SYMBOL-BALANCED QUATERNIONIC PERIODICITY TRANSFORM FOR LATENT PATTERN DETECTION IN DNA SEQUENCES

Andrzej K. Brodzik and Olivia Peters

The MITRE Corporation, {Bedford MA 01730, McLean VA 22102}

{abrodzik, otate}@mitre.org

ABSTRACT

A new approach towards computing periodicity transform of DNA sequences is proposed. The approach is based on mapping of DNA symbols to pure quaternions. The resulting quaternionic periodicity transform outperforms the previously proposed complex periodicity transform due to enhanced, symbol-balanced sensitivity to DNA patterns. The theoretical finding is supported by performance comparison of the two transforms and by an application example.

1. INTRODUCTION

It has been often observed that the occurrence of repetitive structures (or tandem repeats) in genomic data is symptomatic of biological phenomena. Perhaps the best known example of this association is the 3-base repetition of codons, which is characteristic of protein coding regions in DNA sequences of eucaryotic cells. The 3-base repeat is considered large-scale, as it occurs throughout the genome, in contrast to small-scale repeats, typically restricted to individual genes or gene subsets. Other well known large-scale genomic repeats include the 10.5-base repeats that are due to a 3.5 aminoacid repeat in alpha-helical coiled-coil regions in proteins, and the 200-base and the 400-base repeats that are thought to have evolved by fusion of genome segments of nearly identical sizes [18]. Small-scale genomic repeats and repeat changes in the human genome has been associated, among others, with genetic diseases such as Huntington disease, myotonic dystrophy and Friedreich ataxia (CTG, CGG and GAA repeats) [3], and with progression of cancer [14].

Applications of repetitive structures include prediction of gene and exon locations [11], identification of diseases [3], reconstruction of human evolutionary history [17], DNA forensics [5], detection of pathogen exposure [6], and prediction of the relative level of gene expression [16]. The number of known repetitive structures and their applications is certain to grow, as repetitions are estimated to comprise more than one-half of the human genome [10].

The methods used to detect DNA repeats can be classified as either probabilistic or deterministic. Among the deterministic approaches, most rely on spectral analysis of the data, which is typically based either on Fourier [1], Walsh [16], or wavelet transform processing [2]. More recently, a time-domain method, called the periodicity transform has been proposed [13] and applied to genomic feature detection [4]. The periodicity transform enjoys several advantages over the spectral methods, perhaps the most important of which in the context of genomic signal processing is its superior computational efficiency. One of the disadvantages of periodicity transform (which it shares with the spectral methods) is its symbol bias that is inherent in the mapping of DNA symbols to complex numbers, and which results in missed detections of some repetitive structures.

In this paper we propose to replace the complex number set with its algebraic generalization, the set of quaternions [7]. This replacement results in a periodicity transform that is symbol balanced and that detects all repetitive structures. We anticipate that the quaternionic approach can be utilized (via the quaternionic Fourier transform [15]) to improve the spectral methods, and (via the use of higher dimensional hypercomplex number systems, such as octonions and sedenions [9]) to facilitate symbolic signal processing in applications utilizing larger than genomic alphabet sizes.

2. PERIODICITY TRANSFORM

2.1. Periodicity detection

Take $N = PP'$, $P, P' \in Z^+$, and let x be an arbitrary N-point sequence of real numbers, $x = x_0, x_1, ..., x_{N-1}$. Define the periodicity transform (PT) [13] by

$$X_P^N(s) = \frac{1}{P'} \sum_{k=0}^{P'-1} x(s+kP), \;\; 0 \le s < P. \tag{1}$$

Take $\bar{N} = P\bar{P}'$, $\bar{P}'$ and $P'/\bar{P}' \in Z^+$, and $\bar{P}' << P'$. In analogy to the short time Fourier transform, define the short time periodicity transform (STPT) by

$$X_P^{\bar{N},N}(n) = \frac{1}{\bar{P}'} \sum_{k=0}^{\bar{P}'-1} x(n+kP), \tag{2}$$

$$= [X_P^{\bar{N}}(s), X_P^{\bar{N}}(s+P), ..., X_P^{\bar{N}}(s+P(P'-\bar{P}'))], \tag{3}$$

where $0 \le n < N - (\bar{P}' - 1)P$ and $0 \le s < P$. A more robust indicator of presence of a periodic component in a sequence is obtained, when the STPT is normalized by the sequence $\bar{N}$-point segment's 'power' and averaged over P shifts, i.e.,

$$< x >^{\#}(n_0) = \frac{||X_P^{\bar{N}}(n_0)||^2}{\frac{1}{P'}||x(n_0)||^2} = \frac{\sum_{s=0}^{P-1} |X_P^{\bar{N}}(n_0+s)|^2}{\frac{1}{P'}\sum_{n=0}^{\bar{N}-1} |x(n_0+n)|^2}, \tag{4}$$

where $0 \le n_0 < N - \bar{N}$. We call this indicator the periodicity detector (PD). Since $X_P^{\bar{N}}$ is a P-point sequence, and x in (2) is evaluated over an implicit $\bar{N}$-point window, hence the normalization factor $1/\bar{P}'$. If periodicity of the repetitive component is known

Approved for Public Release; Distribution Unlimited: Case # 04-1019

within some specified range, e.g., $P_1 \leq P \leq P_2$, then PD is computed for all P within that range, i.e., $< x >^{\#} (n_0, P),\ P_1 \leq P \leq P_2$, with $\arg \max_{n_0,P}\{< x >^{\#} (n_0, P)\}$ yielding the estimate of the period and the position of the unknown periodic component.

In the next subsection we will investigate properties of PD of a symbolic sequence, evaluated at a single shift. In preparation, we define the single shift PD of a numeric sequence,

$$< x > (n_0 + s) = \frac{|X_P^{\bar{N}}(n_0 + s)|^2}{\frac{1}{\bar{P}'}\sum_{k=0}^{\bar{P}'-1}|x(n_0 + s + kP)|^2} = \frac{|\sum_{k=0}^{\bar{P}'-1} x(n_0 + s + kP)|^2}{\bar{P}'\sum_{k=0}^{\bar{P}'-1}|x(n_0 + s + kP)|^2}, \quad (5)$$

where $0 \leq n_0 < N - \bar{N},\ 0 \leq s < P$. Note, that for convenience, contrary to (4), in (5) normalization is performed over a subset of values of x, so that $0 \leq < x > (n_0 + s) \leq 1$ for all x. As an aside, we observe that, if $\sum_k |x(n_0 + s + kP)|^2 =$const for all s, then $P < x >^{\#} (n_0) = \sum_s < x > (n_0 + s)$.

2.2. Periodic symbol detector

Take an arbitrary NP-point DNA sequence, $x = x_0, x_1, ..., x_{NP-1}$, and choose any stride by P N-point decimation of x, e.g., $x^0 = x_0, x_P, ..., x_{(N-1)P}$. Denote by integers $N_a,\ N_c,\ N_g$ and $N_t,\ N_a + N_c + N_g + N_t = N$, the count of symbols $'a',\ 'c',\ 'g'$ and $'t'$ in x^0. Consider an assignment of DNA symbols to arbitrary (complex or hypercomplex) numbers, e.g.,

$$\begin{aligned} 'a' &\mapsto o_0, \\ 'c' &\mapsto o_1, \\ 'g' &\mapsto o_2, \\ 't' &\mapsto o_3, \end{aligned} \quad (6)$$

$o_0 \neq o_1 \neq o_2 \neq o_3$ and $|o_0| = |o_1| = |o_2| = |o_3| \doteq |o|$. The abstract single shift periodic symbol detector (PSD) of x can then be expressed as

$$< x > [\alpha, \beta, \gamma, \delta] \doteq < x > (0) = \frac{|\alpha o_0 + \beta o_1 + \gamma o_2 + \delta o_3|^2}{|o|^2}, \quad (7)$$

where $[\alpha, \beta, \gamma, \delta] = [N_a/N, N_c/N, N_g/N, N_t/N]$. We require that $< x > [\alpha, \beta, \gamma, \delta]$ satisfies the following conditions:

1. $< x > [\alpha, \beta, \gamma, \delta]$ has a minimum at $[1/4, 1/4, 1/4, 1/4]$.
2. $< x > [\alpha, \beta, \gamma, \delta]$ has a maximum at $[1, 0, 0, 0]$.
3. If $\alpha \geq \beta, \gamma, \delta$ then $< x > [\alpha + 1/N, \beta - 1/N, \gamma, \delta] - < x > [\alpha, \beta, \gamma, \delta] > 0$.
4. $< x > [\alpha, \beta, \gamma, \delta]$ is invariant under permutation of any two symbols.

2.3. Complex assignment of DNA symbols

Consider an assignment of DNA symbols to complex numbers [4], e.g.,

$$\begin{aligned} 'a' &\mapsto p_0 = 1 + i, \\ 'c' &\mapsto p_1 = 1 - i, \\ 'g' &\mapsto p_2 = -1 + i, \\ 't' &\mapsto p_3 = -1 - i. \end{aligned} \quad (8)$$

The complex PSD can then be expressed by the following formula

$$< x >_c [\alpha, \beta, \gamma, \delta] = \frac{|\alpha p_0 + \beta p_1 + \gamma p_2 + \delta p_3|^2}{|p_0|^2} = \frac{(\alpha + \beta - \gamma - \delta)^2 + (\alpha - \beta + \gamma - \delta)^2}{2} \quad (9)$$

Since

$$< x >_c [1/4, 1/4, 1/4, 1/4] = 0. \quad (10)$$

and

$$\begin{aligned} < x >_c [1, 0, 0, 0] &= < x >_c [0, 1, 0, 0] \\ &= < x >_c [0, 0, 1, 0] \\ &= < x >_c [0, 0, 0, 1] = 1, \end{aligned} \quad (11)$$

the complex PSD satisfies conditions 1 and 2. Condition 3, however, is not met since, for example, $< x >_c [1/3 + 1/N, 1/3 - 1/N, 0, 1/3] - < x >_c [1/3, 1/3, 0, 1/3] < 0$ for $N > 3$. Moreover, since, e.g.,

$$< x >_c [1/2, 0, 1/2, 0] = 1 \quad (12)$$

and

$$< x >_c [1/2, 0, 0, 1/2] = 0 \quad (13)$$

it is readily seen that condition 4 is violated as well. In general, the complex PSD given by equations (8-9) is variant under exchange of any two parameters, except for the pairs β and γ, and α and δ. In fact, no assignment of type (8) leads to a detector meeting all four conditions of subsection 2.2.

3. THE QUATERNIONIC APPROACH

Real and complex numbers can be viewed as one- and two-dimensional instances of N-dimensional hypercomplex numbers of the form

$$h = a_0 i_0 + a_1 i_1 + a_2 i_2 + ... + a_N i_N,\ \ N \in Z^+ \cup \{0\}, \quad (14)$$

where $a_j \in \mathcal{R}, 0 \leq j < N$, $i_0 = 1$ and $i_j, 0 < j \leq N$, are symbols called imaginary units. If $N = 0$ then h is real and if $N = 1$ and $i_1^2 = -1$ then h is complex. The best known hypercomplex numbers, apart from the real and the complex numbers, are the four-dimensional quaternions and the eight-dimensional octonions.

The concept of quaternions was introduced by William Hamilton in 1843 [7], who defined a quaternion as a number of the form

$$q = a + bi + cj + dk, \quad (15)$$

where $a, b, c, d \in \mathcal{R}$, and i, j, k are symbols defined by the following set of rules

$$\begin{aligned} i^2 = j^2 = k^2 &= -1, \\ ij &= -ji = k, \\ jk &= -kj = i, \\ ki &= -ik = j. \end{aligned} \quad (16)$$

a is called the real, or scalar, part of q and $q - a$ is called the imaginary, or vector, part of q. $q - a$ is also known as the pure quaternion. Quaternions follow the usual addition rule

$$(a + bi + cj + dk) + (a' + b'i + c'j + d'k) = (a + a') + (b + b')i + (c + c')j + (d + d')k, \quad (17)$$

and a distinct multiplication rule

$$\begin{aligned}&(a+bi+cj+dk)(a'+b'i+c'j+d'k)\\&=(aa'-bb'-cc'-dd')+(ab'+ba'+cd'-dc')i\\&+(ac'+ca'+db'-bd')j+(ad'+da'+bc'-cb')k.\end{aligned}\quad(18)$$

The multiplication is not commutative, i.e., $q_1q_2 \neq q_2q_1$, which is due to the relation between imaginary units (16). The quaternion $q = a + bi + cj + dk$ has a conjugate

$$\bar{q} = a - bi - cj - dk, \quad (19)$$

a norm

$$|q| = \sqrt{q\bar{q}} = \sqrt{a^2+b^2+c^2+d^2}, \quad (20)$$

and an inverse

$$q^{-1} = \frac{\bar{q}}{|q|^2}. \quad (21)$$

Applications of quaternions in signal processing include computer vision, robotics [15], and color and hyperspectral image processing [12,15]. For an elementary treatment of quaternions, see [9].

3.1. Quaternionic assignment of DNA symbols

Consider the assignment of DNA symbols to pure quaternions, e.g.,

$$\begin{aligned}{}'a' &\mapsto q_0 = i+j+k,\\ {}'c' &\mapsto q_1 = i-j-k,\\ {}'g' &\mapsto q_2 = -i-j+k,\\ {}'t' &\mapsto q_3 = -i+j-k.\end{aligned}\quad(22)$$

The quaternionic PSD can then be expressed by the following formula

$$\begin{aligned}&< x >_q [\alpha,\beta,\gamma,\delta] = \frac{|\alpha q_0+\beta q_1+\gamma q_2+\delta q_3|^2}{|q_0|^2}\\&=\frac{(\alpha+\beta-\gamma-\delta)^2+(\alpha-\beta-\gamma+\delta)^2+(\alpha-\beta+\gamma-\delta)^2}{3}\end{aligned}\quad(23)$$

In particular, we have

$$\begin{aligned}< x >_q [1,0,0,0] &= < x >_q [0,1,0,0]\\&= < x >_q [0,0,1,0]\\&= < x >_q [0,0,0,1] = 1,\end{aligned}\quad(24)$$

$$\begin{aligned}< x >_q [1/2,1/2,0,0] &= < x >_q [1/2,0,1/2,0]\\&= < x >_q [1/2,0,0,1/2]\\&= < x >_q [0,1/2,1/2,0]\\&= < x >_q [0,1/2,0,1/2]\\&= < x >_q [0,0,1/2,1/2] = 1/3,\end{aligned}\quad(25)$$

$$\begin{aligned}< x >_q [1/3,1/3,1/3,0] &= < x >_q [1/3,1/3,0,1/3]\\&= < x >_q [1/3,0,1/3,1/3]\\&= < x >_q [0,1/3,1/3,1/3] = 1/9\end{aligned}\quad(26)$$

$$< x >_q [1/4,1/4,1/4,1/4] = 0. \quad (27)$$

From equations (27) and (24) we have that $< x >_q$ satisfies conditions 1 and 2. To verify condition 3, note that $< x >_q [\alpha + 1/N, \beta - 1/N, \gamma, \delta] - < x >_q [\alpha,\beta,\gamma,\delta] = (8/3N)(\alpha - \beta + 1/N) > 0$ if $\alpha + 1/N > \beta$. To verify that $< x >_q$ satisfies conditions 4, note that the numerator in (23) can be written as $3(\alpha^2+\beta^2+\gamma^2+\delta^2) - 2(\alpha\beta+\alpha\gamma+\alpha\delta+\beta\gamma+\beta\delta+\gamma\delta)$.

3.2. Performance comparison

We have compared performance curves and surfaces of the complex and quaternionic detectors in several special cases.

Fig. 1a-1c shows magnitude of complex and quaternionic detectors when either $\gamma = 0$ or $\delta = 0$. Symmetry of the quaternionic detector surface manifests symbol permutation invariance of the quaternionic detector. Lack of symmetry of the complex detector surface manifests symbol imbalance of the complex detector.

Fig. 2a-2c shows performance curves of complex and quaternionic detectors when four, three or two symbol counts are non-zero. Of the non-zero symbol counts, all but one are set to the same value, i.e., in plot a $\beta = \gamma = \delta = \frac{1-\alpha}{3}$, in plot b $\beta = \gamma = \frac{1-\alpha}{2}$, and in plot c $\beta = 1 - \alpha$.

The first plot shows that for $\beta = \gamma = \delta = \frac{1-\alpha}{3}$ the performance of the complex and quaternionic detectors is identical. The second plot (three symbol counts are non-zero) shows that the complex detector performance differs from the quaternionic detector performance, and that the complex detector performance depends on symbol selection. The solid line in Fig. 2b corresponds to the $\alpha + \beta + \gamma = 2\beta + \alpha = 1$ line in Fig. 1a, and the dashed line in Fig. 2b corresponds to the $\alpha + \beta + \delta = 2\beta + \alpha = 1$ line in Fig. 1b. The two complex detectors coincide at points $\alpha = 1/3$ and $\alpha = 1$. At point $\alpha = 0$, the two complex detectors differ by $1/2$, i.e., $< x >_c [0, 1/2, 1/2, 0] = 0$ and $< x >_c [0, 1/2, 0, 1/2] = 1/2$, and at point $\alpha = 1/2$ the two complex detectors differ by $1/8$, i.e., $< x >_c [1/2, 1/4, 1/4, 0] = 1/4$ and $< x >_c [1/2, 1/4, 0, 1/4] = 1/8$. The point $(0, 0)$ marks occurrence of the 'ccgg' string and the point $(0, 1/2)$ marks occurrence of the 'cctt' string. The point $(1/2, 1/4)$ marks occurrence of the 'aacg' string and the point $(1/2, 1/8)$ marks occurrence of the 'aact' string. The third plot (two symbol counts are non-zero) demonstrates different dynamic range of the quaternionic and the two complex detectors, i.e., $< x >_q [1, 0, 0, 0] - < x >_q [1/2, 1/2, 0, 0] = < x >_q [1, 0, 0, 0] - < x >_q [1/2, 0, 0, 1/2] = 2/3$, $< x >_c [1, 0, 0, 0] - < x >_c [1/2, 0, 0, 1/2] = 1$, and $< x >_c [1, 0, 0, 0] - < x >_c [1/2, 1/2, 0, 0] = 1/2$.

Comparison of the quaternionic detector curves in plots a-c illustrates increase in the value of the quaternionic detector (points $(1/4, 0), (1/3, 1/9), (1/2, 1/3)$), as the number of non-zero symbol counts decreases.

Fig. 3 illustrates results of performing complex and quaternionic short time periodicity transform on a random pseudo-DNA sequence with two embedded patterns: $'(agg)_4(ctt)_4'$ and $'(agg)_4(tcc)_4'$. The first pattern contains dominant symbols 'g' and 't', the second pattern contains dominant symbols 'c' and 'g', the periodic content of the two patterns however is identical. Application of the complex detector to the sequence results in detection of the second pattern only; application of the quaternionic detector reveals presence of both patterns.

4. SUMMARY

We have shown that the quaternionic mapping of DNA symbols results in a superior performance of the periodicity transform. We conjecture that an even greater benefit can be derived by employing the quaternionic Fourier transform in the spectral domain approach to DNA pattern detection.

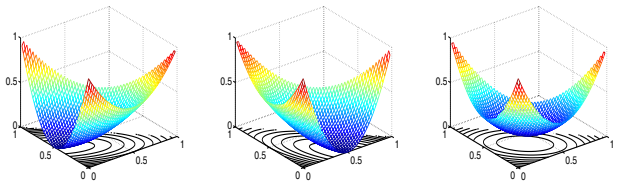

Fig. 1. Performance surfaces of complex and quaternionic detectors: 1) $< x >_c [\alpha, \beta, 1-\alpha-\beta, 0]$, 2) $< x >_c [\alpha, \beta, 0, 1-\alpha-\beta]$ and 3) $< x >_q [\alpha, \beta, 1-\alpha-\beta, 0] = < x >_q [\alpha, \beta, 0, 1-\alpha-\beta]$.

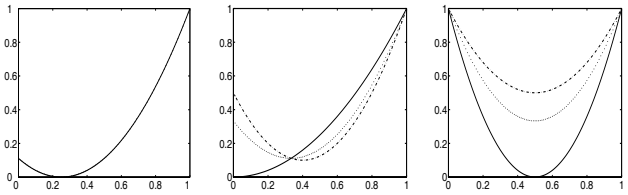

Fig. 2. Performance curves of complex and quaternionic detectors for the following symbol counts: 1) $[\alpha, \frac{1-\alpha}{3}, \frac{1-\alpha}{3}, \frac{1-\alpha}{3}]$, 2) $[\alpha, \frac{1-\alpha}{2}, \frac{1-\alpha}{2}, 0]$ or $[\alpha, \frac{1-\alpha}{2}, 0, \frac{1-\alpha}{2}]$, 3) $[\alpha, 1-\alpha, 0, 0]$ or $[\alpha, 0, 0, 1-\alpha]$, The values of $< x >_c$ (solid and dashed lines) and $< x >_q$ (dotted line) are plotted against the count of symbol 'a', α.

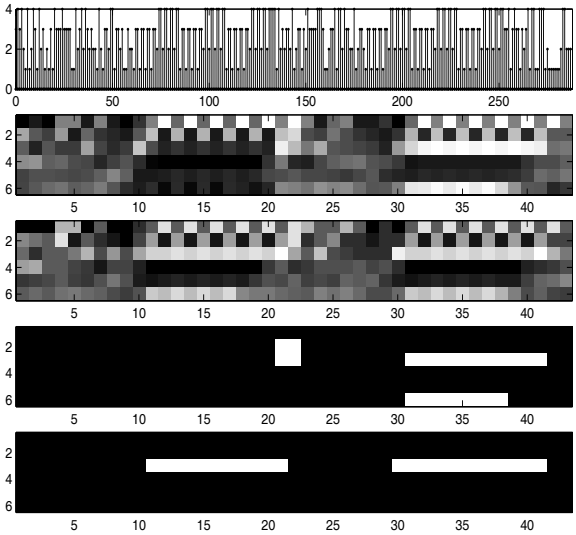

Fig. 3. Random pseudo-DNA sequence with two embedded patterns: $'(agg)_4(ctt)'_4$ and $'(agg)_4(tcc)'_4$ (top plot). Symbols 'a', 'c', 'g' and 't' are marked with stems one, two, three and four units high. Plots two and three are the unprocessed complex and quaternionic PSD's of the DNA sequence. Plots four and five are the complex and quaternionic PSD's after thresholding and an isolated point detection removal. Parameters: $N = 288$, $\bar{N} = 6P$, $0 < P \leq 6$, PSD window shifted in increments of P.

5. ACKNOWLEDGEMENTS

The authors are grateful to John Dileo for discussions of periodic DNA patterns.

6. REFERENCES

[1] D. Anastassiou, "Genomic signal processing", IEEE Trans. SP, Vol. 18, pp. 8-20, July 2001.

[2] A. Arneodo *et al*, "What can we learn with wavelets about DNA sequences?", Physica A, 249, pp. 439-448, 1998.

[3] G. Benson, "Tandem repeat finder: a program to analyze DNA sequences", Nucleic Acid Research, Vol. 27, No. 2, pp. 573-580, 1999.

[4] M. Buchner and S. Janjarasjitt, "Detection and visualization of tandem repeats in DNA sequences", IEEE Trans. SP, Vol. 51, No. 9, pp. 2280-2287, September 2003.

[5] J. Butler, "Forensic DNA typing: biology and technology behind STR markers", Academic Press, 2003.

[6] C. A. Cummings and D. A. Relman, "Microbial Forensics - Cross-Examining Pathogens", Science, Vol. 296, pp.1976-1979, June 2002.

[7] W. R. Hamilton, "Elements of quaternions", London, Longman, 1866.

[8] D. Holste and I. Grosse, "Repeats and correlations in human DNA sequences", Physical Review E, 67, 2003.

[9] I. L. Kantor and A. S. Solodovnikov, "Hypercomplex numbers: an elementary introduction to algebras", New York, Springer-Verlag, 1989.

[10] E. S. Lander *et al*, "Initial sequencing and analysis of the human genome", Nature, Vol. 409, pp. 860-921, February 2001.

[11] J. K. Perkus, "Mathematics of genome analysis", Cambridge University Press, 2002.

[12] S. J. Sangwine, "Fourier transforms of color images using quaternion or hypercomplex numbers", Electron. Lett., Vol. 32, No. 21, pp. 1979-1980, October 1996.

[13] W. A. Sethares and T. W. Staley, "Periodicity Transforms", IEEE Trans. SP, Vol. 47, No. 11, pp. 2953-2964, November 1999.

[14] D. Sidransky, "Nucleic acid-based methods for the detection of cancer", Science, Vol. 278, pp. 1054-1058, November 1997.

[15] G. Sommer (ed.), "Geometric computing with Clifford algebras", New York, Springer-Verlag, 2001.

[16] S. Tavare and B. W. Giddings, "Some statistical aspects of the primary structure of nucleotide sequences", in M. S. Waterman (ed.): Mathematical methods for DNA sequences (pp. 117-131), Boca Raton, CRC Press, 1989.

[17] S. A. Tishkoff *et al*, "Short tandem-repeat polymorphism/alu haplotype variation at the PLAT locus: implications for modern human origins", Am. J. Hum. Genet., 67, pp. 901-925, 2000.

[18] E. N. Trifonov, "3-, 10.5-, 200- and 400-base periodicities in genome sequences", Physica A, 249, pp. 511-516, 1998.