DECOMPOSING STATISTICAL PERIODICITIES

Raman Arora and William Sethares

Department of Electrical and Computer Engineering, University of Wisconsin-Madison, Madison, WI 53706-1691 USA. ramanarora@wisc.edu, sethares@ece.wisc.edu

ABSTRACT

Nonstationary random symbolic sequences are investigated for cyclostationarity and decomposition into constituent cyclostationary sources. Since the symbolic sources do not admit an algebraic structure, a composition of distributions of cyclostationary sources is studied that models the erosion in symbolic sequences, for instance, mutations in gene sequences. This composition gives a rich mathematical structure on the collection of cyclostationary sources and a uniqueness theorem for decomposition of pq-statistically periodic symbolic sources into cyclostationary sources with periods p and q, when p and q are coprime.

Index Terms— Cyclostationarity, symbolic periodicity, symbolic time series, genomic signal processing.

I. INTRODUCTION

SYMBOLIC sequences are time series defined on a finite set with no algebra. In DNA sequences, economic indicator data, and other nominal time series, the only mathematical structure is set membership [1]. An interesting and important behaviour symbolic sequences may exhibit is *periodicity* and finding these periodicities is fundamental to the understanding and determination of the structure of the sequences. In genomic signal processing, for instance, locating hidden periodicities in DNA sequences is a very important task [2]. Repetitions in DNA have been shown to be correlated with several structural and functional roles. For example, base (symbol) periodicity of 21 is associated with α-helical formation for synthesized protein molecules [2] and base periodicity of 3 is identified with protein coding region of the DNA. Such investigations also find applications in diagnosis of genetic disorders (Huntington's disease [3]), DNA forensics, and reconstructing evolutionary history [4].

Symbolic periodicity in DNA sequences comes in several flavors: homologous, eroded, and latent [5]. Homologous periodicities occur when short fragments of DNA are repeated in tandem to give periodic sequences. Imperfect or eroded periodicities [6] result when some of the bases in the homological periodic sequence get replaced or altered so that the tandem repeats are no longer perfect. Latent periodicities [7], [8] occur when the repeating unit is not a fixed sequence but may change in a patterned way (for instance, a DNA sequence in which the nth element is always either A or G) or when there are *indels* (insertions and deletions) in homological periodic sequences. This taxonomy of periodicities is not specific to DNA sequences; it applies to any symbolic sequence.

Most current approaches to detecting periodicities transform the symbolic sequences into a numerical sequence [8], [9], [10]; these techniques are primarily aimed at the detection of homological periodicities [4], [11], [12]. A general approach to the detection of these three classes of periodicities was presented in [13] using a maximum likelihood formulation. In this approach, each element in the DNA sequence is assumed to be generated from an information source with an underlying probability mass function (pmf). The number of sources defines the period and the symbols are drawn from these sources in a cyclic manner. Thus, periodicities in the symbols are represented by repetitions of the pmfs. This can be pictured as in Fig. 1. A rotating carousel (labeled A) contains N_A urns, each with its own distribution of balls (which are labeled A, G, C, or T). At each timestep, a ball is drawn from the urn and the carousel rotates one position. The output of the process is not periodic; rather, the distribution from which the symbols are chosen is periodic. This is called *statistical periodicity* or *strict sense cyclostationarity* [14].

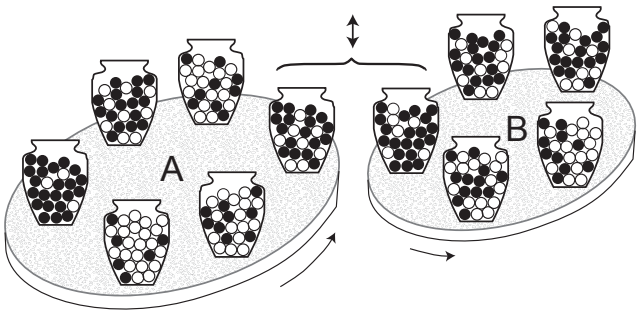

Fig. 1. Each time a ball is removed from one of the N_A urns (indicated by the arrow), platform A rotates, bringing a new urn into position. Similarly, carousel B contains N_B urns, each with its own collection of balls. Draws are made by combining draws from the two aligned urns and results in a $N_A N_B$ statistical periodicity.

Symbolic random variables take values on a finite set called the *alphabet* whose elements are called *symbols*. It is possible to map the symbols to numbers or to define an algebra on the alphabet, but this would impose mathematical structure that was not present to begin with. For instance, the mapping of DNA elements ($T = 0, C = 1, A = 2, G = 3$), suggested in [15], puts an order on the set; the complex representation ($A = 1+j, G = -1+j, C = -1-j, T = 1-j$) used in [12], [8] implies that the euclidean distance between A and C is greater than the distance between A and T [16]. A good survey of various numerical representations for DNA sequences is presented in [17]. Artifacts of such mappings are reported in [11].

In contrast, the formulation in this paper implies no mathematical structure on the alphabet. The focus is on an investigation of multiple periodicities: ways that several periodic symbolic sequences may combine to generate new sequences. In DNA sequences, multiple periodicities have been observed within coding regions [6]. For example, latent periodicities of 120 bases and 126 bases were reported in various genes in [2]. Such longer periods that are multiples of 3 occur in coding regions which have a characteristic period of three. As noted by Korotkov et. al [6], such periodicities can be related to evolutionary origins via multiple duplications.

Multiple periodicities in symbolic random sequences are investigated by defining compositions on the probability measures associated with the sequences. One possibility is to form a Bernoulli mixture of two symbolic sequences; at every instant of time picking symbol from a sequence with probability β and from the other sequence with probability $1-\beta$. If p_t and q_t denote the distributions over the alphabet for the two sequences at instant t, the distribution for the composed sequence is given as $\beta p_t + (1-\beta)q_t$. If the distributions p_t and q_t exhibit periodicities, the Bernoulli mixture may exhibit multiple periodicities. The composition of distributions arises naturally from the underlying experiment, in this case the Bernoulli mixture, and its easily verified that it does not imply an algebraic structure on the alphabet.

This paper focuses on a (different) method of composition that can be interpreted as an erosion or mutation in genes. The corresponding physical experiment is illustrated in Fig. 1, which contains two rotating carousels A and B with N_A and N_B urns respectively. At each timestep, the two carousels rotate into position and an element is drawn from each of the two aligned urns (indicated by the brackets). If the two drawn elements have the same label, the output assumes that label. If the draws give balls with different labels, they are returned to the urns. This continues until an identical pair is drawn. The urns then rotate and the process repeats. The motivation for this model comes from the DNA replication model as discussed in Sect. II. Sect. III shows that this method of composition gives a rich mathematical structure in which to study statistical periodicities with multiple hidden periodicities. The figure shows how two cyclostationary sequences with periods N_A and N_B may combine to form a new sequence with period $N_A N_B$.

Sect. IV investigates the inverse problem: given a cyclostationary symbolic sequence, how can it be decomposed into constituent cyclostationary subsequences. These investigations provide a structured way of attacking the problem of locating such hidden periodicities. While the DNA sequencing application provides motivation for this work, the underlying mathematics is general enough to easily include any symbolic set with any (finite) number of elements.

II. DNA REPLICATION AND MUTATION

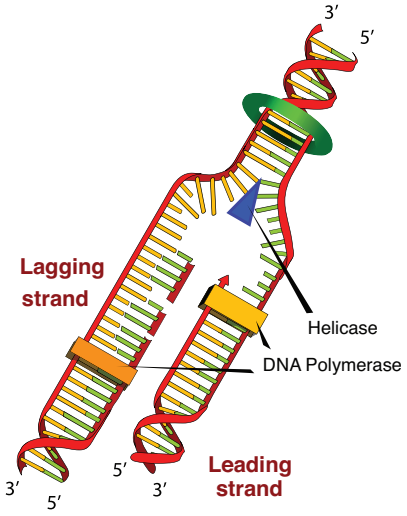

Fig. 2. DNA replication: The two strands of the DNA fork and the polymerase recreates the DNA by attaching the complementary bases to each separated strand thus creating two copies of the original DNA sequence [18].

DNA is a long polymer composed of four repeating bases or nucleotides: A (adenine), G (guanine), T (thymine), and C (cytosine). It exists as a tightly entwined pair of strands in the shape of a helix. The two strands are held together by hydrogen bonds between complementary bases: the nucleotide A in one strand bonding only with T of the other and G bonding only with C. The changes in the base-pair sequences of DNA are mutations which may occur when the DNA is being replicated.

DNA replication begins with helical unwinding. The two strands are pulled apart like a zipper as shown in Fig. 2, resulting into two separate strands. The DNA sequence of the forked strands is recreated by the enzyme *polymerase* in accordance with rules of complementary base pairing. However, substitution errors may occur, and the error rate is approximately 10^{-3} to 10^{-5} [19]. The overall mutation

rate is reported to be 10^{-10}, due largely to the inbuilt proof-reading mechanism of the replication process explained next.

A substitution error in the replication process causes a kink in the DNA sequence due to an imbalance of the sizes of the *purines* (A, G) and the *pyrimidines* (C, T). If a mismatch is detected, the replication stops till the polymerase restores the correct nucleotide [16]. This DNA evolution (or mutation) process provides the motivation for the two-carousel-model above. The original DNA sequence can be modeled as a sequence of non-stationary random variables. Often these sequences exhibit cyclostationarity, for instance the 3-base-pair periodicity is characteristic of the exons. The complementary nucleotides generated by the polymerase (when recreating the forked strands) are modeled as a second sequence of nonstationary random variables. The two sequences may have the same distributions in the absence of the mutations. If there are mutations, the distribution of the secondary sequence is affected and the indels may cause the secondary sequence to exhibit different statistical periodicity. led as a second sequence of nonstationary random variables. The two sequences may have the same distributions in the absence of the mutations. If there are mutations, the distribution of the secondary sequence is affected and the indels may cause the secondary sequence to exhibit different statistical periodicity.

Finally, the composition of pmfs of the two sequences captures the proof-reading mechanism in the DNA replication which progresses only if bases drawn from the two sources are complementary. The two carousel model of Fig. 1 provides a simplified analogy: the former defines an event as identical balls drawn from the two urns, the latter defines an event as complementary base pairs attached to two strands. The analogy is strengthened since each nucleotide uniquely determines the complementary base.

III. PERIODIC SUBSPACES

Let $\mathcal{A} = \{a_1, \ldots, a_M\}$ be a finite set with cardinality M. Let X be an $\mathcal{A}$-valued random variable with probability mass function (pmf) μ_X, i.e. for $a \in \mathcal{A}$, $\mu_X(a)$ denotes the probability $P(X = a)$. Let $\mathbb{X}$ denote the collection of all random variables on the alphabet $\mathcal{A}$. For $n, p \in \mathbb{Z}^+$ let $\hat{n}_p$ denote the positive integer $1 + ((n-1) \mod p)$.

Define a symbolic (random) sequence taking values on the set $\mathcal{A}$ to be a sequence of independent random variables $S : \mathbb{Z}^+ \to \mathbb{X}$. The symbolic sequence S is said to be *p-statistically periodic* if p is a positive integer such that the random variables S_n and $S_{\hat{n}_p}$ are identically distributed for all $n \in \mathbb{Z}^+$. Let $\mathcal{P}_p$ be the collection of $p-$statistically periodic sequences. Then $\mathcal{P} = \bigcup_{p \in \mathbb{Z}^+} \mathcal{P}_p$ is the set of all statistically periodic sequences of random variables on the alphabet $\mathcal{A}$. A sequence S in $\mathcal{P}_p$ can also be described by an $M \times p$ column-stochastic matrix $\mathbf{Q}^S$ whose i^{th} column, denoted $\mathbf{q}_i^S$, gives the pmf of S_{np+i} for all $n \in \mathbb{Z}^+$, i.e.

$$P(S_{np+i} = a_j) = P(S_i = a_j) = \mathbf{Q}^S_{ji} \equiv \mathbf{q}_i^S(j) \qquad (1)$$

where $j = 1, \cdots, M$. Therefore $\mathcal{P}_p$ can also be identified as the set of all $M \times p$ column stochastic matrices. With a slight abuse of notation, the elements of $\mathcal{P}_p$ may be referred to as a random symbolic sequence X or as the corresponding pmf $\mathbf{Q}^X$. The law of composition on the pmfs of the random symbolic sequences that captures the experiment in Fig. 1 defines

$$\begin{array}{rcl} \oplus : \mathcal{P} \times \mathcal{P} & \to & \mathcal{P} \\ (X, Y) & \mapsto & Z \end{array} \qquad (2)$$

on $\mathcal{P}$ as follows. Let $X, Y \in \mathcal{P}$ be sequences with statistical periodicities p and q respectively. Then $Z = X \oplus Y$ is the sequence of random variables such that for all $a \in \mathcal{A}$

$$P\Big(Z_n = a\Big) = P\Big(X_{\hat{n}_p} = a, Y_{\hat{n}_q} = a \Big| X_{\hat{n}_p} = Y_{\hat{n}_q}\Big). \quad (3)$$

Again, this is a slight abuse of notation since the binary operation is defined on the matrices $\mathbf{Q}^X, \mathbf{Q}^Y$ but is expressed in terms of the symbolic sequences X, Y.

Lemma 1. *Let $X \in \mathcal{P}_p$ and $Y \in \mathcal{P}_q$. Let $Z = X \oplus Y$. Then $Z \in \mathcal{P}_r$, where r is the lowest common multiple of p and q.*

Proof: Let $m = n + rs$ where r is the lowest common multiple of p and q and s is any positive integer. Then $\hat{m}_p = \hat{n}_p$ and $\hat{m}_q = \hat{n}_q$. Thus for all $a \in \mathcal{A}$, $P\Big(Z_m = a\Big) = P\Big(X_{\hat{n}_p} = a, Y_{\hat{n}_q} = a \Big| X_{\hat{n}_p} = Y_{\hat{n}_q}\Big) = P\Big(Z_n = a\Big)$. ∎

Corollary 1. *Let $X, Y \in \mathcal{P}_p$. Then $X \oplus Y$ is p-statistically periodic.*

In Lemma 1, if p and q are mutually prime then $Z \in \mathcal{P}_{pq}$. If $\mathbf{Q}^X, \mathbf{Q}^Y$ and $\mathbf{Q}^Z$ denote the stochastic matrices of X, Y and Z, respectively, then by definition (3), the n^{th} column of the $M \times pq$ matrix $\mathbf{Q}^Z$ is

$$\mathbf{q}_n^Z = \frac{1}{C} \begin{bmatrix} \mathbf{q}^X_{\hat{n}_p}(1)\mathbf{q}^Y_{\hat{n}_q}(1) \\ \vdots \\ \mathbf{q}^X_{\hat{n}_p}(M)\mathbf{q}^Y_{\hat{n}_q}(M) \end{bmatrix} \qquad (4)$$

where $C = \sum_{j=1}^M \mathbf{q}^X_{\hat{n}_p}(j)\mathbf{q}^Y_{\hat{n}_q}(j)$ is the normalization factor.

Example 1. *Consider an example of composition of two cyclostationary sources with statistical periods 2 and 3. Eqn. (4) gives*

$$\underbrace{\begin{bmatrix} .25 & .6 \\ .25 & .2 \\ .25 & .1 \\ .25 & .1 \end{bmatrix}}_{X \in \mathcal{P}_2} \oplus \underbrace{\begin{bmatrix} .3 & .1 & 1 \\ 0 & .1 & 0 \\ .3 & .2 & 0 \\ .4 & .6 & 0 \end{bmatrix}}_{Y \in \mathcal{P}_3}$$

$$= \underbrace{\begin{bmatrix} 0.3 & 0.375 & 1 & 0.72 & 0.1 & 1 \\ 0 & 0.125 & 0 & 0 & 0.1 & 0 \\ 0.3 & 0.125 & 0 & 0.12 & 0.2 & 0 \\ 0.4 & 0.375 & 0 & 0.16 & 0.6 & 0 \end{bmatrix}}_{Z \in \mathcal{P}_6}$$

Note that the first source in the sequence X is the identity of the composition and the last source of the sequence Y acts like an infinity of the operation. ■

If $X = Y$, then $Z = X \oplus Y$ is in $\mathcal{P}_p$ with

$$\mathbf{q}_n^Z(k) = (\mathbf{q}_n^X(k))^2 / \sum_{j=1}^{M} (\mathbf{q}_n^X(k))^2,$$

for $k = 1, \ldots, M$ and $n = 1, \ldots, p$. The operation of composing a symbolic sequence with itself can also be expressed as multiplication by the scalar 2; write $Z = X \oplus X = 2 \circ X$. This definition can be extended to multiplication by any scalar. For $r \in \mathbb{R}$ and $X \in \mathcal{P}$ define

$$\begin{array}{rcl} \circ : \mathbb{R} \times \mathcal{P} & \rightarrow & \mathcal{P} \\ (r, X) & \mapsto & Z \end{array} \tag{5}$$

so that $Z = r \circ X$ is the random symbolic sequence with

$$P\big(Z_n = a\big) = \frac{P(X_n = a)^r}{\sum_{b \in \mathcal{A}} P(X_n = b)^r} \tag{6}$$

for all $a \in \mathcal{A}$ with $P(X_n = a) \neq 0$. When $P(X_n = a) = 0$, $P(Z_n = a)$ is defined to be 0. If $X \in \mathcal{P}_p$, $Z \in \mathcal{P}_p$.

Example 2. *Consider an example of scalar multiplication. Let X be a cyclostationary symbolic sequence with X_i distributed as $\mathbf{q}_i^X = [\frac{1}{2} \ \frac{1}{4} \ \frac{1}{4} \ 0]^T$. If $Y = 2 \circ X$ then Y_i is distributed as $\mathbf{q}_i^Y = [\frac{2}{3} \ \frac{1}{6} \ \frac{1}{6} \ 0]^T$.*

Theorem 1. *The set $\mathcal{P}$ forms an abelian group under the binary operation $\oplus : \mathcal{P} \times \mathcal{P} \rightarrow \mathcal{P}$.*

Proof: The closure of $\mathcal{P}$ under $\oplus$ follows by Lemma 1 and the operation is commutative by definition. Associativity is easy to check: let $X, Y, Z \in \mathcal{P}$ have statistical periodicities p, q and r respectively. Let $V = X \oplus (Y \oplus Z)$ and $W = (X \oplus Y) \oplus Z$. Then $\mathbf{Q}_{ji}^V$ can be rewritten as

$$\frac{\mathbf{Q}_{ji_p}^X \left(\mathbf{Q}_{ji_q}^Y \mathbf{Q}_{ji_r}^Z \right)}{\sum_j \mathbf{Q}_{ji_p}^X \left(\mathbf{Q}_{ji_q}^Y \mathbf{Q}_{ji_r}^Z \right)} = \frac{\left(\mathbf{Q}_{ji_p}^X \mathbf{Q}_{ji_q}^Y \right) \mathbf{Q}_{ji_r}^Z}{\sum_j \left(\mathbf{Q}_{ji_p}^X \mathbf{Q}_{ji_q}^Y \right) \mathbf{Q}_{ji_r}^Z} = \mathbf{Q}_{ji}^W$$

for $j = 1, \ldots, M$ and $i = 1, \ldots, pq$. The unique identity element, denoted E, is the stationary or 1-statistically periodic random sequence such that $P(E = a_j) = \frac{1}{M}$ for all $a_j \in \mathcal{A}$. Finally, for $X \in \mathcal{P}$ if $Y = (-1) \circ X$ then it is easy to verify that $X \oplus Y = E$. Thus every $X \in \mathcal{P}$ has an inverse. ■

Theorem 2. *$(\mathcal{P}, \oplus, \circ)$ is a vector space over $\mathbb{R}$.*

Proof: The closure of $\mathcal{P}$ under $\circ$ follows by definition and the identity element is $1 \in \mathbb{R}$ since $1 \circ X = X$. The distributive properties are easy to check: for $\alpha \in \mathbb{R}$, $X \in \mathcal{P}_p$ and $Y \in \mathcal{P}_q$, $\alpha \circ (X \oplus Y) = (\alpha \circ X) \oplus (\alpha \circ Y)$ and for $\alpha, \beta \in \mathbb{R}$ and $X \in \mathcal{P}_p$, $(\alpha + \beta) \circ X = (\alpha \circ X) \oplus (\beta \circ X)$. Finally, scalar multiplication is compatible with multiplication in the field of scalars: $\alpha \circ (\beta \circ X) = (\alpha\beta) \circ X$. ■

Corollary 2. *For $p \in \mathbb{Z}^+$, $\mathcal{P}_p$ is a subspace of $\mathcal{P}$.*

IV. DECOMPOSING PERIODICITIES

A fundamental problem in symbolic signal processing [8] is identifying the periodic structure of the symbolic sources. Given a realization of the symbolic sequence, the maximum likelihood estimates of the statistical periodicity and the corresponding pmf were derived in [13]. This section investigates the problem of decomposing the discovered symbolic source into various smaller components.

Assume that an observed sequence $Z \in \mathcal{P}_{pq}$ was originally composed of sequences $X \in \mathcal{P}_p$ and $Y \in \mathcal{P}_q$, i.e. $Z = X \oplus Y$. Then $Z_n = X_{\hat{n}_p} \oplus Y_{\hat{n}_q}$, for $n = 1, \ldots, pq$. The pq equations can be expressed in matrix form as

$$\begin{bmatrix} Z_1 \\ \vdots \\ Z_{pq} \end{bmatrix}_{pq \times 1} = \underbrace{\begin{bmatrix} I_p & I_q \\ \vdots & \vdots \\ I_p & I_q \end{bmatrix}}_{\mathbf{T}_{pq \times (p+q)}} \circ \begin{bmatrix} X_1 \\ \vdots \\ X_p \\ Y_1 \\ \vdots \\ Y_q \end{bmatrix}_{(p+q) \times 1} . \tag{7}$$

Theorem 3. *For mutually prime p and q, the matrix $\mathbf{T}$ above has rank $p + q - 1$. The null space of $\mathbf{T}$ is spanned by the vector $[\underbrace{-1 \ldots -1}_{p} \underbrace{1 \ldots 1}_{q}]$*

Theorem 3 shows that if $Z \in \mathcal{P}_{pq}$ can be decomposed into $Z = X \oplus Y$ for some $X \in \mathcal{P}_p$ and $Y \in \mathcal{P}_q$, then it can also be decomposed as

$$(X \oplus \delta_p) \oplus (Y \ominus \delta_q) = Z$$

where $Y \ominus \delta_q = Y \oplus (-1 \circ \delta_q)$ and $\delta_r = [\overbrace{\delta, \ldots, \delta}^{r}]$ for some $\delta \in \mathcal{P}_1$ and $r = p, q$. Thus there is a class of decompositions of Z. In words, a pq-periodic symbolic source Z can be decomposed into p and $q-$periodic components X, Y unique up to an additive factor $\delta \in \mathcal{P}_1$.

Example 3. *With the same X and Y as in example 1,*

$$\underbrace{\begin{bmatrix} 2/10 & 12/23 \\ 3/10 & 6/23 \\ 3/10 & 3/23 \\ 2/10 & 2/23 \end{bmatrix}}_{X' = X \oplus \delta} \oplus \underbrace{\begin{bmatrix} 1/3 & 1/9 & 1 \\ 0 & 2/27 & 0 \\ 2/9 & 4/27 & 0 \\ 4/9 & 2/3 & 0 \end{bmatrix}}_{Y' = Y \ominus \delta = Y \oplus (-1 \circ \delta)}$$

$$= \underbrace{\begin{bmatrix} 0.3 & 0.375 & 1 & 0.72 & 0.1 & 1 \\ 0 & 0.125 & 0 & 0 & 0.1 & 0 \\ 0.3 & 0.125 & 0 & 0.12 & 0.2 & 0 \\ 0.4 & 0.375 & 0 & 0.16 & 0.6 & 0 \end{bmatrix}}_{Z}$$

where $\delta = \left[\frac{2}{10} \ \frac{3}{10} \ \frac{3}{10} \ \frac{2}{10}\right]^T$ and $-1 \circ \delta = \left[\frac{3}{10} \ \frac{2}{10} \ \frac{2}{10} \ \frac{3}{10}\right]^T$ ■

V. DISCUSSION

The investigation of multiple periodicities in symbolic sequences and their decomposition into smaller periodic components may be useful as a way to understand the underlying generative mechanism. For instance, in DNA sequences, the decomposition may provide insight into the underlying evolutionary process that determines the structure of the sequences. The investigation is challenging at least in part due to the lack of an algebraic structure. The approach used here models the symbolic sequence as a nonstationary random process on a finite alphabet and then studies the (de)composition of the distributions. In particular, the decomposition of DNA sequences are studied under a composition rule that is inspired by the biological model for gene replication and mutation.

The formulation of the problem in this paper is different from the classical stochastic techniques where distributions are estimated by averaging over various ensembles or realizations. Often, it is impractical or impossible to obtain more than one realization and an engineer's solution is to perform averaging over single realization of data. This temporal averaging may be justified when the data exhibits cyclostationarity over long periods or when it is reasonable to assume ergodicity. An interesting discussion about the two approaches may be found in [20].

VI. REFERENCES

[1] Wei Wang and Don H Johnson, "Computing linear transforms of symbolic signals," *IEEE Trans. On Signal Processing*, vol. 50, no. 3, pp. 628–634, March 2002.

[2] E V Korotkov and N Kudryaschov, "Latent periodicity of many genes," *Genome Informatics*, vol. 12, pp. 437 – 439, 2001.

[3] The Huntington's Disease Collaborative Research Group, "A novel gene containing a trinucleotide repeat that is expanded and unstable on huntington's disease chromosomes," *Cell*, vol. 72, pp. 971–983, March 1993.

[4] Andrzej K. Brodzik, "Quaternionic periodicity transform: an algebraic solution to the tandem repeat detection problem," *Bioinformatics*, vol. 23, no. 6, pp. 694–700, Jan 2007.

[5] M B Chaley, E V Korotkov, and K G Skryabin, "Method revealing latent periodicity of the nucleotide sequences modified for a case of small samples," *DNA Research*, vol. 6, pp. 357 – 363, Feb. 1999.

[6] E V Korotkov and D A Phoenix, "Latent periodicity of DNA sequences of many genes," in *Proceedings of Pacific Symposium on Biocomputing 97*, 1997, pp. 222–229.

[7] E V Korotkov and M A Korotova, "Latent periodicity of dna sequences of some human genes," *DNA Sequence*, vol. 5, pp. 353, 1995.

[8] Dimitris Anastassiou, "Genomic signal processing," *IEEE Signal Processing*, vol. 18, pp. 8–20, Jul 2001.

[9] V R Chechetkin, L A Knizhnikova, and A Yu Turygin, "Three-quasiperiodicity, mutual correlatuions, ordering and long modulations in genomic nucleotide sequences viruses," *Journal of biomolecular structure and dynamics*, vol. 12, pp. 271, 1994.

[10] E A Cheever, D B Searls, W Karunaratne, and G C Overton, "Using signal processing techniques for dna sequence comparison," in *Proc. of the 1989 Fifteenth Annual Northeast Bioengineering Conference*, Boston, MA, Mar 1989, pp. 173 – 174.

[11] Ravi Gupta, Divya Sarthi, Ankush Mittal, and Kuldip Singh, "Exactly periodic subspace decomposition based approach for identifying tandem repeats in dna sequences," in *Proc. of the* 14^{th} *European Signal Processing Conference (EUSIPCO)*, Florence, Italy, Sep 2006.

[12] Marc Buchner and Suparerk Janjarasjitt, "Detection and visualization of tandem repeats in dna sequences," *IEEE Transactions on Signal Processing*, vol. 51, no. 9, pp. 2280–2287, Sep 2003.

[13] Raman Arora and William A. Sethares, "Detection of periodicities in gene sequences: a maximum likelihood approach," in *GENSIPS (Genomic Signal Processing and Statistics)*, Tuusula, Finland, June 2007.

[14] W. A. Gardner, A. Napolitano, and L. Paura, "Cyclostationarity: half a century of research," *Signal Processing*, vol. 86, pp. 639–697, 2006.

[15] P. D. Cristea, "Genetic signal representation and analysis," in *Proceeding SPIE Conference, International Biomedical Optics Symposium (BIOS02)*, 2002, pp. 77–84.

[16] Gail L. Rosen, *Signal Processing for biologically-inspired gradient source localization and DNA sequence anlysis*, PhD Thesis, Georgia Institute of Technology, 2006.

[17] Mahmood Akhtar, Julien Epps, and Eliathamby Ambikairajah1, "On dna numerical representations for period-3 based exon prediction," in *GENSIPS (Genomic Signal Processing and Statistics)*, Tuusula, Finland, June 2007.

[18] DNA, [Online]. Available on Wikipedia at http://en.wikipedia.org/wiki/DNA.

[19] Roy H Burdon, *Genes and the Environment*, Taylor and Francis Inc., PA, 1999.

[20] William A. Gardner, *Cyclostationarity in Communications and Signal Processing*, IEEE press, NY, 1994.