Physica A 389 (2010) 5698–5705

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Global transposable characteristics in the complete DNA sequence of the yeast

Zuo-Bing Wu

State Key Laboratory of Nonlinear Mechanics, Institute of Mechanics, Chinese Academy of Sciences, Beijing 100190, China

ARTICLE INFO

Article history:
Received 4 July 2010
Received in revised form 16 August 2010
Available online 15 September 2010

Keywords:
Yeast
DNA sequences
Coherence structure
Metric representation
Recurrence plot

ABSTRACT

Global transposable characteristics in the complete DNA sequence of the Saccharomyces cevevisiae yeast is determined by using the metric representation and recurrence plot methods. On the basis of the correlation distance of nucleotide strings, 16 chromosome sequences of the yeast, which are divided into 5 groups, display 4 kinds of the fundamental transposable characteristics: a short increasing period, a long increasing quasi-period, a long major value and hardly relevant.

1. Introduction

The recent complete DNA sequences of many organisms are available for systematic search of genome structures. For a large amount of DNA sequences, developing quantitative methods for the extraction of meaningful information is a major challenge in bioinformatics, which paves the way for a new branch of application of statistical mechanics to biological systems [1]. To understand the one-dimensional symbolic sequences composed of the four letters 'A', 'C', 'G' and 'T' (or 'U'), some statistical and geometrical methods were developed [2–12]. The outcome of these studies on DNA sequences of many organisms has resulted in determining the nontrivial statistical characteristics, such as long-range correlations, short-range correlations and fractal features. In particular, chaos game representation (CGR) [13], 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 images were described as genomic signatures, and the classification of species in the whole bacteria genome was analyzed by making a Euclidean metric between two CGR images [14]. Based on the genomic signature, the distance between two DNA sequences, depending on the length of nucleotide strings, was presented [15] and the horizontal transfers in prokaryotes and eukaryotes were detected and characterized [16,17].

In general, symbolic dynamics and recurrence plots are basic methods for analyzing complex systems [18,19]. Although conventional science has made great strides in understanding genetic patterns, they are required to analyze the so-called junk DNA with complex functions, governing mutations [20]. Recently, a one-to-one metric representation of the DNA sequence [21], which was borrowed from symbolic dynamics, makes an ordering of subsequences in a plane. The suppression of certain nucleotide strings in the DNA sequences leads to a self-similarity of pattern that is seen in the metric representation of DNA sequences. The self-similarity limits of genomic signatures were determined as an optimal string length for generating genomic signatures [22]. Moreover, by using the metric representation method, the recurrence plot technique of DNA sequences was established, and employed to analyze the correlation structure of nucleotide strings [23].

As a eukaryotic organism, yeast is one of the premier industrial microorganisms, because of its essential role in brewing, baking, and fuel alcohol production. In addition, yeast has proven to be an excellent model organism for the study of a

E-mail address: wuzb@lnm.imech.ac.cn.

0378-4371/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2010.08.026

variety of biological problems involving the fields of genetics, molecular biology, cell biology and other disciplines within the biomedical and life sciences. In April 1996, a complete DNA sequence of the yeast (Saccharomyces cevevisiae) genome, consisting of 16 chromosomes with 12 million basepairs, had been released to provide a resource of genome information of a single organism. However, only 43.3% of all 6000 predicted genes in the Saccharomyces cevevisiae yeast were functionally characterized, even though a complete sequence of the yeast genome was available [24]. Moreover, it was found that DNA transposable elements have the ability to move from one place to another and make many replicas within the genome via the transposition [25,26]. Therefore, the complete DNA sequence of the yeast remains a topic to be studied, with respect to its genome architecture structure in the whole sequence.

In this paper, using the metric representation and recurrence plot methods, we analyze global transposable characteristics in the complete DNA sequence of the yeast, i.e., 16 chromosome sequences. It is conducive to understanding the complexity of the DNA sequence (global and local statistical features) and exploring possible biological functions (heredity and variation).

2. Metric representation and recurrence plot methods

For a given DNA sequence $s_1 s_2 \cdots s_i \cdots s_N$ ($s_i \in \{A, C, G, T\}$), a plane metric representation is generated by making the correspondence of symbol s_i to a number μ_i or $\nu_i \in \{0, 1\}$ and calculating values (α_k, β_k) of all subsequences $\Sigma_k = s_1 s_2 \cdots s_k$ ($1 \le k \le N$) defined as follows

$$\begin{aligned} \alpha_k &= 2\sum_{j=1}^{k} \mu_{k-j+1} 3^{-j} + 3^{-k} = 2\sum_{i=1}^{k} \mu_i 3^{-(k-i+1)} + 3^{-k}, \\ \beta_k &= 2\sum_{j=1}^{k} \nu_{k-j+1} 3^{-j} + 3^{-k} = 2\sum_{i=1}^{k} \nu_i 3^{-(k-i+1)} + 3^{-k}, \end{aligned} \tag{1}$$

where μ_i is 0 if $s_i \in \{A, C\}$ or 1 if $s_i \in \{G, T\}$ and ν_i is 0 if $s_i \in \{A, T\}$ or 1 if $s_i \in \{C, G\}$. Thus, the one-dimensional symbolic sequence is partitioned into N subsequences Σ_k and mapped in the two-dimensional plane (α_k, β_k). Subsequences with the same ending l-nucleotide string (a constant string $s_{k-l+1} \cdots s_k$ in Σ_k), which are labeled by Σ^l, correspond to points in the zone encoded by the l-nucleotide string. Taking a subsequence $\Sigma_i \in \Sigma^l$, we calculate

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

where Θ is the Heaviside function [$\Theta(x) = 1$, if $x > 0$; $\Theta(x) = 0$, if $x \le 0$] and Σ_j is a subsequence ($j \ge l$). When $\Theta(\epsilon_l - |\Sigma_i - \Sigma_j|) = 1$, i.e., $\Sigma_j \in \Sigma^l$, a point (i, j) is plotted in a plane. Thus, repeating the above process from the beginning of one-dimensional symbolic sequence and shifting forward, we obtain a recurrence plot of the DNA sequence. In Eq. (1), the metric representation is constructed on the basis of the scaling rate 1/3, so the correspondent zone size ϵ_l is 3^{-l}. The maximal length of the available l-nucleotide strings depends on the type of real variables and the computer-aided accuracy. For example, $\epsilon_{11} = 5.6 \times 10^{-6}$, $\epsilon_{15} = 7.0 \times 10^{-8}$ and $\epsilon_{30} = 4.9 \times 10^{-15}$. If $l = 11$ or 15 or 30 is chosen, the computer should be able to identify the number ϵ_{11} or ϵ_{15} or ϵ_{30}.

For presenting a correlation structure in the recurrence plot plane, a correlation intensity is defined at a given correlation distance d

$$\Xi(d) = \sum_{i=1}^{N-d} \Theta(\epsilon_l - |\Sigma_i - \Sigma_{i+d}|). \tag{3}$$

The quantity displays the transference of l-nucleotide strings in the DNA sequence. To further determine positions and lengths of the transposable elements, we analyze the recurrent plot plane. From $\Theta(\epsilon_l - |\Sigma_i - \Sigma_j|) = 1$, we have Σ_i and $\Sigma_j \in \Sigma^l$, where $\Sigma_i = s_1 s_2 \cdots s_{i-l+1} \cdots s_i$ and $\Sigma_j = s_1 s_2 \cdots s_{j-l+1} \cdots s_j$. The same ending l-nucleotide string $s_{i-l+1} \cdots s_i$ or $s_{j-l+1} \cdots s_j$ is the starting part of a transposable element, so the transposable element has the length l at least. From the recurrence plot plane, we calculate the maximal value of x to satisfy

$$\Theta(\epsilon_l - |\Sigma_{i+x} - \Sigma_{j+x}|) = 1, \quad x = 0, 1, 2, \ldots, x_{\max}, \tag{4}$$

i.e., Σ_{i+x} and $\Sigma_{j+x} \in \Sigma^l$. In Σ_{i+x} and Σ_{j+x}, the same ending l-nucleotide string $s_{i+x-l+1} \cdots s_{i+x}$ or $s_{j+x-l+1} \cdots s_{j+x}$ may be different from the above one. Finally, the same ending l-nucleotide string $s_{i+x_{\max}-l+1} \cdots s_{i+x_{\max}}$ or $s_{j+x_{\max}-l+1} \cdots s_{j+x_{\max}}$ is the closing part of the transposable element. Thus, the transposable element has the length $L = l + x_{\max}$ and is placed at the positions $(i - l + 1, i + x_{\max})$ and $(j - l + 1, j + x_{\max})$. Since the element $s_{i-l+1} \cdots s_{i+x_{\max}}$ or $s_{j-l+1} \cdots s_{j+x_{\max}}$ is moved from the position $(i - l + 1, i + x_{\max})$ to the position $(j - l + 1, j + x_{\max})$, the correction distance is $d = (j - l + 1) - (i - l + 1) = (j + x_{\max}) - (i + x_{\max}) = j - i$.

Table 1
Transference of nucleotide strings with lengths $L(\geq 100)$ for YEAST I with 230 209 bases.

No.	String	Position 1	Position 2	L	d	Note
1	$t^2a\cdots act$	11 745–11 969	24 177–24 401	225	12 432	
2	$ctg\cdots a^2t$	12 258–12 396	24 711–24 849	139	12 453	
3	$g^2a\cdots g^2a$	12 988–13 171	25 153–25 336	184	12 165	
4	$c^2g\cdots cgt$	25 715–25 851	26 255–26 391	137	540	$4d_b$
5	$at^2\cdots ac^2$	25 739–25 851	26 414–26 526	113	675	$5d_b$
6	$gta\cdots ac^2$	25 751–25 851	26 561–26 661	101	810	$6d_b$
7	$gta\cdots ac^2$	25 751–25 851	26 696–26 796	101	945	$7d_b$
8	$t^2g\cdots g^2t$	25 853–25 968	26 393–26 508	116	540	$4d_b$
9	$atg\cdots gtg$	25 925–26 035	26 060–26 170	111	135	d_b
10	$atg\cdots agt$	25 925–26 058	26 195–26 328	134	270	$2d_b$
11	$agt\cdots gtg$	26 050–26 170	26 185–26 305	121	135	d_b
12	$at^2\cdots gac$	26 279–26 406	26 414–26 541	128	135	d_b
13	$gta\cdots gac$	26 291–26 406	26 561–26 676	116	270	$2d_b$
14	$gta\cdots gac$	26 291–26 406	26 696–26 811	116	405	$3d_b$
15	$gta\cdots gtg$	26 426–26 710	26 561–26 845	285	135	d_b
16	$tga\cdots aca$	160 239–160 575	165 827–166 163	337	5 588	
17	$cac\cdots tac$	204 518–204 802	204 653–204 937	285	135	d_b
18	$g^2t\cdots tac$	204 567–204 667	205 512–205 612	101	945	$7d_b$
19	$g^2t\cdots tac$	204 702–204 802	205 512–205 612	101	810	$6d_b$
20	$g^2t\cdots a^2t$	204 837–204 949	205 512–205 624	113	675	$5d_b$
21	$ac^2\cdots c^2a$	204 855–204 969	205 395–205 509	115	540	$4d_b$
22	$ctc\cdots cat$	205 042–205 168	205 312–205 438	127	270	$2d_b$
23	$cac\cdots act$	205 058–205 178	205 193–205 313	121	135	d_b
24	$cac\cdots cat$	205 193–205 303	205 328–205 438	111	135	d_b
25	$atg\cdots t^2c$	205 758–205 879	206 433–206 554	122	675	$5d_b$

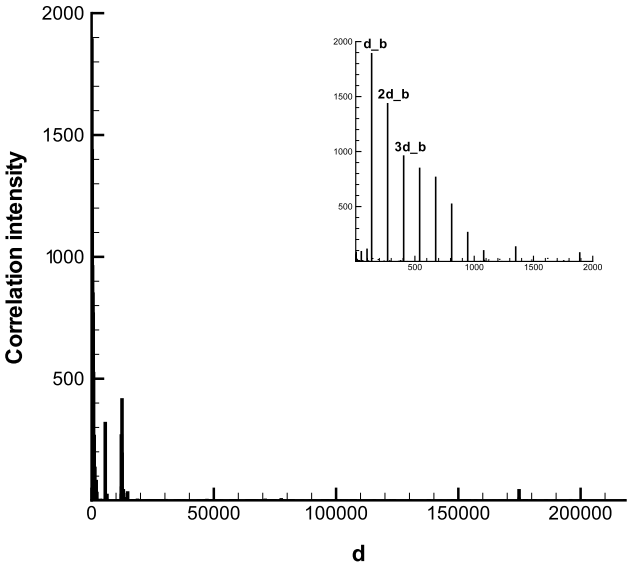

Fig. 1. A plot of correlation intensity $\Xi(d)$ versus correlation distance d for YEAST I.

3. Global transposable characteristics in the complete DNA sequence of the yeast

The Saccharomyces cevevisiae yeast has 16 chromosomes, which are denoted as YEAST I to XVI. Using the metric representation and recurrence plot methods, we analyze correlation structures of 16 DNA sequences. In accordance with the characteristics of the correlation structures, we summarize the results as follows:

(1) The correlation distance has a short increasing period. YEAST I, IX and XI have such characteristics. Let us take YEAST I as an example to analyze. Fig. 1 displays the correlation intensity at different correlation distance $d(\leq N - l)$ with $l = 15$. A local region is magnified in the figure. It is clearly evident that there exist some equidistant parallel lines with a basic correlation distance $d_b = 135$. Using Eq. (4), we determine positions and lengths of the transposable elements in Table 1, where their lengths are limited to $L \geq 100$. Many nucleotide strings have a correlation distance, which is an integral multiple of d_b. They mainly distribute in two local regions of the DNA sequence (25 715–26 845) and (204 518–206 554) or (11.2%–11.7%) and (88.8%–89.7%) expressed in percentage. YEAST IX and XI have similar behaviors. YEAST IX has a basic

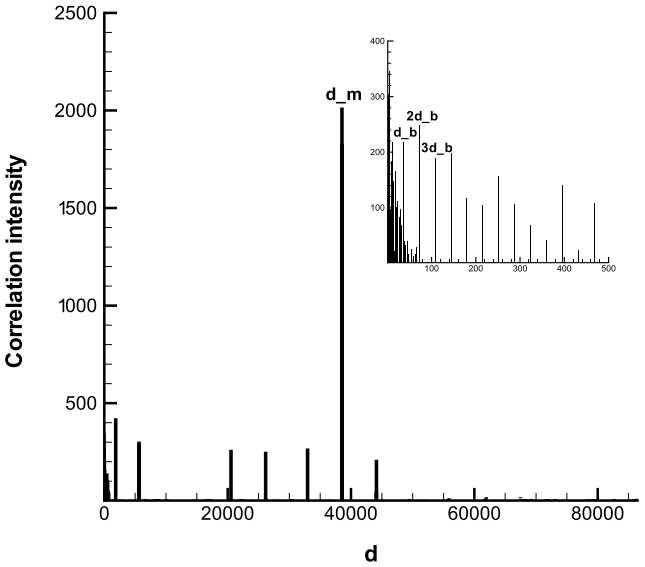

Fig. 2. A plot of correlation intensity $\Xi(d)$ versus correlation distance d for YEAST II.

Table 2
Transference of nucleotide strings with lengths $L(\geq 50)$ for YEAST II with 813 142 bases. Due to the limited spacing, 22 nucleotide strings out of a total number 57 are not presented.

No.	String	Position 1	Position 2	L	d	Note
1	$tag\cdots agt$	1584–1650	1692–1758	67	108	$3d_b$
2	$tag\cdots gct$	1620–1674	2016–2070	55	396	$11d_b$
3	$tag\cdots gct$	1620–1674	2268–2322	55	648	$18d_b$
4	$tgc\cdots tg^2$	1815–1870	1851–1906	56	36	d_b
5	$gct\cdots gta$	1860–1936	1932–2008	77	72	$2d_b$
6	$tgc\cdots gta$	1887–1936	2355–2404	50	468	$13d_b$
7	$tgc\cdots gtg$	1959–2018	2355–2414	60	396	$11d_b$
8	$gca\cdots agt$	2012–2082	2264–2334	71	252	$7d_b$
9	$tg^2\cdots agt$	2022–2190	2094–2262	169	72	$2d_b$
10	$cag\cdots gta$	2037–2104	2325–2392	68	288	$8d_b$
11	$tg^2\cdots agt$	2094–2154	2274–2334	61	180	$5d_b$
12	$cag\cdots gta$	2109–2176	2325–2392	68	216	$6d_b$
13	$tg^2\cdots agt$	2166–2226	2274–2334	61	108	$3d_b$
14	$cag\cdots gta$	2181–2248	2325–2392	68	144	$4d_b$
33	$a^2t\cdots c^2t$	221 182–221 231	259 718–259 767	50	38 536	$\approx d_m$
35	$gta\cdots tct$	221 249–221 308	259 783–259 842	60	38 534	d_m
37	$gat\cdots tca$	222 669–222 827	261 203–261 361	159	38 534	d_m
38	$ta^2\cdots tgc$	222 829–223 096	261 363–261 630	268	38 534	d_m
39	$cat\cdots gct$	223 098–223 609	261 632–262 143	512	38 534	d_m
40	$a^2c\cdots g^2a$	223 611–223 966	262 145–262 500	356	38 534	d_m
41	$tc^2\cdots c^2g$	223 968–224 563	262 502–263 097	596	38 534	d_m
42	$ga^2\cdots tac$	224 736–225 308	263 273–263 845	573	38 537	$\approx d_m$
43	$ata\cdots aca$	225 310–225 493	263 847–264 030	184	38 537	$\approx d_m$
44	$ac^2\cdots a^2c$	225 841–226 231	264 378–264 768	391	38 537	$\approx d_m$
45	$acg\cdots t^3$	226 233–226 797	264 770–265 334	565	38 537	$\approx d_m$

correlation distance $d_b = 18$. Many nucleotide strings ($L \geq 50$) with an integral multiple of d_b mainly distribute in a local region of the DNA sequence (391 337–393 583) or (89.0%–89.5%) expressed in percentage. YEAST XI has a basic correlation distance $d_b = 189$. Many nucleotide strings ($L \geq 50$) with an integral multiple of d_b mainly distribute in a local region of the DNA sequence (647 101–647 783) or (97.1%–97.2%) expressed in percentage.

(2) The correlation distance has a long major value and a short increasing period. YEAST II, V, VII, VIII, X, XII, XIII, XIV, XV and XVI have such characteristics. Let us take YEAST II as an example to analyze. Fig. 2 displays the correlation intensity at different correlation distances $d(\leq N - l)$ with $l = 15$. The maximal correlation intensity appears at correlation distances $d_m = 38\,534$. A local region is magnified in the figure. It is clearly evident that there exist some equidistant parallel lines with a basic correlation distance $d_b = 36$. In Table 2, positions and lengths ($L \geq 50$) of the transposable elements are given. The maximal transposable elements mainly distribute in two local regions of the DNA sequence (221 249–224 565, 259 783–263 097) or (27.2%–27.6%, 31.9%–32.4%) expressed in percentage. Near the positions, there also exist some

Table 3
Transference of nucleotide strings with lengths $L(\geq 50)$ for YEAST III with 315 341 bases.

No.	String	Position 1	Position 2	L	d	Note
1	$c^3\cdots ac^2$	90–141	233–284	52	143	
2	$a^2t\cdots ca^2$	1190–1253	4083–4146	64	2 893	
3	$t^2g\cdots gta$	11 499–11 764	197 402–197 667	266	185 903	d_{m1}
4	$tag\cdots cta$	11 766–11 908	197 669–197 811	143	185 903	d_{m1}
5	$ta^2\cdots gac$	11 910–12 185	197 813–198 088	276	185 903	d_{m1}
6	$at^2\cdots gtg$	12 187–13 810	198 090–199 713	1624	185 903	d_{m1}
7	$tac\cdots tat$	12 268–12 340	291 794–291 866	73	279 526	$\approx d_{m3}$
8	$ata\cdots at^2$	12 325–12 932	291 853–292 460	608	279 528	d_{m3}
9	$a^2t\cdots gtg$	13 691–13 810	293 114–293 233	120	279 423	
10	$cg^2\cdots ca^2$	13 812–14 006	199 715–199 909	195	185 903	d_{m1}
11	$cg^2\cdots t^2c$	13 812–13 893	293 235–293 316	82	279 423	
12	$tgt\cdots a^2c$	83 677–83 802	84 419–84 544	126	742	
13	$tgt\cdots a^2c$	83 677–83 802	90 049–90 174	126	6 372	
14	$cta\cdots a^2c$	83 724–83 802	83 951–84 029	79	227	
15	$ca^2\cdots tg^2$	83 804–83 902	84 031–84 129	99	227	
16	$ca^2\cdots tg^2$	83 804–83 902	84 546–84 644	99	742	
17	$ca^2\cdots tg^2$	83 804–83 902	90 176–90 274	99	6 372	
18	$cta\cdots tct$	83 951–84 230	84 466–84 745	280	515	
19	$cta\cdots tat$	83 951–84 189	90 096–90 334	239	6 145	
20	$tgt\cdots tat$	84 419–84 704	90 049–90 334	286	5 630	
21	$cac\cdots t^3$	123 942–124 058	142 661–142 777	117	18 719	
22	$tac\cdots tat$	198 171–198 243	291 794–291 866	73	93 623	$\approx d_{m2}$
23	$ata\cdots at^2$	198 228–198 835	291 853–292 460	608	93 625	d_{m2}
24	$a^2t\cdots t^2c$	199 594–199 796	293 114–293 316	203	93 520	$\approx d_{m2}$
25	$g^2a\cdots g^3$	267 365–267 574	267 692–267 901	210	327	

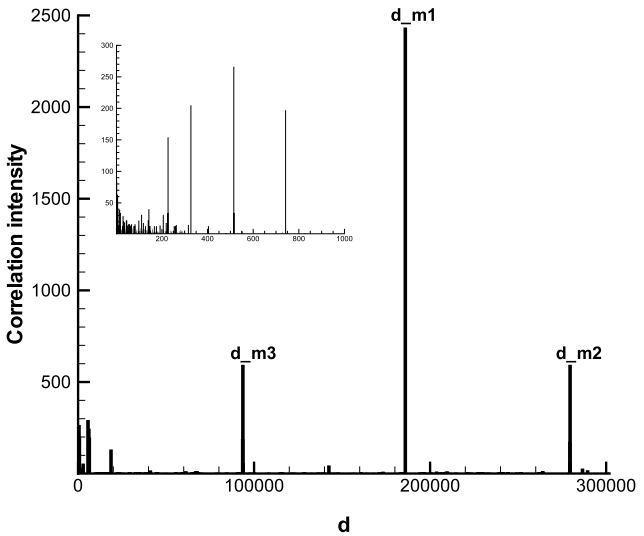

Fig. 3. A plot of correlation intensity $\Xi(d)$ versus correlation distance d for YEAST III.

transposable elements with approximate values for d_m. Moreover, many nucleotide strings have correlation distances, which are an integral multiples of d_b. They mainly distribute in a local region of the DNA sequence (391 337–393 583) or (89.0%–89.5%) expressed in percentage. In the other 9 DNA sequences, YEAST V, X, XII, XIII, XIV, XV and XVI have the same basic correlation distance $d_b = 36$, and a similar behavior with different major correlation distances $d_m = 49\,099$, 5584, 9137, 12 167, 5566, 447 110 and 45 988, respectively. YEAST VII and VIII have different basic correlation distances $d_b = 12$ and 135, and a similar behavior with the major correlation distances $d_m = 255\,548$ and 1998, respectively.

(3) The correlation distance has a long increasing quasi-period. YEAST III has such characteristics. Fig. 3 displays the coherence intensity at different correlation distances $d(\leq N - l)$ with $l = 15$. The correlation intensity has the maximal value at the correlation distances $d_{m1} = 185\,903$ and two vice-maximal values at the correlation distances $d_{m2} = 93\,625$ and $d_{m3} = 279\,528$. Since $d_{m2} \approx d_{m1}/2 \approx d_{m3}/3$, the coherence distance has an increasing quasi-period. A local region is magnified in the figure. There does not exist any clear short increasing period for the correlation distance. Using Eq. (4), we determine positions and lengths ($L \geq 50$) of the transposable elements in Table 3. The maximal and vice-maximal transposable elements mainly distribute in local regions of the DNA sequence (11 499–13 810, 197 402–199 713),

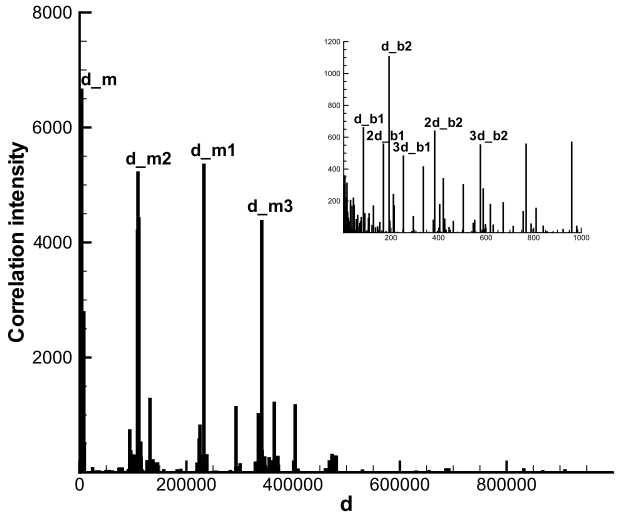

Fig. 4. A plot of correlation intensity $\Xi(d)$ versus correlation distance d for YEAST IV.

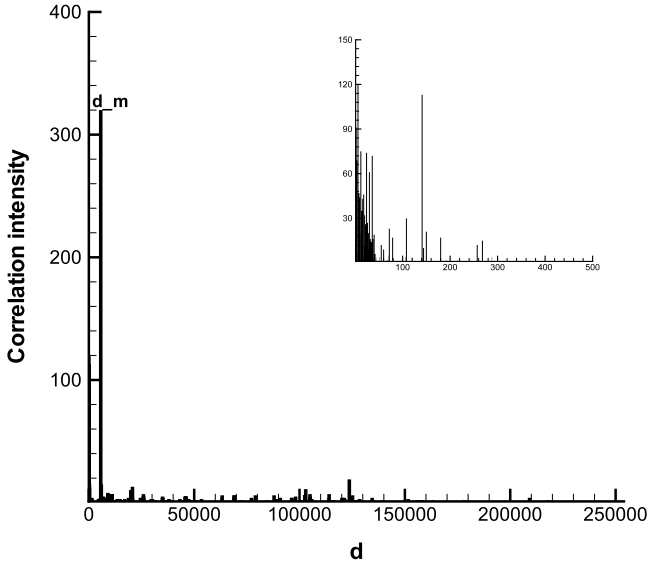

Fig. 5. A plot of correlation intensity $\Xi(d)$ versus correlation distance d for YEAST VI.

(198 171–199 796, 291 794–293 316) and (12 268–12 932, 291 794–292 460) or (3.6%–4.4%, 62.6%–63.6%), (62.8%–63.4%, 92.5%–93.0%) and (3.9%–4.1%, 92.5%–92.7%) expressed in percentage.

(4) The correlation distance has a long major value and a long increasing quasi-period and two short increasing periods. YEAST IV has such characteristics. Fig. 4 displays the coherence intensity at different correlation distances $d(\leq N - l)$ with $l = 15$. The maximal coherence intensity appears at the correlation distance $d_m = 3885$. There also exist three vice-maximal values at the correlation distances $d_{m1} = 232\,800$, $d_{m2} = 109\,349$ and $d_{m3} = 341\,221$, which forms a long increasing quasi-period of the correlation distance, i.e., $d_{m2} \approx d_{m1}/2 \approx d_{m3}/3$. A local region is magnified in the figure. It is clearly evident that there exist two short increasing periods with $d_{b1} = 84$ and $d_{b2} = 192$ in the correlation distance. In Table 4, positions and lengths ($L \geq 100$) of the transposable elements are determined by using Eq. (4). All correlation distances with the long major value and the long increasing quasi-period and two short increasing periods are denoted. The transposable elements with d_m, d_{m1}, d_{m2}, d_{m3}, d_{b1} and d_{b2} mainly distribute in local regions of the DNA sequence (527 570–538 236), (871 858–876 927, 981 207–986 276), (645 646–651 457, 878 346–884 257), (646 379–651 032, 987 600–992 253), (1 307 733–1 308 591) and (758 135–759 495) or (34.4%–35.1%), (56.9%–57.2%, 64.0%–64.4%), (42.1%–42.5%, 57.3%–57.7%), (42.2%–42.5%, 64.4%–64.8%), (85.36%–85.41%) and (49.5%–49.6%) expressed in percentage.

(5) The DNA sequence is hardly relevant. YEAST VI has such characteristics. Fig. 5 displays the coherence intensity at different correlation distances $d(\leq N - l)$ with $l = 15$. The maximal coherence intensity appears at the correlation distance $d_m = 5627$. A local region is magnified in the figure. The sequence has not a short increasing period for the coherence

Table 4
Transference of nucleotide strings with lengths $L(\geq 100)$ for YEAST IV with 1531977 bases. Due to the limited spacing, 112 nucleotide strings without any notes in the total number 176 are not presented.

No.	String	Position 1	Position 2	L	d	Note
28	$gct\cdots a^3$	527 570–527 835	531 455–531 720	266	3 885	d_m
29	$gct\cdots a^2t$	527 570–527 781	535 340–535 551	212	7 770	
30	$gt^2\cdots c^2a$	527 891–528 066	531 776–531 951	176	3 885	d_m
31	$gt^2\cdots c^2a$	527 891–528 066	535 661–535 836	176	7 770	
32	$ag^2\cdots a^2t$	528 094–531 666	531 979–535 551	3573	3 885	d_m
35	$ct^2\cdots t^2g$	531 668–532 364	535 553–536 249	697	3 885	d_m
36	$gca\cdots c^2a$	532 366–534 351	536 251–538 236	1986	3 885	d_m
40	$ata\cdots agt$	645 546–646 118	878 346–878 918	573	232 800	d_{m2}
41	$ag^2\cdots aca$	645 635–645 873	992 440–992 678	239	346 805	
42	$agc\cdots atc$	646 120–646 336	878 920–879 136	217	232 800	d_{m2}
43	$ct^2\cdots gta$	646 338–646 726	879 138–879 526	389	232 800	d_{m2}
44	$t^3\cdots atg$	646 379–646 564	987 600–987 785	186	341 221	d_{m3}
45	$t^2c\cdots cat$	646 787–647 179	879 587–879 979	393	232 800	d_{m2}
46	$tat\cdots t^2c$	646 964–647 288	988 185–988 509	325	341 221	d_{m3}
47	$t^3\cdots ta^2$	647 310–647 470	880 110–880 270	161	232 800	d_{m2}
48	$ta^2\cdots tga$	647 468–647 693	988 689–988 914	226	341 221	d_{m3}
49	$gt^2\cdots cgt$	647 472–648 042	880 272–880 842	571	232 800	d_{m2}
50	$gag\cdots aga$	647 695–649 519	988 916–990 740	1825	341 221	d_{m3}
51	$gat\cdots cg^2$	648 165–648 265	880 965–881 065	101	232 800	d_{m2}
52	$tc^2\cdots atg$	648 486–649 715	881 286–882 515	1230	232 800	d_{m2}
53	$cgt\cdots atg$	649 521–649 715	990 742–990 936	195	341 221	d_{m3}
54	$ctg\cdots t^3$	649 853–650 096	991 074–991 317	244	341 221	d_{m3}
55	$tag\cdots ctc$	649 946–650 073	882 746–882 873	128	232 800	d_{m2}
56	$gat\cdots aca$	650 075–651 457	882 875–884 257	1383	232 800	d_{m2}
57	$ctg\cdots ct^2$	650 098–651 032	991 319–992 253	935	341 221	d_{m3}
60	$ag^2\cdots aca$	651 219–651 457	992 440–992 678	239	341 221	d_{m3}
63	$ac^2\cdots c^2a$	757 478–757 581	757 670–757 773	104	192	d_{b2}
68	$cta\cdots act$	758 135–758 259	758 519–758 643	125	384	$2d_{b2}$
69	$cta\cdots act$	758 135–758 259	758 711–758 835	125	576	$3d_{b2}$
70	$cta\cdots act$	758 135–758 259	759 479–759 603	125	1 344	$7d_{b2}$
71	$gca\cdots g^2a$	758 219–758 343	758 411–758 535	125	192	d_{b2}
72	$gca\cdots a^2t$	758 219–758 347	759 179–759 307	129	960	$5d_{b2}$
73	$gca\cdots g^2a$	758 219–758 343	759 371–759 495	125	1 152	$6d_{b2}$
74	$ctg\cdots act$	758 349–758 451	758 925–759 027	103	576	$3d_{b2}$
75	$ctg\cdots g^2a$	758 349–758 535	759 117–759 303	187	768	$4d_{b2}$
76	$ctg\cdots cat$	758 349–758 683	759 309–759 643	335	960	$5d_{b2}$
77	$cta\cdots cat$	758 495–758 683	758 687–758 875	189	192	d_{b2}
78	$cta\cdots aga$	758 687–758 901	759 455–759 669	215	768	$4d_{b2}$
79	$gca\cdots act$	758 795–759 027	758 987–759 219	233	192	d_{b2}
80	$cta\cdots act$	758 903–759 027	759 287–759 411	125	384	$2d_{b2}$
81	$gca\cdots aga$	758 987–759 093	759 563–759 669	107	576	$3d_{b2}$
82	$cta\cdots g^2a$	759 095–759 303	759 287–759 495	209	192	d_{b2}
99	$tgt\cdots a^3$	871 858–872 030	981 207–981 379	173	109 349	d_{m1}
103	$tgc\cdots ca^2$	872 202–872 307	981 551–981 656	106	109 349	d_{m1}
104	$a^2g\cdots cag$	872 309–872 592	981 658–981 941	284	109 349	d_{m1}
105	$at^2\cdots tat$	872 737–872 871	982 086–982 220	135	109 349	d_{m1}
106	$ag^2\cdots ca^2$	873 022–873 286	982 371–982 635	265	109 349	d_{m1}
107	$tga\cdots cat$	873 378–874 087	982 727–983 436	710	109 349	d_{m1}
108	$tca\cdots g^2t$	874 089–874 236	983 438–983 585	148	109 349	d_{m1}
109	$tac\cdots tgc$	874 238–874 593	983 587–983 942	356	109 349	d_{m1}
110	$aga\cdots atc$	874 595–874 764	983 944–984 113	170	109 349	d_{m1}
111	$t^3\cdots ca^2$	874 853–875 247	984 202–984 596	395	109 349	d_{m1}
112	$gat\cdots a^2c$	875 249–875 604	984 598–984 953	356	109 349	d_{m1}
113	$ca^2\cdots tga$	875 637–876 474	984 986–985 823	838	109 349	d_{m1}
114	$agc\cdots tga$	876 491–876 927	985 840–986 276	437	109 349	d_{m1}
117	$g^2a\cdots aga$	877 085–877 385	986 434–986 734	301	109 349	d_{m1}
118	$ga^2\cdots a^3$	877 387–877 657	986 736–987 006	271	109 349	d_{m1}
169	$cgt\cdots ac^2$	1 307 733–1 307 835	1 308 321–1 308 423	103	588	$7d_{b1}$
170	$ac^2\cdots g^2c$	1 307 749–1 307 874	1 308 505–1 308 630	126	756	$9d_{b1}$
171	$gt^2\cdots atc$	1 307 753–1 307 871	1 307 921–1 308 039	119	168	$2d_{b1}$
172	$gt^2\cdots atc$	1 307 921–1 308 039	1 308 509–1 308 627	119	588	$7d_{b1}$
173	$gt^2\cdots atc$	1 308 089–1 308 249	1 308 341–1 308 501	161	252	$3d_{b1}$
174	$ac^2\cdots atc$	1 308 169–1 308 333	1 308 253–1 308 417	165	84	d_{b1}
175	$ac^2\cdots ac^2$	1 308 337–1 308 507	1 308 421–1 308 591	171	84	d_{b1}

Table 5
Transference of nucleotide strings with lengths $L(\geq 50)$ for YEAST VI with 270 148 bases.

No.	String	Position 1	Position 2	L	d	Note
1	$tat\cdots aca$	137 905—138 238	143 532—143 865	334	5627	d_m
2	$ga^2\cdots t^3$	178 016—178 086	178 157—178 227	71	141	
3	$ca^2\cdots gtc$	178 088—178 157	178 229—178 298	70	141	
4	$tgt\cdots gtg$	210 332—210 391	210 334—210 393	60	2	

distance. In Table 5, positions and lengths ($L \geq 50$) of the transposable elements are given. Only one nucleotide string with the length 337 has the correlation distance d_m. YEAST VI is almost irrelevant, so YEAST VI approaches a random sequence.

4. Conclusion

Global transposable characteristics in the complete DNA sequence of the yeast is determined by using the metric representation and recurrence plot methods. Positions and lengths of all transposable nucleotide strings in the 16 chromosome DNA sequences of the yeast are determined. On the basis of correlation distances of nucleotide strings, the fundamental transposable characteristics display a short increasing period, a long increasing quasi-period, a long major value and hardly relevant. The 16 chromosome sequences are divided into 5 groups, which have one or several of the 4 kinds of the fundamental transposable characteristics.

Acknowledgements

We thank the IMECH and SHENTENG 7000 research computing facilities for assisting us in the computation. The work is partially supported by Contract No. KJCX2-YW-H18 of the Chinese Academy of Sciences.

References

[1] L.T. Wille, New Directions in Statistical Physics: Econophysics, Bioinformatics, and Pattern Recognition, Springer Press, 2004.
[2] C.K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H.E. Stanley, Long-range correlations in nucleotide sequence, Nature 356 (1992) 168.
[3] B.-L. Hao, Fractals from genomes—exact solutions of a biology-inspired problem, Physica A 282 (2000) 225.
[4] D. Qi, A.J. Cuticchia, Compositional symmetries in complete genomes, Bioinformatics 17 (2001) 557.
[5] P. Katsaloulis, T. Theoharis, A. Provata, Statistical distributions of oligonucleotide combinations: applications in human chromosomes 21 and 22, Physica A 316 (2002) 380.
[6] D. Holste, I. Grosse, S. Beirer, P. Schieg, H. Herzel, Repeats and correlations in human DNA sequences, Phys. Rev. E 67 (2003) 061913.
[7] S. Garte, Fractal properties of the human genome, J. Theoret. Biol. 230 (2004) 251.
[8] P.W. Messer, P.F. Arndt, M. Lässig, Solvable sequence evolution models and genomic correlations, Phys. Rev. Lett. 94 (2005) 138103.
[9] P. Katsaloulis, T. Theoharis, W.-M. Zheng, B.-L. Hao, A. Bountis, Y. Almirantis, A. Provata, Long-range correlations of RNA polymerase II promoter sequences across organisms, Physica A 366 (2006) 308.
[10] C. Vaillant, B. Audit, A. Arneodo, Experiments confirm the influence of genome long-range correlations on nucleosome positioning, Phys. Rev. Lett. 99 (2007) 218103.
[11] T. Oikonomou, A. Provata, U. Tirnakli, Nonextensive statistical approach to non-coding human DNA, Physica A 387 (2008) 2653.
[12] J.S. Almeida, S. Vinga, Biological sequences as pictures—a generic two dimensional solution for iterated maps, BMC Bioinformatics 10 (2009) 100.
[13] H.J. Jeffrey, Chaos game representation of gene structure, Nucleic Acids Res. 18 (1990) 2163.
[14] P.J. Deschavanne, A. Giron, J. Vilain, G. Fagot, B. Fertil, Genomic signature: characterization and classification of species assessed by chaos game representation of sequences, Mol. Biol. Evol. 16 (1999) 1391.
[15] Y. Wang, K. Hill, S. Singh, L. Kari, The spectrum of genomic signatures: from dinucleotides to chaos game representation, Gene 346 (2005) 173.
[16] C. Dufraigne, B. Feitil, S. Lespinats, A. Giron, P. Deschavanne, Detection and characterization of horizontal transfers in prokaryotes using genomic signature, Nucleic Acids Res. 33 (2005) e6.
[17] L.V. Mallet, J. Becq, P. Deschavanne, Whole genome evaluation of horizontal transfers in the pathogenic fungus Aspergillus fumigatus, BMC Genomics 11 (2010) 171.
[18] B.-L. Hao, W.-M. Zheng, Applied Symbolic Dynamics and Chaos, World Scientific, Singapore, 1998.
[19] N. Marwan, M.C. Romano, M. Thiel, J. Kurths, Recurrence plots for the analysis of complex systems, Phys. Rep. 438 (2007) 237.
[20] S.J. Guastello, Progress of applied nonlinear dynamics: welcome to NDPLS Volume 8, Nonlinear Dynam. Psychol. Life Sci. 8 (2004) 1.
[21] Z.-B. Wu, Metric representation of DNA sequences, Electrophoresis 21 (2000) 2321.
[22] Z.-B. Wu, Self-similarity limits of genomic signatures, Fractals 11 (2003) 19.
[23] Z.-B. Wu, Recurrence plot analysis of DNA sequences, Phys. Lett. A 232 (2004) 250.
[24] H.W. Mewes, K. Albermann, M. Bähr, et al., Overview of the yeast genome, Nature 387 (1997) 7.
[25] H. Ochman, J.G. Lawrence, E.A. Groisman, Lateral gene transfer and the nature of bacterial innovation, Nature 405 (2000) 299.
[26] J.L. Bennetzen, Transposable element contributions to plane gene and genome evolution, Plant Mol. Biol. 42 (2000) 251.