Available online at www.sciencedirect.com

SciVerse ScienceDirect

Genomics Proteomics Bioinformatics 10 (2012) 230–236

GENOMICS PROTEOMICS & BIOINFORMATICS

www.elsevier.com/locate/gpb

Application Note

Wavelet Analysis of DNA Walks on the Human and Chimpanzee MAGE/CSAG-palindromes

Yanjiao Qi [1,*], Nengzhi Jin [2], Duiyuan Ai [3]

[1] *Department of Chemical Engineering, Northwest University for Nationalities, Lanzhou 730030, China*
[2] *Gansu Computing Center, Lanzhou 730000, China*
[3] *Gansu Agricultural University, Lanzhou 730070, China*

Received 24 November 2011; revised 18 February 2012; accepted 2 March 2012
Available online 6 August 2012

Abstract

The palindrome is one class of symmetrical duplications with reverse complementary characters, which is widely distributed in many organisms. Graphical representation of DNA sequence provides a simple way of viewing and comparing various genomic structures. Through 3-D DNA walk analysis, the similarity and differences in nucleotide composition, as well as the evolutionary relationship between human and chimpanzee MAGE/CSAG-palindromes, can be clearly revealed. Further wavelet analysis indicated that duplicated segments have irregular patterns compared to their surrounding sequences. However, sequence similarity analysis suggests that there is possible common ancestor between human and chimpanzee MAGE/CSAG-palindromes. Based on the specific distribution and orientation of the repeated sequences, a simple possible evolutionary model of the palindromes is suggested, which may help us to better understand the evolutionary course of the genes and the symmetrical sequences.

Keywords: Palindrome; DNA walks; Wavelet analysis; Phylogenic relationship

Introduction

In nature, symmetry can be found everywhere, including macroscopic and microcosmic objects. The palindrome, as one of the symmetrical sequences, consists of two arms of similar DNA-with one inverted and complemented relative to the other around a central point, usually nonhomologous spacer. Previous studies report that palindromic sequences are frequently observed and important for the structure and/or function of several classes of proteins [1,2]. However, extracting statistical features within the palindrome still needs to be further explored, and this may improve understanding of the organization of candidates for immunotherapy [3,4], and the evolution of life on the genomic level [5,6].

Nowadays, methods of signal processing are becoming increasingly popular for various applications in bioinformatics as they may facilitate the exploration of intrinsic structural features. For example, mutual information functions [7,8], autocorrelation functions [9,10], power spectra [11,12], “DNA walk” representation [13], Zipf analysis [14], Fourier analysis [15] and so on have revealed many interesting physical properties of DNA molecules. However, the mosaic structure of DNA sequences is one of the main obstacles to intricate statistical analysis [16]. These patches appear in the DNA walk landscapes and are likely to introduce some breaking of scale invariance [17,18]. Wavelets have been widely applied to a variety of biomedical problems with great success [19–21]. In addition, wavelets are also well suited to visualizing patterns in DNA sequences and extracting regions with biological interest [22]. Wavelet transform has been found useful for verifying the existence of long-range correlations in intronic DNA sequences [23], and characterizing the scaling properties of sequences [24,25].

With the completion of the human genome sequence, comparative genomics has become a powerful approach

[*] Corresponding author.
E-mail: qiajiao@163.com (Qi Y).

http://dx.doi.org/10.1016/j.gpb.2012.07.004

to extracting genetic information from large stretches of nucleotide sequences. The chimpanzee, our closest living relative, may help us to understand humans thoroughly both in function and statistical features.

Fourier transformation is often used to convert a signal to the frequency domain. However, there are some disadvantages that restrict its wide application. For example, Fourier transformation cannot provide information about a simple discontinuity signal spectrogram and time-localization [26]. Compared with Fourier transformation, wavelets show advantages in analyzing signals that contain discontinuities and sharp spikes. This is mainly due to the fact that the wavelet transform incorporates in its definition two basic features, time and scale, which are important to fractal processes. Therefore, there is a growing interest in using wavelets in the sequence analysis. One of the easiest ways to extract information from a DNA sequence is to “view” it. In this study, we used the Daubechies wavelet (db1) 1-D Daubechies binary discrete wavelet analysis of DNA walks on the palindromes that contains some MAGEA and CSAG cancer/testis family genes (so called “MAGE/CSAG-palindrome”) [27]. By digital signal processing tools, we try to address the pattern irregularities in the palindromes, which are often associated with biological function. Our data suggest that segmental duplications and their reverse complemented sequences, which are located in either the two arms or the spacers of the palindromes, display distinct pattern regularities under wavelet analysis. In addition, a simple evolutionary model is proposed for the evolutionary relationship between human and chimpanzee based on the symmetrical structure of the MAGE/CSAG-palindromes.

Results and discussion

The orthologous segmental repeats have higher similarities than the paralogous sequences

We performed dot matrix program alignment for the palindrome sequences in human and chimpanzee. It was found that the sequence length is 111.399 kb for the palindrome from human, and 107.974 kb for that from chimpanzee. The left and right arms of human palindrome H_IR are 51.214 and 51.191 kb, respectively, with spacer of 8.994 kb. However, the arms of the homologous X-palindrome in chimpanzee are shorter (44.294 and 44.308 kb, respectively), while the spacer is lager (19.372 kb). The arm-to-arm similarity in palindrome is 99.7% and 99.8% in the chimpanzee and human, respectively. However, the similarity of orthologous sequences between human and chimpanzee is a little bit lower, which is 97.4%, 97.3% and 94.3% for the left arms, right arms and the spacer regions, respectively.

As seen in **Figure 1**, the dot plot shows that there are five segmental repeats, r1 to r5, on the MAGE/CSAG-palindrome from both human (H_r1~H_r5) and chimpanzee (P_r1~P_r5). These include two inverted repeats and two direct repeats on both arms of the palindrome and one segmental repeat in the spacers which is completely reversed. Using the Martinez-NW Method, we calculated the similarity between these repeats. The results showed that the similarities between the repeats within the same species were mostly less than 90% (Table S1A, B), while the similarities were more than 90% between orthologous repeats of the human and chimpanzee palindromes (Table S1C). For example, the similarity between r1 and the other repeats (r2 to r5) in human is 80.60%, 82.20%, 80.60% and 99.50%, respectively while the r1 repeat from human and chimpanzee shares 97.8% similarity.

Three-dimensional DNA walks of the human and chimpanzee X palindromes

Graphical representations of DNA sequences are helpful because they allow visual observations of base patterns, sequence composition and evolution. The four-letter genome alphabet is firstly converted into some numerical pattern. By increasing the dimensionality of numerical DNA sequence, DNA walk representation is no longer limited to strictly binary classifications only [13], and may be useful to reflect structural information.

In this study, we adopt an image representation for nucleotides based on the 3D DNA walks analysis. Similar trends of the repetitive regions on the both arms of palindromes, and dissimilar tendencies in the spacers can be observed easily using the complex walk representation. The resulting walk sequences were plotted in the 3D Cartesian coordinate system with the index k for sequence Y plotted along the Z axis (**Figure 2**). The DNA walk plots provide a direct graphical representation for these direct repeats and inverted repetitive sequences. In particular, the graphical representation highlights the similarity and difference of the palindromes between human and chimpanzee. The remarkable differences were further investigated by wavelet analysis to visualize the fractal pattern along the human and chimpanzee palindromes.

Wavelet analysis of DNA walks of the palindromes

Figure 3 display the 3D DNA walks wavelet analysis of the MAGE/CSAG-palindromes for human and chimpanzee, respectively. The top portion of each of the figures produced by the Wavelet toolbox plots the Ca, τ wavelet coefficients versus the base index number, while the bottom portion is a “temperature” plot of $|Ca, \tau|^2$ versus the scale and base index number. High temperatures correspond to high intensities of $|Ca, \tau|^2$ [22]. Regions of high intensity in the wavelet transform are correlated with segmental duplications, which contain testis/cancer MAGEA/CSAG genes. As shown in Figure 3A, regions of high intensity in the yellow boxes are matched to these segmental repeats, which include all introns and exons of MAGE-A genes [27], such as MAGEA 6, MAGEA 2B, MAGEA 12, MAGEA 2 and MAGEA 3 genes. However, some CSAG genes fell out of the scope,

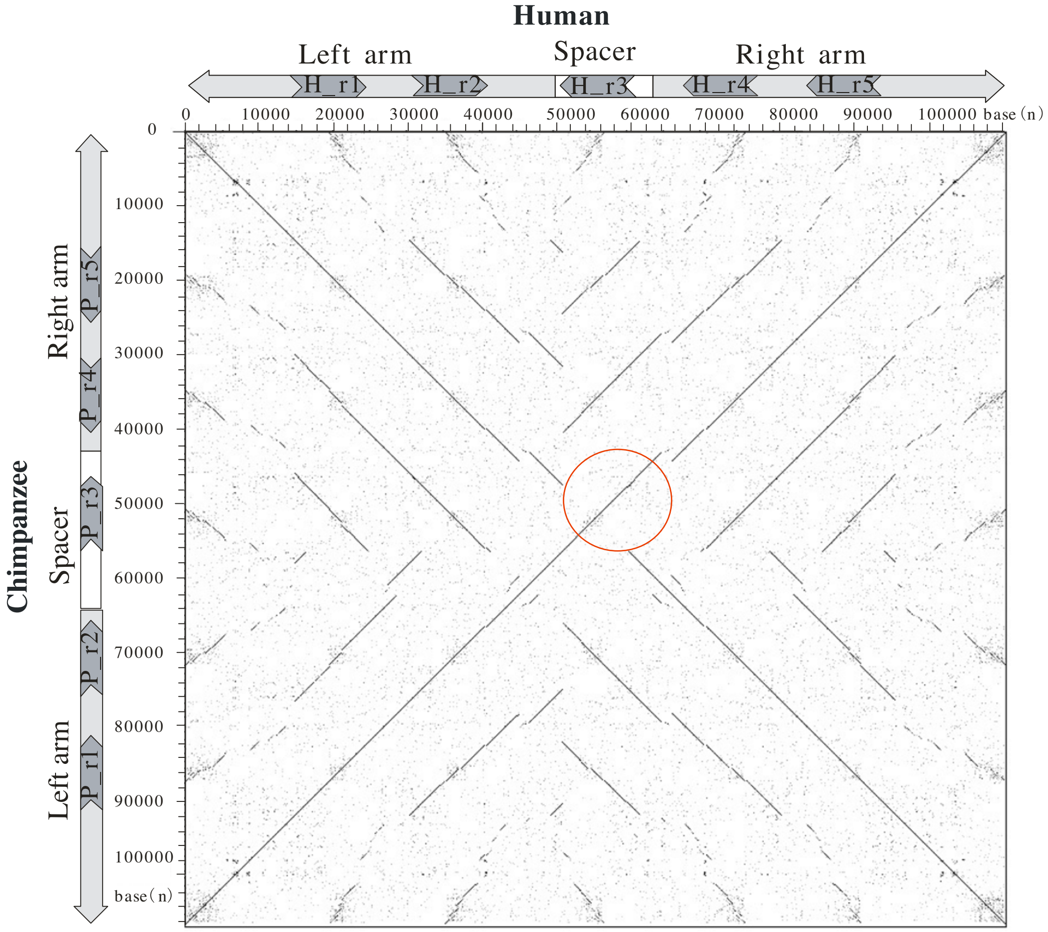

Figure 1 Sketch map of segmental repeats on the palindrome in human and chimpanzee
H_r1~H_r5 and P_r1~P_r5 indicate the segmental repeats (1–5) on the MAGE/CSAG-palindrome in the human (*Homo sapiens*) and chimpanzee (*Pan troglodytes*), respectively. Gray arrowheads denote the arms of the palindromes. The segmental repeats (H_r3 and P_r3) in the spacer regions are reverse complementary between human and chimpanzee (showed in the red circle). The position of these repeats are acquired by dot matrix program alignment (H_r1: 14816~24813, H_r2: 30428~41411, H_r3: 51211~61752, H_r4: 66086~80973, H_r5: 82719~96568; P_r1: 14765~24744, P_r2: 30378~40998, P_r3: 46630~57177, P_r4: 63065~77581, P_r5: 79327~93220).

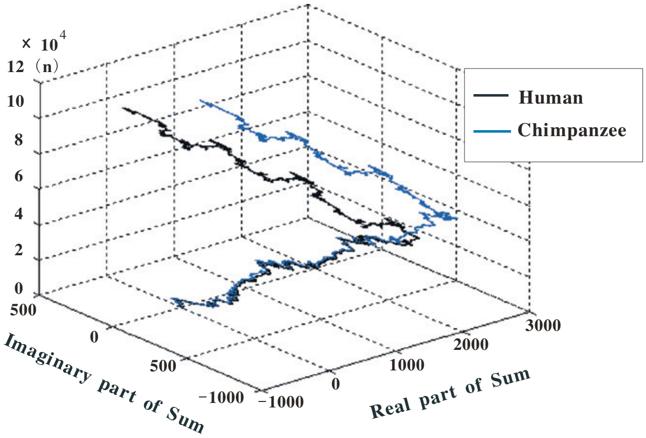

Figure 2 3D DNA walks of palindromes on the X chromosomes
DNA walk for palindromes on the human and chimpanzee X chromosomes was shown in black and blue, respectively. Here, we define the value $x[k] = +1$ for A, $x[k] = -1$ for T, $x[k] = +j$ for C, $x[k] = -j$ for G. For any position k we have a cumulative sum. So the sum of A − T is a real part of the sum, and the sum of C − G is the imaginary part of the sum. The base numbering is shown on Z axis.

such as CSAG2, CSAG4, CSAG1 and CSAG2B. A similar phenomenon was observed in the MAGE/CSAG-palindrome of chimpanzee (Figure 3B). Indeed, these segmental repeats generate irregular patterns in the DNA walk, which show high intensity in the wavelet transform. The region of high-intensity appearing on the right side of the fifth rectangular box may be due to its high GC components. This suggested that the segmental duplications with biological functions of the palindromes confer this distinct wavelet transform pattern from their surrounding sequences without biological functions.

Phylogenetic relationship between the human and chimpanzee palindromes

The similarity between the whole human and chimpanzee palindrome structure is 94.5%. In order to look for evidence that the MAGE/CSAG-palindrome was present in the common ancestor of human and chimpanzee, we further analyzed the two 400 bp inner boundaries (the sequences between the left arm and the spacer, and between the spacer

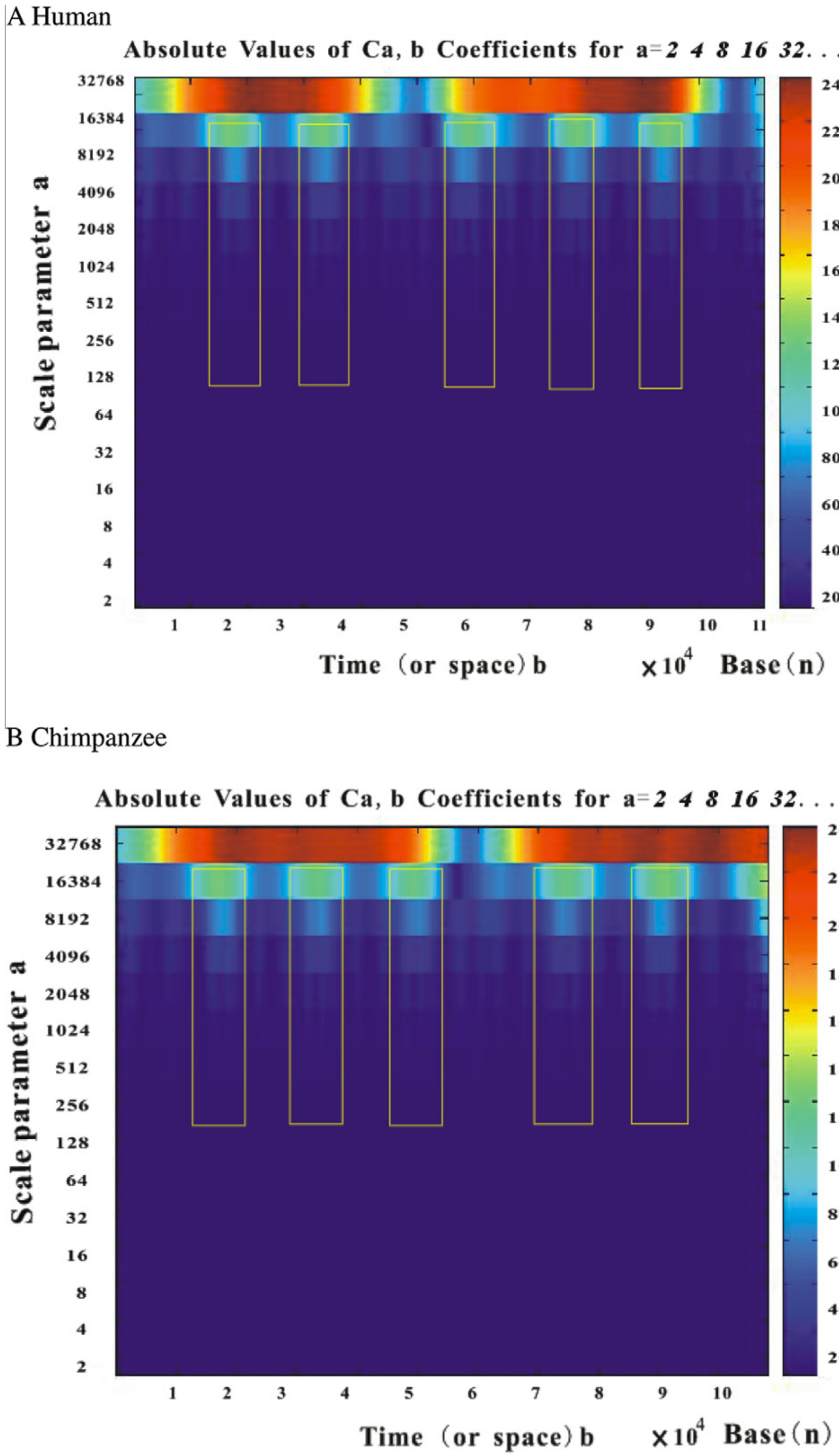

Figure 3 Wavelet transform analysis for 3D DNA walk of X-palindrome

A. Yellow boxes (high temperature) represent the location of segmental repeats on human palindrome, including two repetitive segments on each arm and one in the spacer region. **B.** Yellow boxes represent the location of segmental repeats on chimpanzee palindrome, including two direct repetitive segments on the left arm, one in the spacer, and two inverted repetitive segments on the right arm.

and the right arm) and the two 500 bp outer boundaries (the sequence between the left/right boundary and the left/right arm). Results showed that the identities of the orthologous outer boundaries are remarkably high between human and chimpanzee, but the orthologous inner boundaries have very low similarities. For example, the similarity between the left outer boundary of H_IR (L_O_H) and the left outer boundary of P_IR (L_O_P) is 99.20%, but the similarity between the left inner boundary of H_IR (L_I_H) and the left inner boundary of P _IR (L_I_P) is 34.40%. However, we observed extremely low similarity between the paralogous two outer/inner boundaries within the same palindrome (Table S2). For example, the similarity between the L_O_H and the right outer boundary of H_IR (R_O_H) is 31.60%, the similarity between the L_O_P and the right outer boundary of P_IR (R_O_P) is 31.60%. These findings suggested that the palindromes were already present in the common ancestor of humans and chimpanzees. Furthermore, it was also found that the similarities between the orthologous arms are less than that of the paralogous arms both in human and chimpanzee. In this study, all repetitive segments contain MAGE-A/CSAG genes, and the same gene order is maintained in the chimpanzee as in the human [28,29]. Therefore, we speculated that the palindromes might pre-date separation of the human and chimpanzee lineages, and the paired arms of the palindromes evolved in concert.

However, it is still a challenge for researchers to interpret the recent origin of these segments and their role in primate genome evolution [29–34]. Segmental duplications have been shown to be associated with genome rearrangement events during species evolution [35,36]. Unequal crossing over may happen between these repetitive segments with high similarities during evolution after human diverged from chimpanzee, which may have strong effects on long-scale correlation observed in the original palindromes. Therefore, a simple model of the evolutionary relationship between the MAGE/CSAG-palindromes of human and chimpanzee is proposed in **Figure 4**. The common ancestral palindrome may have three symmetrical repetitive segments on each arm. When unequal chromatid exchange occurs between tandem arrays of sequence, contraction and expansion of the array can homogenize the sequence repeats. Finally, the produced palindromes include different structural patterns. This may suggest that the evolutionary mechanism and structures of palindromes on the X chromosome differ from those of the palindromes on the Y chromosome [1,2,37]. However, further experiments are still needed to study the structure, biological function and the relationship between human and chimpanzee.

Conclusion

The graphic comparison of the sequences by DNA walks using wavelet analysis may help us to better investigate and characterize the features and structures, and offer a comprehensive understanding for regions of biological interest. From the 3-D DNA walk plots, we can visualize the similarities and differences between human and chimpanzee during the evolutionary process. Further wavelet analysis indicates that the sequences with biological significance have different patterns compared to surrounding sequences, and the duplicated segments present in the palindromes reflect the evolutionary course. Based on the specific distribution and possible occurrences, such as crossing over, gene conversion and so on, we proposed a simple evolutionary model of the MAGE/CSAG-palindromes on the human and the chimpanzee X chromosomes, although the numbers of palindromes on the human X chromosome considered here are not plenteous, and our knowledge about their real evolutionary origin and hidden significance is incomplete. The model allows for a direct visualization of irregular genomic structural patterns, and may offer a new vision to better understand the special structure and evolution of the X-palindromes. Further experiments are still needed to verify the particular characteristics and biological functions, as well as the phylogenetic relationship to other primates.

Materials and methods

Subjects and data

The sequences of human complete palindrome and chimpanzee BAC clones (AC145689, AC144384) were downloaded from the NCBI website (http://www.ncbi.nlm.nih.gov/) (NCBI, Build Number: 37.1). The palindrome locates on the human X chromosome 151847041∼

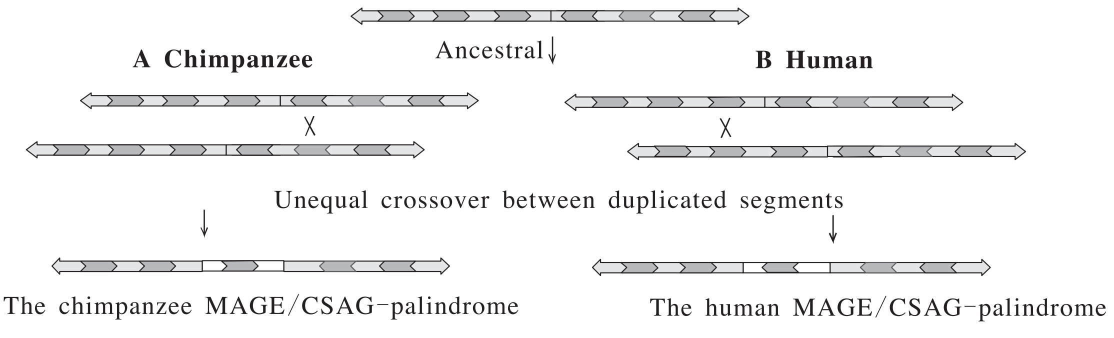
Ancestral
A Chimpanzee
B Human
X
X
Unequal crossover between duplicated segments
The chimpanzee MAGE/CSAG-palindrome
The human MAGE/CSAG-palindrome

Figure 4 Proposed evolutionary model of the MAGE/CSAG-palindromes
The common ancestral palindrome may have three symmetrical repetitive segments on each arm. The chimpanzee palindrome **(A)** and the human palindrome **(B)** were produced after unequal crossover between duplicated segments. There are two inverted repetitive segments (black regions with the right arrow) and three direct repetitive segments (black regions with the left arrow) on the chimpanzee MAGE/CSAG-palindrome. In contrast, three inverted repetitive segments and two direct repetitive segments exist on the human MAGE/CSAG-palindrome.

151958439, containing some testis/cancer genes [37]. Location of the palindrome on the chimpanzee X chromosome, as well as the arms of the two palindromes, was obtained, by performing dot matrix program alignment of the BAC clones [38].

DNA walk analysis

The random 2D DNA walk plot provides a tool to exhibit periodic patterns in a sequence [13]. By increasing the dimensionality of the numerical DNA sequence, the graphical representation is no longer limited strictly to binary classifications. Here we adopt a distinct representation for nucleotides based on their mapping into the four cardinal points $\{+1, -1, +j, -j\}$ of the complex plane. Let $\chi = \{x[i]; i = 1, 2, \ldots, N\}$ denote as a DNA sequence of length N. For a position k within the DNA sequence, we define the value $x[k] = +1$ for A, $x[k] = -1$ for T, $x[k] = +j$ for C, $x[k] = -j$ for G. Furthermore, let us denote the 3D DNA walk sequence as $Y = \{y[i]; i = 1, 2, \ldots, N\}$, where for any position k we have a cumulative sum of the $x[i]$ for $1 \leqslant i \leqslant k$ described by:

$$y[k] = \sum_{i=1}^{k} x[i] \tag{1}$$

In this method, four cardinal directions in (x, y) coordinate system are chosen to represent the content of the four bases in DNA sequences. Among the existing methods of DNA sequence visualization, 3D walk [39] is the most popular method.

Wavelet-based analysis

Wavelet-based tools are well suited to multi-resolution analysis and local feature extraction of non-stationary signals, such as locating different patterns in genome sequences [40]. The continuous wavelet transform (CWT) of a signal $B_H(t)$ with respect to the wavelet $\psi(t)$ is defined as:

$$Ca, \tau \equiv \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} B_H(t) \psi \left[\frac{t - \tau}{a} \right] dt \tag{2}$$

where a (the scale parameter) > 0 and τ (the translation parameter) is a real number.

Generally, by using the CWT with discrete values of a and τ, the discrete wavelet transform (DWT) is simply determined. Here we merely choose the Daubechies wavelet (db1) 1-D Daubechies, and denote $a = 2^j, j \in Z$.

Authors' contributions

YQ conceived the idea, collected the datasets, carried out the analysis, interpreted the data and drafted the manuscript. NJ provided the corresponding computer programs about wavelet transform, and revised the manuscript. DA revised the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

Acknowledgements

We would like to thank Yi Wan for helpful discussions about computer programs. We thank Maoyi Li for discussions about the results. We also thank the anonymous reviewers for their valuable advice. This work was supported by the Fundamental Research Funds for the Central Universities (Program No. zyz2012063).

Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.gpb.2012.07.004.

References

[1] Kuroda-Kawaguchi T, Skaletsky H, Brown LG, Minx PJ, Cordum HS, Waterston RH, et al. The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. Nat Genet 2001;29:279–86.

[2] Skaletsky H, Kuroda-Kawaquchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, et al. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature 2003;423:825–37.

[3] Andrade VC, Vettore AL, Felix RS, Almeida MS, Carvalho F, Oliveira JS, et al. Prognostic impact of cancer/testis antigen expression in advanced stage multiple myeloma patients. Cancer Immun 2008;8:2.

[4] Kondo T, Zhu X, Asa SL, Ezzat S. The cancer/testis antigen melanoma-associated Antigen-A3/A6 is a novel target of fibroblast growth factor receptor 2-IIIb through histone H3 modifications in thyroid cancer. Clin Cancer Res 2007;13:4713–20.

[5] Beckmann JS, Trifonov EN. Splice junctions follow a 205-base ladder. Proc Natl Acad Sci U S A 1991;88:2380–3.

[6] Lobzin VV, Chechetkin VR. Order and correlations in genomic DNA sequences: the spectral approach. Phys Usp 2000;43:55–78.

[7] Li W, Kaneko K. Long-range correlation and partial 1/f$^{\alpha}$ spectrum in a non-coding DNA sequence. Europhys Lett 1992;17:655–60.

[8] Dehnert M, Helm WE, Hütt MT. Information theory reveals large-scale synchronization of statistical correlations in eukaryote genomes. Gene 2005;345:81–90.

[9] Goncharov AF, Mazin II II, Eggert JH, Hemley RJ, Mao HK. Invariant points and phase transitions in deuterium at megabar pressures. Phys Rev Lett 1995;75:2514–7.

[10] Bernaola-Galván P, Carpena P, Román-Roldán R, Oliver JL. Study of statistical correlations in DNA sequences. Gene 2002;300:105–15.

[11] Voss RF. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. Phys Rev Lett 1992;68:3805–8.

[12] Fukushima A, Ikemura T, Kinouchi M, Oshima T, Kudo Y, Mori H, et al. Periodicity in prokaryotic and eukaryotic genomes identified by power spectrum analysis. Gene 2002;300:203–11.

[13] Peng CK, Buldyrev SV, Goldberger AL, Havlin S, Sciortino F, Simons M, et al. Long-range correlations in nucleotide sequences. Nature 1992;356:168–70.

[14] Mantegna RN, Buldyrev SV, Goldberger AL, Havlin S, Peng CK, Simons M, et al. Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics 1995;52:2939–50.

[15] Tiwari S, Ramachandran S, Bhattacharya A, Bhattacharya S, Ramaswamy R. Prediction of probable genes by Fourier analysis of genomic sequences. Comput Appl Biosci 1997;13:263–70.

[16] Fickett JW, Torney DC, Wolf DR. Base compositional structure of genomes. Genomics 1992;13:1056–64.

[17] Nee S. Uncorrelated DNA walks. Nature 1992;357:450–3.

[18] Karlin S, Brendel V. Patchiness and correlations in DNA sequences. Science 1993;259:677–80.

[19] Unser M, Aldroubi A. A review of wavelets in biomedical applications. Proc IEEE 1996;84:626–38.

[20] Liò P. Wavelets in bioinformatics and computational biology: state of art and perspectives. Bioinformatics 2003;19:2–9.

[21] Arneodo A, d'Aubenton-Carafa Y, Bacry E, Graves PV, Muzy JF, Thermes C. Wavelet based fractal analysis of DNA sequences. Physica D 1996;96:291–320.

[22] Haimovich AD, Byrne B, Ramaswamy R, Welsh WJ. Wavelet analysis of DNA walks. J Comput Biol 2006;13:1289–98.

[23] Arneodo A, Bacry E, Graves PV, Muzy JF. Characterizing long-range correlations in DNA sequences from wavelet analysis. Phys Rev Lett 1995;74:3293–6.

[24] Audit B, Bacry E, Muzy JF, Arneodo A. Wavelet-based estimators of scaling behavior. IEEE Trans Inf Theory 2002;48:2938–54.

[25] Audit B, Thermes C, Vaillant C, d'Aubenton-Carafa Y, Muzy JF, Arneodo A. Long-range correlations in genomic DNA: a signature of the nucleosomal structure. Phys Rev Lett 2001;86:2471–4.

[26] Mandal BN, Chakrabarti A. A generalization to the hybrid Fourier transform and its application. Appl Math Lett 2003;16:703–8.

[27] Bredenbeck A, Hollstein VM, Trefzer U, Sterry W, Walden P, Losch FO. Coordinated expression of clustered cancer/testis genes encoded in a large inverted repeat DNA structure. Gene 2008;415:68–73.

[28] Losch FO, Bredenbeck A, Hollstein VM, Walden P, Wrede P. Evidence for a large double-cruciform DNA structure on the X chromosome of human and chimpanzee. Hum Genet 2007;122:337–43.

[29] Saionz JR. Palindromes on the human X chromosome: testis-biased transcription, gene conversion and evolution. PhD Thesis, Massachusetts Institute of Technology; 2005.

[30] Chen FC, Li WH. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am J Hum Genet 2001;68:444–56.

[31] Smith GP. Evolution of repeated DNA sequences by unequal crossover. Science 1976;191:528–35.

[32] Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, et al. Recent segmental duplications in the human genome. Science 2002;297:1003–7.

[33] Eichler EE. Recent duplication, domain accretion and the dynamic mutation of the human genome. Trends Genet 2001;17:661–9.

[34] Samonte RV, Eichler EE. Segmental duplications and the evolution of the primate genome. Nat Rev Genet 2002;3:65–72.

[35] Armengol L, Pujana MA, Cheung J, Scherer SW, Estivill X. Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements. Hum Mol Genet 2003;12:2201–8.

[36] Bailey JA, Baertsch R, Kent WJ, Haussler D, Eichler EE. Hotspots of mammalian chromosomal evolution. Genome Biol 2004;5:R23.

[37] Warburton PE, Giordano J, Cheung F, Gelfand Y, Benson G. Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. Genome Res 2004;14:1861–9.

[38] Sonnhammer EL, Durbin R. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 1995;167:GC1–GC10.

[39] Gate MA. A simple way to look at DNA. J Theor Biol 1986;119:319–28.

[40] Liò P, Vannucci M. Finding pathogenicity islands and gene transfer events in genome data. Bioinformatics 2000;16:932–40.