide string $g^2ag^2ctgag^2cag^2aga^2t$, repeats two times, i.e., the first one occurs in the range of 5797 to 5816 in the diagonal labeled by 1 and the second one occurs in the range of 44962 to 44981 labeled by 4. The five nucleotide strings have different correlation distance, that leads to the global transference of nucleotide strings. When many neighboring nucleotide strings are transfered with the same correlation distance, a long parallel line appears in the recurrence plot plane. It displays the local transference of nucleotide strings. From the recurrence plot, transference of long nucleotide strings can be determined. In order to preserve the total number of long nucleotide strings is not too large, we take the cut-off threshold of 30 nucleotides. In Table 1, all repeated $k(\geqslant 30)$-nucleotide strings and their correspondent positions in HUMHBB are presented. Many long nucleotide strings have correlation distances 4916 and 4936, which correspond to long parallel lines in the center of Fig. 1(b). The longest one, which has the correlation distance 4936, is a 1058-nucleotide string $t^2a \ldots gtg$ positioned in the ranges of 34503 to 35560 and 39439 to 40496.

Fig. 2 displays the correlation intensity $\Xi(l)$ at different correlation distance l for $k = 15$. When k is taken as 9, the global behavior in Fig. 2 is still preserved. The maximal correlation intensity appears at $l = 4936$, i.e., transference of 15-nucleotide strings over 4936 is the most powerful. At the correlation distance $l = 4916$, the correlation intensity reaches the second local maximal value. For other correlation distance, the correlation intensity is not larger than one

Table 1
Transference of nucleotide strings with lengths $k(\geqslant 30)$ for HUMHBB

No.	String	Length	Position 1	Position 2	l
1	$ctc \ldots g^3$	41	8032–8072	44802–44842	36770
2	$g^2t \ldots c^3$	31	8194–8224	52153–52183	43959
3	$t^2g \ldots tga$	30	10780–10809	66933–66962	56153
4	$gtg \ldots atg$	30	13037–13066	26061–26090	13024
5	$agc \ldots gca$	31	19925–19955	55180–55210	35255
6	$ctg \ldots agt$	48	19936–19983	34926–34973	14990
7	$ctg \ldots agt$	48	19936–19983	39862–39909	19926
8	$aga \ldots c^2a$	44	33551–33594	38489–38532	4938
9	$tga \ldots tga$	38	33769–33806	38707–38744	4938
10	$tca \ldots t^2a$	73	34007–34079	38943–39015	4936
11	$ct^2 \ldots g^2t$	41	34081–34121	39017–39057	4936
12	$a^3 \ldots atc$	31	34123–34153	39059–39089	4936
13	$a^3 \ldots a^3$	35	34172–34206	39108–39142	4936
14	$tg^2 \ldots g^2t$	112	34208–34319	39144–39255	4936
15	$c^2t \ldots gag$	181	34321–34501	39257–39437	4936
16	$t^2a \ldots gtg$	1058	34503–35560	39439–40496	4936
17	$atg \ldots tga$	43	34818–34860	45995–46037	11177
18	$ca^2 \ldots gct$	32	35651–35682	40567–40598	4916
19	$gca \ldots gct$	55	35717–35771	40633–40687	4916
20	$g^2t \ldots ctg$	175	35773–35947	40689–40863	4916
21	$agt \ldots agc$	35	35949–35984	40865–40900	4916
22	$cag \ldots gct$	60	36000–36059	40916–40975	4916
23	$tct \ldots t^3$	31	36100–36130	41016–41046	4916
24	$ctc \ldots g^3$	30	36348–36377	41253–41282	4905
25	$a^3 \ldots ca^2$	31	37015–37045	41788–41818	4773
26	$atg \ldots tga$	43	39754–39796	45995–46037	6241
27	$a^2c \ldots a^3$	37	44880–44916	52074–52110	7194
28	$gtg \ldots gca$	37	54761–54797	62158–62194	7397
29	$agt \ldots t^2a$	40	54858–54897	62255–62294	7397
30	$aga \ldots tct$	43	54939–54981	62336–62378	7397
31	$ctg \ldots tc^2$	54	55014–55067	62413–62466	7399
32	$ctc \ldots ga^2$	49	55069–55117	62468–62516	7399
33	$g^2t \ldots cac$	47	55125–55171	62524–62570	7399
34	$ctg \ldots gtc$	58	55182–55239	62581–62638	7399
35	$tgt \ldots tgt$	31	59457–59487	59459–59489	2

tenth of the maximal one. The properties give an evidence that the transference of nucleotide strings in HUMHBB has a major correlation distance.

3.2. *Correlation structure of Yeast1*

Recurrence plots of Yeast1 (*Saccharomyces cerevisiae* yeast, chromosome 1) with 230209 bases for $k = 11$ and 15 are displayed in Fig. 3(a) and (b). Most parts of the pattern in Fig. 3(a) are similar to those in Fig. 1(a). In the comparison of coarse-grained recurrence plots, a difference of Fig. 3(b) from Fig. 1(b) is two square sets of points near the diagonal, which are labeled by 1 and 2. A square set of points consists of many diagonal parallel lines, which correspond to many neighboring repeated nucleotide strings with a basic correlation distance. The transference of nucleotide strings is a local behavior. For example, in the square set 1, a 95-nucleotide string $at^2 \ldots g^2t$ repeats two times at its neighboring region. They are distributed in the ranges of 25739 to 25833 and 25874 to 25968 and have a correlation distance 135. We take the correlation distance 135 as a basic one. A 101-nucleotide string $gta \ldots ac^2$ repeats three times at its neighboring region. The first one exists at position 25751 to 25851, the second one exists at position 26561

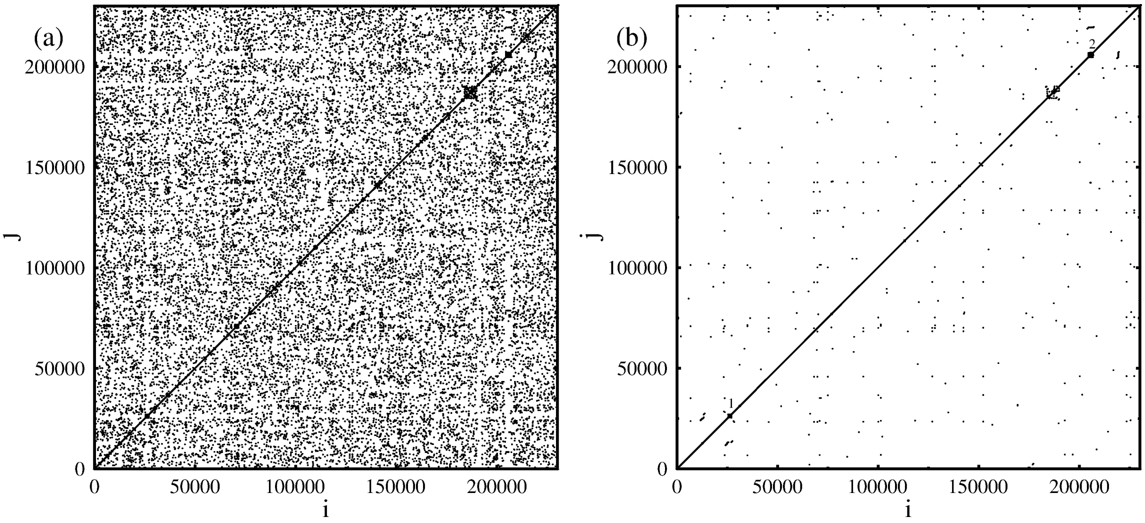

Fig. 3. Recurrence plots of Yeast1 for (a) 11-nucleotide strings; (b) 15-nucleotide strings.

Table 2
Transference of nucleotide strings with lengths $k(\geqslant 90)$ for Yeast1

No.	String	Length	Position 1	Position 2	l
1	$t^2a\ldots act$	225	11745–11969	24177–24401	12432
2	$ctg\ldots a^2t$	139	12258–12396	24711–24849	12453
3	$g^2a\ldots g^2a$	184	12988–13171	25153–25336	12165
4	$c^2g\ldots ac^2$	137	25715–25851	26255–26391	540
5	$at^2\ldots g^2t$	95	25739–25833	25874–25968	135
6	$at^2\ldots ac^2$	113	25739–25851	26414–26526	675
7	$gta\ldots ac^2$	101	25751–25851	26561–26661	810
8	$gta\ldots ac^2$	101	25751–25851	26696–26796	945
9	$t^2g\ldots g^2t$	116	25853–25968	26393–26508	540
10	$at^2\ldots g^2t$	95	25874–25968	26279–26373	405
11	$atg\ldots gtg$	111	25925–26035	26060–26170	135
12	$atg\ldots agt$	134	25925–26058	26195–26328	270
13	$agt\ldots gtg$	121	26050–26170	26185–26305	135
14	$at^2\ldots gac$	128	26279–26406	26414–26541	135
15	$gta\ldots gac$	116	26291–26406	26561–26676	270
16	$gta\ldots gac$	116	26291–26406	26696–26811	405
17	$gta\ldots gtg$	285	26426–26710	26561–26845	135
18	$tga\ldots aca$	337	160239–160575	165827–166163	5588
19	$cac\ldots tac$	285	204518–204802	204653–204937	135
20	$g^2t\ldots tac$	101	204567–204667	205512–205612	945
21	$g^2t\ldots tac$	101	204702–204802	205512–205612	810
22	$g^2t\ldots a^2t$	113	204837–204949	205512–205624	675
23	$ac^2\ldots c^2a$	115	204855–204969	205395–205509	540
24	$ctc\ldots cat$	127	205042–205168	205312–205438	270
25	$cac\ldots act$	121	205058–205178	205193–205313	135
26	$cac\ldots cat$	111	205193–205303	205328–205438	135
27	$ac^2\ldots a^2t$	95	205395–205489	205530–205624	135
28	$atg\ldots t^2c$	122	205758–205879	206433–206554	675
29	$ac^2\ldots tg^2$	92	205911–206002	206181–206272	270

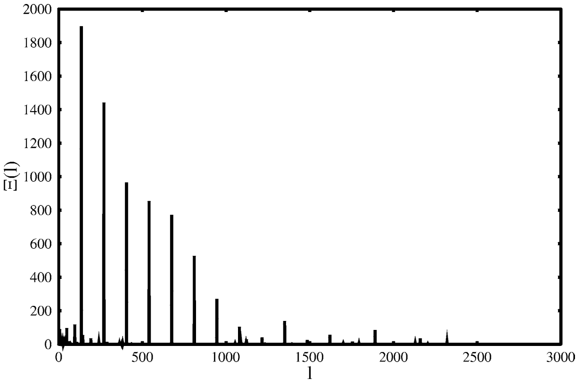

Fig. 4. A plot of correlation intensity $\Xi(l)$ versus correlation distance l calculated from Fig. 3(b).

to 26661, and the third one exists at position 26696 to 26796. Their correlation distances are 810 and 945, which are 6 and 7 times of the basic correlation distance. For the square set 2, in the same way, a 101-nucleotide string $g^2t\ldots tac$ is copied three times in its neighboring region. The first one occurs in the range from 204567 to 204667, the second one occurs in the range from 204702 to 204802, and the third one occurs in the range from 205512 to 205612. Their correlation distances are 135 and 945, which are equal to the basic correlation distance and its 7 times. In Table 2, we present the transference of $k(\geqslant 90)$-nucleotide strings and their correspondent positions in Yeast1. In the same way, we take the cutoff threshold of 90 nucleotides. The repeated longest nucleotide string $tga\ldots aca$ has 337 letters and the correlation distance 5588. It distributes in the ranges of 160239 to 160575 and 165827 to 166163. Many long nucleotide strings have correlation distances, which are integer times of the basic correlation distance.

Fig. 4 displays the correlation intensity $\Xi(l)$ at different correlation distance l with $k = 15$. The global behavior is preserved even when k is changed to 11. In a difference from Fig. 2, there exist some parallel lines with the same distance. The maximal correlation intensity appears at $l = 135$. Then, when the correlation distance increases with 135, the correlation intensity arrives at a local maximum. In the global properties, the local maximal correlation intensity decreases monotonically. The properties give an evidence that in the transference of nucleotide strings of Yeast1, its correlation distance has a constant increasing.

4. Conclusion

In summary, by using metric representation, recurrence plot technique is extended to analyze correlation structure of nucleotide strings in DNA sequences. In the correlation structure, some diagonal parallel lines correspond to global and local transference of nucleotide strings in the sequences. The correlation structure is quantified by correlation intensity, which can be used to display transference of nucleotide strings in the DNA sequences. It is found that, in the transference of nucleotide strings, HUMHBB has a major correlation distance, but Yeast1's correlation distance has a constant increasing.

Acknowledgements

This work was supported in part by the National Key Program for Developing Basic Science G1999-032801-11. The author thanks the referee for valuable suggestions.

References

[1] H.J. Jeffrey, Nucleic Acids Res. 18 (1990) 2163.
[2] P.J. Deschavanne, A. Giron, J. Vilain, G. Fagot, B. Fertil, Mol. Biol. Evol. 16 (1999) 1391.
[3] B.-L. Hao, H.C. Lee, S.-Y. Zhang, Chaos Solitons Fractals 11 (2000) 825.
[4] J. Qi, B. Wang, B.-L. Hao, J. Mol. Evol. 58 (2004) 1.
[5] Z.-B. Wu, Electrophoresis 21 (2000) 2321.
[6] Z.-B. Wu, Fractals 11 (2003) 19.
[7] J.D. Farmer, Physica D 4 (1982) 366.
[8] J.-P. Eckmann, S.O. Kamphorst, D. Ruelle, Europhys. Lett. 4 (1987) 973.
[9] C.L. Webber, J.P. Zbilut, J. Appl. Physiol. 94 (1994) 965.
[10] P. Faure, H. Korn, Proc. Natl. Acad. Sci. USA 94 (1997) 6506.
[11] J.P. Zbilut, A. Giuliani, C.L. Webber, A. Colosimo, Protein Eng. 11 (1998) 87.
[12] H. Ochman, J.G. Lawrence, E.A. Groisman, Nature 405 (2000) 299.
[13] J.L. Bennetzen, Plant Mol. Biol. 42 (2000) 251.