Latent Periodicities in Genome Sequences

Raman Arora, *Student Member, IEEE*, William A. Sethares, *Member, IEEE*, and James A. Bucklew, *Senior Member, IEEE*

***Abstract*—A novel approach is presented for the detection of periodicities in DNA sequences. A DNA sequence can be modelled as a nonstationary stochastic process that exhibits various statistical periodicities over different regions. The coding part of the DNA, for instance, exhibits statistical periodicity with period three. Such regions in DNA are modelled as generated from a collection of information sources (with an underlying probability distribution) in a cyclic manner, thus exhibiting cyclostationarity. The maximum likelihood estimates are developed for the distributions of the information sources and for the statistical period of the DNA sequence. Such probabilistic sources are further investigated for decomposition into constituent cyclostationary sources. Since the symbolic sources do not admit an algebraic structure, a composition of cyclostationary probabilistic sources is studied that models DNA replication process. This composition is shown to give a rich mathematical structure on the collection of cyclostationary sources and allows a uniqueness theorem for the decomposition.**

***Index Terms*—Cyclostationarity, gene replication, genomic signal processing, symbolic periodicity, symbolic sequences.**

I. Introduction

SYMBOLIC sequences consist of strings of elements (or symbols) drawn from a finite set (or alphabet), typically with no algebraic structure. In DNA sequences, economic indicator data, and other nominal time series, the only mathematical structure is the set membership [1]. Such symbolic sequences may exhibit various kinds of repetitions and regularities, and finding such features is fundamental to understanding the structure of the sequences. In genomic signal processing, latent periodicities in DNA sequences have been shown to be correlated with several structural and functional roles [2]. For example, a base (symbol) periodicity of 21 is associated with α-helical formation for synthesized protein molecules [2] and a base periodicity of three is identified with protein coding regions of the DNA. Such investigations also find application in the diagnosis of genetic disorders like Huntington's disease [3], DNA forensics and in the reconstruction of evolutionary history [4], [5].

Symbolic periodicities in DNA sequences may be classified into homologous, eroded and latent [6]. Homologous periodicities occur when short fragments of DNA are repeated in tandem to give periodic sequences. Imperfect or eroded periodicities [7] result when some of the bases in the homologous sequence get replaced or undergo insertions and deletions, so that the tandem repeats are not identical. Latent periodicities [8], [9] occur when the repeating unit is not fixed but may change in a patterned way. For instance, an observed latent period of nucleotides may be

$$[(\mathrm{A/C})\,(\mathrm{T/G})\,(\mathrm{T/A})\,(\mathrm{G/T})\,(\mathrm{C/G/A})\,(\mathrm{G/A})] \qquad (1)$$

which specifies the first element as either A or C, the second as either T or G, and so on. The latent periods in DNA sequences may provide insights into the changing nature of the sequences. For instance, in mRNA, the latent period [GCU] is believed to be a sequence fossil of ancient codons which dominated the earliest stages of evolution [10].

Most current approaches to detecting periodicities transform the symbolic sequences into numerical sequences and compute Fourier transform [9], [11]–[13] or perform a periodic subspace decomposition (EPSD) [14]. Though this is computationally convenient, it imposes a mathematical structure that is not present in the data. For instance, the mapping of DNA elements ($\mathrm{T}=0$, $\mathrm{C}=1$, $\mathrm{A}=2$, $\mathrm{G}=3$) suggested in [15] puts a total order on the set; the complex representation ($\mathrm{A}=1+j$, $\mathrm{G}=-1+j$, $\mathrm{C}=-1-j$, $\mathrm{T}=1-j$) used in [9], [16] implies that the euclidean distance between A and C is greater than the distance between A and T [17]. Such numerical mappings may introduce artifacts in the spectrum of the sequence. For example, consider the symbolic sequence ACTACTACTACT with the numerical representation ($\mathrm{T}=0$, $\mathrm{C}=1$, $\mathrm{A}=2$, $\mathrm{G}=3$). Due to the order present in the numerical representation, a mutation of any symbol to G results in larger noise than other mutations. If the first and the last occurrence of T both flip to G, the spectral energy leaks from the bin corresponding to period three resulting in a dominant peak corresponding to period two. Similar artifacts may occur in the presence of noise for other representations; also see [14]. A survey of various numerical mappings for DNA sequences is presented in [18], most of which are aimed primarily at the detection of homological periodicities [5], [14], [16].

In contrast, the formulation in this paper implies no mathematical structure on the alphabet and presents a general approach to the detection of periodicities. Each symbol of the sequence is assumed to be generated by an information source with some underlying probability mass function (pmf). The sequence is generated by drawing symbols from a collection of such 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 may be 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; instead, the distribution from which the symbols are chosen is periodic. This is called *statistical*

Manuscript received September 7, 2007; revised March 8, 2008. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ioan Tabus.

The authors are with the Department of Electrical and Computer Engineering, University of Wisconsin, Madison, WI 53706-1691 USA (e-mail: ramanarora@wisc.edu; sethares@ece.wisc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2008.923861

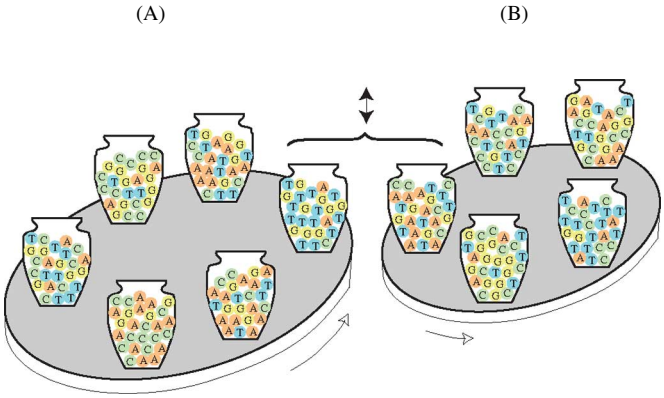


Fig. 1. Each time a ball is removed from one of the N_A urns (indicated by the arrow), platform A rotates, bringing a new urn into position. Similarly, carousel B contains N_B urns, each with its own collection of balls. The urns are the information sources and the cyclostationary sequences generated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

periodicity or *strict sense cyclostationarity* [19]. The number of sources is equal to the latent period in the sequence. The carousel model is a special case of a first order hidden Markov model in which the path is deterministic since a given state transitions with probability one to the next state in the cycle. The model captures all three notions of periodicities in symbolic sequences: tandem repeats result from information sources with trivial zero-one pmfs while the eroded and latent periodicities correspond to a larger and more general class of pmfs.

Multiple periodicities have also been observed in DNA sequences [7]. Latent multiple periodicities of 120 and 126 base-pairs were reported in various genes in [2]. Such longer periods that are multiples of three tend to occur in coding regions. As noted by Korotkov *et al.* [7], these periodicities can be related to evolutionary origins via multiple duplications. Multiple periodicities in symbolic random sequences can be investigated by defining compositions on the probability distributions associated with the sequences. One possibility is to form a Bernoulli mixture of two symbolic sequences; for each base location pick a symbol from the first sequence with probability β and from the other with probability $1-\beta$. If p_t and q_t denote the distributions over the common alphabet for the two sequences at location 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 parameter β itself may vary with base location. This composition arises naturally from the underlying experiment (in this case the Bernoulli mixture) and the binary operation is easily extended to a finite number of sequences. However, the operation is not associative and the order in which the sequences are composed determines the outcome.

This paper presents a (different) method of composition in analogy with the DNA replication process. The corresponding physical experiment, illustrated in Fig. 1, comprises of simultaneous draws from 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 bracket). If elements with different labels are drawn, they are returned to the urns and the draws continue until an identical pair is drawn. If the drawn elements have the same label, the output assumes that label. The urns then rotate and the process repeats. The motivation for this model comes from the DNA replication process. DNA exists as a tightly entwined pair of strands in the shape of a helix. DNA replication begins with helical unwinding in which the two strands are pulled apart like a zipper 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 [20]. 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 [17]. The analogy between DNA replication and the two carousel model is that the former defines an event as recreation of complementary base pairs by the action of polymerase and the latter defines an event as identical balls drawn from the two urns. The analogy is strengthened since each nucleotide uniquely determines the complementary base. The evolved DNA sequence results from combination of the strand of the original sequence with the recreated sequence of complementary nucleotides. However, even with error-correction mechanism in operation, mutations occur in the two strands being combined in the form of base-mispairing, replication slippage, insertions and deletions. In fact, a major fraction of total mutations in the genome are caused during the replication process. These mutations manifest themselves by altering the statistical periodicity profile of the single strands being combined. If the statistical periodicities vary significantly for the two strands, the resulting sequence may exhibit multiple cyclostationarities. This method of composition defines a rich mathematical framework (as detailed in Section IV) in which to study multiple latent cyclo-stationarities. The composition law is associative thus making the extension to a finite number of sequences trivial and the order of composition irrelevant.

The paper is organized as follows. The problem of detecting latent periodicities in symbolic sequences is formulated mathematically in the next section. The maximum likelihood estimate of the dominant period is developed in Section II-A. The estimates are improved by incorporating a complexity term with the likelihood function in Section II-B. The penalized maximum likelihood estimator is justified by the application of minimum description length (MDL) principle to the model selection problem. Section III presents experimental results with the proposed algorithm applied to finding periodicities and localizing periodic segments in both simulated sequences and DNA sequence data. Section IV focuses on investigation of multiple periodicities and composition of sequences as discussed above. The corresponding inverse problem, how a cyclostationary symbolic sequence can be decomposed into constituent cyclostationary subsequences, is also addressed. While the DNA sequences provide motivation for this work, the underlying mathematics is general enough to easily include any symbolic set with finite number of elements. Some parts of this paper were previously presented in [21]–[23].

II. Statistical Periodicity

Let $\mathcal{A} = \{a_1, \ldots, a_L\}$ be a finite alphabet of size L. For DNA sequences, $\mathcal{A} = \{\mathrm{A}, \mathrm{G}, \mathrm{C}, \mathrm{T}\}$ where the symbols represent nucleotides Adenine, Guanine, Cytosine and Thymine respectively. Let $\mathcal{A}^n$ denote the n-fold cartesian product of $\mathcal{A}$ and $x^n \in \mathcal{A}^n$ denote a sequence of length n.

A *probabilistic source* is defined as a sequence of probability distributions $P^{(1)}, P^{(2)}, \ldots$ on corresponding sequence of alphabets $\mathcal{A}^1, \mathcal{A}^2, \ldots$ such that for all n, and for all $x^n \in \mathcal{A}^n$, $P^{(n)}(x^n) = \sum_{y \in \mathcal{A}} P^{(n+1)}(x^n, y)$. Let $X^N = X_1 X_2 \ldots X_N$ be a sequence of $\mathcal{A}$-valued random variables (or information sources) corresponding to the probabilistic source $P^{(N)}$. A realization of X^N is a symbolic sequence $\mathbf{s} = \mathbf{s}_1 \mathbf{s}_2 \ldots \mathbf{s}_N$ of length N. If the symbolic sequence $\mathbf{s}$ is generated by repeated concatenation of realizations of a probabilistic source $P^{(K)}$, the *statistical period* of $\mathbf{s}$ is defined to be K. In other words, a *cyclostationary* symbolic sequence $\mathbf{s}$ with period K is generated by K information sources denoted as $X_1, \ldots, X_K$, in a cyclic fashion. X_i's are states of a first order hidden Markov model with the trivial cyclic transition matrix (since X_i transitions to X_{i+1} with probability one). Consequently, the likelihood of observing a sequence $\mathbf{s}$ can be expressed solely in terms of the emission probabilities of the states. The emission probabilities of X_i are described by a probability mass function P_i; X_i emits the j^{th} symbol in $\mathcal{A}$ with probability $P_i(j) = P(X_i = a_j)$ for $j = 1, \ldots, |\mathcal{A}|$ where $|\mathcal{A}|$ is the cardinality of the alphabet (four for DNA sequences). Collecting the $|\mathcal{A}| \times 1$ dimensional vectors P_i into a matrix $\mathbf{Q}^{(k)} = [P_1, \ldots, P_k]$ gives a compact description of the k-periodic cyclostationary source $P^{(n)}$ for all n.

The *dominant period* of a K-periodic cyclostationary sequence is defined to be the substring of consensus bases in a period. It is described by the symbolic sequence $\mathbf{s}^* = \mathbf{s}_1^* \ldots \mathbf{s}_K^*$ of length K such that the i^{th} symbol in every period is more likely to be $\mathbf{s}_i^*$ than any other symbol from the alphabet. Mathematically, $\mathbf{s}_i^* = \mathcal{A}_{i^*}$ where $i^* = \arg\max_{1 \le j \le |\mathcal{A}|} P_i(j)$. If $\mathbf{s}_i^*$ is not unique then the following notation is adopted: the dominant period $[\mathrm{A}(\mathrm{G}/\mathrm{C})(\mathrm{T})]$ denotes a 3-periodic cyclostationary sequence where the first symbol is most likely A, the second symbol is equally likely to be a G or C and the third symbol is always a T.

Let K denote the true period and k be the hypothesized period. The number of complete statistical periods in an N-symbol long k-periodic cyclostationary sequence $\mathbf{s}$ are $M = \lfloor N/k \rfloor$, where $\lfloor x \rfloor$ denotes the largest integer less than or equal to x. Define

$$\lfloor i \rfloor_k = 1 + ((i-1) \bmod k) \tag{2}$$

where $(x \bmod \mathrm{y})$ denotes the remainder after division of x by y. Then for $1 \le i \le N$, the symbol $\mathbf{s}_i$ is generated by the random variable $X_{\lfloor i \rfloor_k}$. The parameters, period k and pmfs $P_1, \ldots, P_k$ of corresponding information sources, are unknown. The search space for k is the set $\mathcal{K} = \{1, \ldots, N_0\}$, for some $N_0 < N$ and for corresponding probabilistic source $\mathrm{Q}^{(k)}$ the search space is the subset $\mathcal{Q}^{(k)} \subseteq [0,1]^{|\mathcal{A}| \times k}$ of column stochastic matrices (for $\mathcal{Q} \in \mathcal{Q}^{(k)}$, $\mathcal{Q}_{ji} \in [0,1]$ and $\sum_{j=1}^{|\mathcal{A}|} \mathbf{Q}_{ji} = 1$ for $i = 1, \ldots, k$).

The likelihood of observing the sequence $\mathbf{s}$ is given as

$$P(\mathbf{s}|k, \mathbf{Q}^{(k)}) = \prod_{i=1}^{N} P(X_{\lfloor i \rfloor_k} = \mathbf{s}_i | k, \mathbf{Q}^{(k)}). \tag{3}$$

Conditioned on k, the maximum likelihood estimate of $\mathbf{Q}^{(k)}$ can be expressed as

$$\mathbf{Q}_{\mathrm{ML}}^{(k)} = \arg \max_{\mathbf{Q} \in \mathcal{Q}^{(k)}} P(\mathbf{s}|k, \mathbf{Q}). \tag{4}$$

Finally, the plug-in maximum likelihood estimate for K is computed as

$$\begin{aligned} K_{\mathrm{ML}} &= \arg \max_{k \in \mathcal{K}} P(\mathbf{s}|k, \mathbf{Q}_{\mathrm{ML}}^{(k)}) \\ &= \arg \max_{k \in \mathcal{K}} \prod_{i=1}^{N} \sum_{j=1}^{|\mathcal{A}|} \left[\mathbf{Q}_{\mathrm{ML}}^{(k)}\right]_{j, \lfloor i \rfloor_k} \cdot \mathbf{1}\{\mathbf{s}_i = \mathcal{A}_j\} \end{aligned} \tag{5}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function and $[\mathbf{Q}]_{j,i}$ denotes the $(j,i)^{th}$ element of the matrix $\mathbf{Q}$. The maximum likelihood estimates (MLE) for the unknown parameters are developed in the next section. However, as seen from the experimental results on simulated and real DNA sequences, the MLE tends to overfit the data. To address the problem of over-fitting, a penalized maximum likelihood estimator is suggested in Section II-B using the refined minimum description length (MDL) principle.

A. The Maximum Likelihood Estimate

The derivation of the MLE is simplified by representing the data-sequence $\mathbf{s} = \mathbf{s}_1 \ldots \mathbf{s}_N$ by a sequence of vectors $\mathbf{w} = \mathbf{w}_1 \ldots \mathbf{w}_N$ where each $\mathbf{w}_i$ is a $|\mathcal{A}| \times 1$ column-vector with j^{th} entry equal to one if $\mathbf{s}_i = \mathcal{A}_j$ and zero otherwise. For DNA sequences, if the i^{th} symbol in the sequence $\mathbf{s}$ is C, i.e., the third symbol of the alphabet $\mathcal{A}$, then the i^{th} vector $\mathbf{w}_i$ in the sequence $\mathbf{w}$ is $[0\ 0\ 1\ 0]'$. The $|\mathcal{A}| \times k$ stochastic matrix $\mathbf{Q}^{(k)}$ comprises of entries $\mathbf{Q}_{ji}^{(k)} = P(X_i = \mathcal{A}_j)$. The columns of the matrix $\mathbf{Q}^{(k)}$ denote the pmfs of the information sources; the entry $\mathbf{Q}_{ji}^{(k)}$ denotes the probability that the i^{th} source generates the j^{th} symbol of the alphabet $\mathcal{A}$. Then

$$P(X_{\lfloor i \rfloor_k} = \mathbf{s}_i | k, \mathbf{Q}^{(k)}) = \prod_{j=1}^{|\mathcal{A}|} (\mathbf{Q}_{j \lfloor i \rfloor_k}^{(k)})^{\mathbf{w}_{ji}}. \tag{6}$$

The likelihood (3) can therefore be written as

$$\begin{aligned} P(\mathbf{w}|k, \mathbf{Q}^{(k)}) &= \prod_{i=1}^{N} P(X_{\lfloor i \rfloor_k} = \mathbf{s}_i | k, \mathbf{Q}^{(k)}) \\ &= \prod_{i=1}^{N} \prod_{j=1}^{|\mathcal{A}|} \left(\mathbf{Q}_{j \lfloor i \rfloor_k}^{(k)}\right)^{\mathbf{w}_{ji}} \\ &= \prod_{m=1}^{M} \prod_{\lfloor i \rfloor_k = 1}^{k} \prod_{j=1}^{|\mathcal{A}|} \left(\mathbf{Q}_{j \lfloor i \rfloor_k}^{(k)}\right)^{\mathbf{w}_{j \lfloor i \rfloor_k^m}} \\ &\quad \times \prod_{\lfloor i \rfloor_k = 1}^{N-Mk} \prod_{j=1}^{|\mathcal{A}|} \left(\mathbf{Q}_{j \lfloor i \rfloor_k}^{(k)}\right)^{\mathbf{w}_{j \lfloor i \rfloor_k^{M+1}}} \end{aligned} \tag{7}$$

where $\lfloor i \rfloor_k^m = (m-1)k + \lfloor i \rfloor_k$. The first term on the right hand side of (7) captures the observations in M complete periods (given the period k) while the second product captures the observations over the last incomplete cycle. The corresponding log-likelihood is

$$\log P(\mathbf{w}|\mathbf{Q}^{(k)}, k) = \sum_{m=1}^{M} \sum_{\lfloor i \rfloor_k = 1}^{k} \sum_{j=1}^{|\mathcal{A}|} \mathbf{w}_{j \lfloor i \rfloor_k^m} \log \left(\mathbf{Q}^{(k)}_{j \lfloor i \rfloor_k} \right) + \sum_{\lfloor i \rfloor_k = 1}^{N-Mk} \sum_{j=1}^{|\mathcal{A}|} \mathbf{w}_{j \lfloor i \rfloor_k^{M+1}} \log \left(\mathbf{Q}^{(k)}_{j \lfloor i \rfloor_k} \right). \quad (8)$$

The MLE for $\mathbf{Q}^{(k)}$ is first derived and then substituted in (5) to form the plug-in maximum-likelihood-estimator for K. For a fixed k, the MLE for $\mathbf{Q}^{(k)}$ is given as

$$\mathbf{Q}^{(k)}_{\mathrm{ML}} = \arg \max_{\mathbf{Q}^{(k)} \in \mathcal{Q}^{(k)}} \log P(\mathbf{w}|k, \mathbf{Q}^{(k)}). \quad (9)$$

Equivalently

$$\mathbf{Q}^{(k)}_{\mathrm{ML}} = \arg \min_{\mathbf{Q}^{(k)} \in \mathcal{Q}^{(k)}} - \log P(\mathbf{w}|k, \mathbf{Q}^{(k)}). \quad (10)$$

The log-likelihood in (8) is a concave function of variables $\mathbf{Q}^{(k)}_{j \lfloor i \rfloor_k}$ which also satisfy the constraints: $\sum_{j=1}^{|\mathcal{A}|} \mathbf{Q}^{(k)}_{j \lfloor i \rfloor_k} = 1$ for $\lfloor i \rfloor_k = 1, \ldots, k$. Constrained optimization using Lagrange multipliers gives the $(j, \lfloor i \rfloor_k)^{th}$ element of the matrix $\mathbf{Q}^{(k)}_{\mathrm{ML}}$ as

$$\left[\mathbf{Q}^{(k)}_{\mathrm{ML}} \right]_{j, \lfloor i \rfloor_k} = \begin{cases} \dfrac{\sum_{m=1}^{M+1} \mathbf{w}_{j \lfloor i \rfloor_k^m}}{M+1}, & \lfloor i \rfloor_k = 1, \ldots, N - Mk \\ \dfrac{\sum_{m=1}^{M} \mathbf{w}_{j \lfloor i \rfloor_k^m}}{M}, & \lfloor i \rfloor_k = N - Mk + 1, \ldots, k \end{cases} \quad (11)$$

for $j = 1, \ldots, |\mathcal{A}|$. The MLE for the probability mass functions of the random sources is intuitive. Given the period is k, it first amounts to segmentation of the data sequence into k nonoverlapping subsequences. For instance, if the hypothesized statistical period in a gene sequence is 3 then the second subsequence comprises of every third symbol starting with the second symbol. Then the pmf of the m^{th} information source is given by the empirical probabilities of each symbol in the m^{th} subsequence.

The estimates of $\mathbf{Q}^{(k)}$ determine the MLE for the period

$$K_{\mathrm{ML}} = \arg \min_{k \in \mathcal{K}} - \log P(\mathbf{w}|k, \mathbf{Q}^{(k)}_{\mathrm{ML}}). \quad (12)$$

This is a simple plug-in estimator where the search is over a collection of models with complexity of models increasing with k. Within each model, indexed by k, the best fit for the sequence is picked—this is $\mathbf{Q}^{(k)}_{\mathrm{ML}}$. This set of MLEs from different models are then compared for the goodness-of-fit in terms of the likelihood of observing the sequence.

B. Regularized Maximum Likelihood Estimator

The plug-in-estimator for the latent period in its current form (12) may overfit the data since it compares models with increasing complexity and more complex models tend to fit the data better. The model selection is regularized using the minimum description length (MDL) principle. The fundamental idea behind MDL is that more regular the data is, the easier it is to compress and thus learn [24]. For instance in a homological sequence, a single period captures the entire data whereas a sequence of coin-tosses is uniformly random and there may not be any shorter description of the data than the data itself. Most real data lies somewhere in between—it is not completely regular but it is not completely random either. The MDL principle embodies several desired features. Most importantly, MDL avoids overfitting automatically by trading off complexity with the goodness of fit. Given the data and a collection of hypothesis $\mathcal{Q}$, the MDL principle picks the model that compresses the data most with respect to the description method.

As in previous section, $\mathbf{s}$ denotes the sequence data and $\mathcal{Q}^{(1)}, \mathcal{Q}^{(2)}, \ldots, \mathcal{Q}^{(N_0)}$ is a list of candidate models. Recall that $\mathcal{Q}^{(k)}$ is the set of $|\mathcal{A}| \times k$ column-stochastic matrices. The best estimate of the cyclostationary period of sequence $\mathbf{s}$ is the $k \in \mathcal{K}$ that minimizes the description length

$$\mathbb{L}(\mathbf{s}; k) = \mathbb{L}(\mathcal{Q}^{(k)}) + \mathbb{L}(\mathbf{s}|\mathbf{Q}^{(k)}_{\mathrm{ML}}) \quad (13)$$

where $\mathbb{L}(\mathcal{Q}^{(k)})$ is the description length (in bits) of the hypothesis $\mathcal{Q}^{(k)}$ and $\mathbb{L}(\mathbf{s}|\mathbf{Q}^{(k)}_{\mathrm{ML}})$ is the length (in bits) of the description of the data when encoded by the best ML hypothesis $\mathbf{Q}^{k}_{\mathrm{ML}} \in \mathcal{Q}^{(k)}$. The term $\mathbb{L}(\mathcal{Q}^{(k)})$ is the *parametric complexity* of the model and $\mathbb{L}(\mathbf{s}|\mathbf{Q}^{(k)}_{\mathrm{ML}})$ is the *stochastic complexity* of the sequence given the model. The MDL model selection involves a trade-off between the goodness-of-fit and the complexity.

The second term $\mathbb{L}(\mathbf{s}|\mathbf{Q}^{(k)}_{\mathrm{ML}})$ in (13) is the codelength of the data when encoded with the hypothesis $\mathbf{Q}^{(k)}_{\mathrm{ML}}$. Assuming the hypotheses are probabilistic, the Shannon–Fano codes are optimal in terms of the expected codelength. Thus, $\mathbb{L}(\mathbf{s}|\mathbf{Q}^{(k)}_{\mathrm{ML}}) = -\log P(\mathbf{s}|\mathbf{Q}^{(k)}_{\mathrm{ML}})$, where $P(\mathbf{s}|\mathbf{Q}^{(k)}_{\mathrm{ML}})$ is the probability of observing $\mathbf{s}$ conditioned on the hypothesis $\mathbf{Q}^{(k)}_{\mathrm{ML}}$. The codelength is therefore the negative-log-likelihood of having observed the data $\mathbf{s}$. This term is exactly the same as in (12).

The following code may be adopted for the description of the hypothesis. First encode k using $\lceil \log k \rceil$ 1's followed by a 0 which is followed by another $\lceil \log k \rceil$ bits for binary representation of k. This a prefix code that requires $2\lceil \log k \rceil + 1$ bits. The parameters of $\mathbf{Q} \in \mathcal{Q}^{(k)}$ are described by $k' = k|\mathcal{A}|$ frequencies or probabilities that are determined by the counts in the set $\{0, 1, \ldots, \lceil N/k \rceil\}$, thus taking $k' \log(\lceil N/k \rceil + 1)$ bits. The total codelength for the code is therefore

$$\mathbb{L}(\mathcal{Q}^{(k)}) + \mathbb{L}(\mathbf{s}|\mathbf{Q}^{(k)}_{\mathrm{ML}}) = 2\lceil \log k \rceil + 1 + k|\mathcal{A}| \log \left\lceil \frac{N}{k} \right\rceil - \log P(\mathbf{s}|\mathbf{Q}^{(k)}_{\mathrm{ML}}) \quad (14)$$

for $\mathbf{Q}^{(k)}_{\mathrm{ML}} \in \mathcal{Q}^{(k)}$. Summarizing, the MDL estimator is given as

$$K_{\mathrm{MDL}} = \arg \min_{k \in \mathcal{K}} \left(2\lceil \log k \rceil + k|\mathcal{A}| \log \left\lceil \frac{N}{k} \right\rceil - \log P(\mathbf{s}|\mathbf{Q}^{(k)}_{\mathrm{ML}}) \right). \quad (15)$$

It is clear from (15) that the MDL principle yields a penalized ML estimate. The code used here is a *universal code* and implies

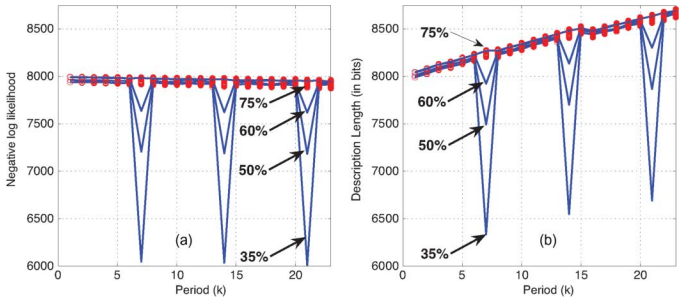


Fig. 2. (a) Negative log-likelihood for the ML estimate plotted against period for a simulated symbolic sequence of length 4000, with period 7 under 35%, 50%, 60% and 75% erosion, (b) description length (in bits) plotted for the ML estimate in $\mathcal{Q}^{(k)}$ plotted against k for corresponding sequences. The CNC permutations are plotted in red.

a universal prior on the hypothesis. In other words, the penalized ML estimator in (15) is essentially the maximum-a-posteriori estimator with respect to the universal prior.

III. Experimental Results

This section presents several examples of cyclostationary structures within symbolic DNA sequences. Section III-A applies the methods of Section II to both simulated and real gene sequences. The methods are extended to consider changing periodicities in symbolic DNA sequences using a windowed approach in Section III-B.

A. Finding Periodicities in DNA Sequences

A homological symbolic sequence from the set $\mathcal{A} = \{\mathrm{A}, \mathrm{G}, \mathrm{C}, \mathrm{T}\}$ with period $K = 7$ was generated and the algorithm was tested with various degrees of erosion introduced by flipping the symbols at randomly chosen points in the sequence. The negative log-likelihood is plotted against the period in Fig. 2(a). The periodic behavior is evident from the plots. Also notable are the subharmonics, i.e. the integer multiples of the true period. The plots strongly support a statistical periodicity of 7 even with 60% erosion. The noise level in the plots increases with erosion and at 75% erosion the sequence exhibits no repetitive behavior. The erosion levels denote the fraction of total symbols that have mutated. The dotted red plot is obtained by a variant of computational negative controls (CNC) strategy proposed in [25]. It corresponds to negative log-likelihood for various permutations of the original sequence. It provides a good reference for comparison since random permutations destroy regular sequential structure. The CNC variant for fifty different permutations is plotted for all the experiments in this paper. Only the features that fall below the family of these curves (when seeking a minima) are deemed significant.

The algorithm was also tested with the protein coding region of chromosome III of $\mathcal{S}.\ cerevisiae$ [26]. The 1629 base-pair (bp) long sequence (bp: 6,571–8,199) shows a latent periodicity of period three in Fig. 3(a). The period-3 behavior of protein coding genes is expected since amino acids are coded by trinucleotide units called *codons* [9], [27]. For comparison with Fourier based methods, the symbolic sequence is transformed into a numerical sequence using the complex mapping developed in [9] for identification of protein coding regions (A $= 0.1 + 0.12j$, G $= 0.45 - 0.19j$, C $= 0$, T $= -0.3 - 0.2j$). The magnitude of the 1629-point DFT of numerical sequence of poly-nucleotides is plotted against the frequency in Fig. 3(b). The peaks at $f_1 = 0.33$ and $f_2 = 0.167$ correspond to 3– and 6-periodic behavior, respectively; however, some other peaks are simply the artifacts, perhaps of the numerical mapping.

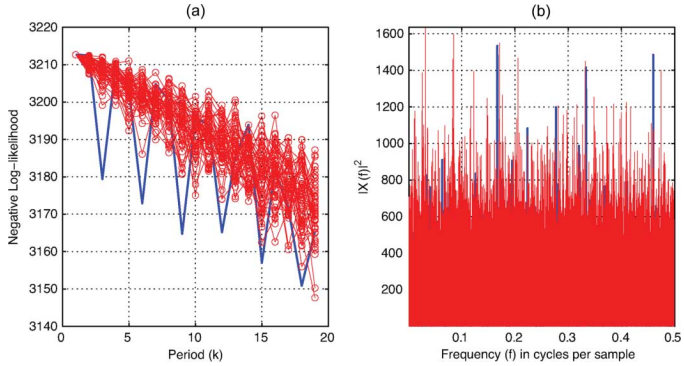


Fig. 3. (a) Negative log-likelihood for ML estimate plotted against period for the 1629 base-pair long sequence from the protein-coding region of chromosome III (bp: 6,571–8,199) of $\mathcal{S}.\ cerevisiae$ genome and (b) the magnitude of DFT of numerical sequence derived from the same sequence. The CNC variants are plotted in red.

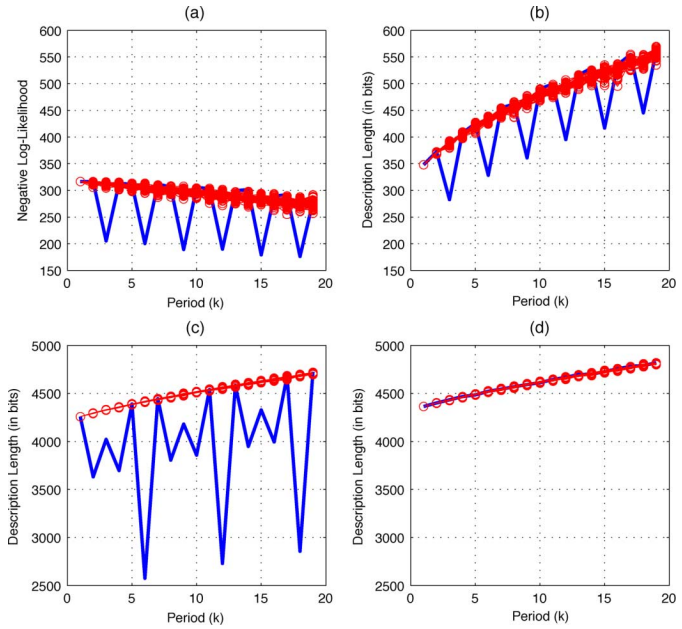


Fig. 4. (a) Negative log-likelihood for ML estimate plotted against the period for the protein coding region of chromosome XVI (bp: 521,009–521,199) of $\mathcal{S}.\ cerevisiae$ genome. Description length (in bits) for the penalized ML estimate in $\mathcal{Q}^{(k)}$ plotted against k for (b) the protein coding region of chromosome XVI (bp: 521,009–521,199) of $\mathcal{S}.\ cerevisiae$ genome, (c) a simulated symbolic sequence of length 2160 with latent period 6, (d) a uniformly random symbolic sequence. The CNC variants are plotted in red circles.

The MLE is compared with the MDL estimator in Fig. 2 for simulated sequences and in Fig. 4(a)–(b) for 191 base-pair-long sequence from Chromosome XVI (bp: 521,009–521,199) of the $\mathcal{S}.$ cerevisiae genome [26]. The problem of *overfitting* is evident from the negative tilt of "valleys" in the plots. This behavior is manifested by (12), giving the largest integer multiple of $K \in \mathcal{K}$. However, the MDL estimator resolves the issue by

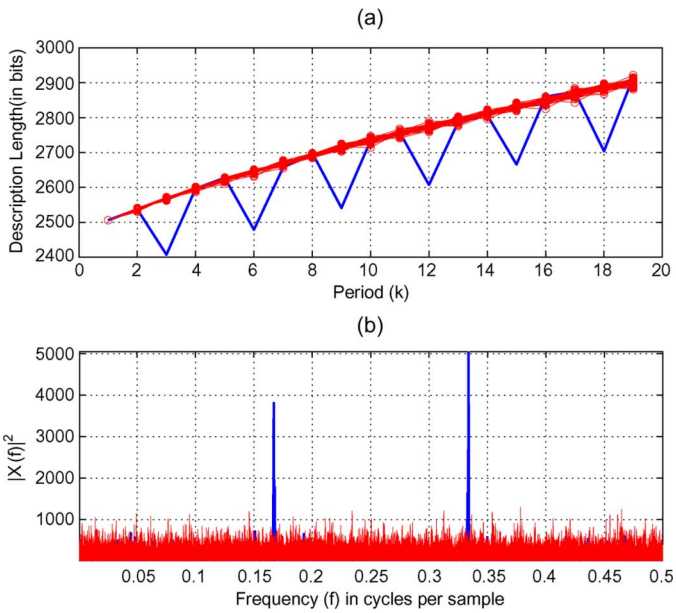


Fig. 5. (a) Description length (in bits) for the ML estimate in $\mathcal{Q}^{(k)}$ plotted against k for the 1305 base pair long protein coding region of chromosome 20 of human genome (bp: 22 557 488–22 558 792); (b) The magnitude of DFT of numerical sequence derived from the protein coding region of chromosome 20 of human genome. The CNC variants are plotted in red.

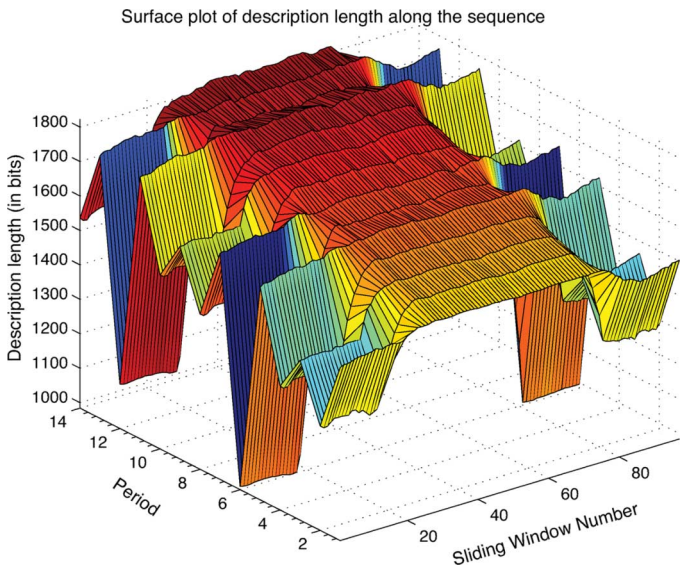


Fig. 6. Description length (in bits) for the ML estimate in $\mathcal{Q}^{(k)}$ plotted against period k along the simulated sequence with latent period given by (1). The surface plot exhibits varying cyclostationarities in the underlying probabilistic source.

penalizing the models commensurately with their complexity. If two models fit data equally well, it picks the simpler one.

Fig. 4(c) shows results with simulated symbolic sequence exhibiting the latent period given by (1). The plot reveals a strong six-periodic behavior and the detected dominant period (the minimum of the curve) coincides with the true latent period. In contrast, when a random sequence is used (i.e., when each source generates all symbols with equal frequency), Fig. 4(d) shows that no significant periodicities are detected.

Although the method of Anastassiou [9] and other numerical representation techniques combined with Fourier transform perform poorly at severe mutation rates [see Fig. 3(b)], their performance in low noise conditions is comparable to the MDL estimator. Fig. 5 shows results for 1305 base-pair-long sequence from Chromosome 20 (bp:22,557,488–22,558,792) of the human genome [26]. The gradual roll-off of valleys in the description length and low noise floor in the DFT plots provide the evidence of high signal to noise ratio. Nonetheless, it should be remarked that the numerical mappings are typically obtained by solving an optimization problem aimed at enhancing particular aspect of the behavior of the sequences, the three-periodic nature for instance. Consequently, such tailored techniques run a risk for being too specific and perform poorly at finding new periodicities.

B. Localizing Periodic Regions in DNA Sequences

The cyclostationarity profile of DNA sequences varies with location. The varying periodicities in DNA can be discovered by using a sliding window and a statistical test may be devised to detect the change in periodicity profiles. The penalized MLE is applied to various simulated symbolic sequences and real gene sequences. In order to detect changes in periodicity profile in a sequence of N symbols, the estimates are computed in a sliding window of size $N_w < N$ with an overlap of N_c symbols between successive windows. The method presented here is similar to windowed Fourier transform techniques for generating the spectrogram in [16], [28], [29], except that no numerical mapping is imposed.

Fig. 6 shows results for a simulated 8000-symbols long DNA sequence that has latent periodicity of period 6 for subsequences with indices 1-2000 and 6001-8000 and is uniformly random in the middle. Thus there are two *change points* in the sequence. The dominant period of the cyclostationary part of the sequence is [(A/C)(T/G)(T/A)(G/T)(C/G/A)(G/A)]. The window size was chosen to be 750 symbols and the overlap was 675 symbols. The description length (Z-axis) is plotted for the ML hypothesis corresponding to each period (Y-axis) along the sequence (X-axis). Note that both change points are detected in the surface plot. Also the six-periodic behavior is evident from the plot as are the subharmonics.

The sliding window method was applied to chromosome 20 of the human genome [26]. The 9748 base-pair long sequence (bp 22,553,000–22,562,747) contains 1305 long (bp 22,557,488–22 558,792) protein coding region (*exon*) flanked by noncoding parts (*introns*) on both sides. The contour plot in Fig. 7 shows a latent periodicity of period three beginning at sliding window number 60 which corresponds to bp 22,557,427 ($N_w = 750$, $N_c = 675$).

The window size N_w determines the usual trade-off between the resolution and the accuracy of the estimates. The larger the window size, the better the estimates since the averaging in the empirical estimator is taken over more data. On the other hand, smaller windows give better resolution since the estimates along the sequence depend only on the symbols in a small neighborhood.

The periodicity profile transforms shape near the change points while in other regions the profile remains constant except for some small fluctuations due to the noise in data.

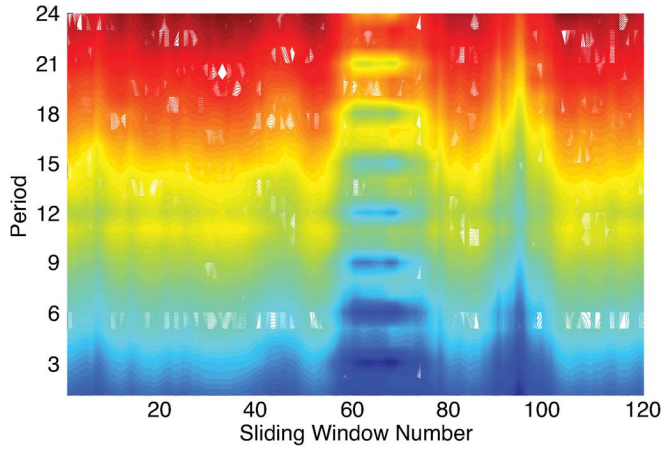


Fig. 7. Contour plot of description length (in bits) for the ML estimate in $\mathcal{Q}^{(k)}$ plotted against period k along the 9748 base-pair long sequence from chromosome 20 of human genome. The sequence contains a short *exon* identified with dark spectral lines in the *spectrogram*.

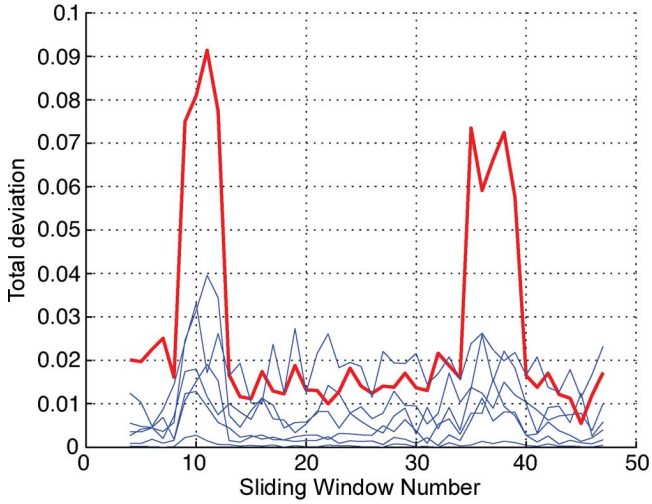


Fig. 8. $\mathbf{\Theta}_t^{(K)}$ plotted for the sequence from Fig. 6. $\mathbf{\Theta}_t^{(6)}$ is plotted in red ($N_w = 750$, $N_c = 600$, $T = 3$).

Thus a powerful statistical test may be constructed based on the positive inflection rate over multiple successive windows. If the maximum likelihood period reported is K then the alternate composite hypothesis is that the period is no longer K. The formulation is similar to the change-point problem in statistics. The test proposed here is based on a cumulative sum approach. The null hypothesis that there is no change is rejected if

$$\mathbf{\Theta}_t^{(K)} = \min_{m\in\{1,\ldots,T\}} \left|\mathbf{Q}_{\mathrm{ML},t}^{(K)} - \mathbf{Q}_{\mathrm{ML},t-m}^{(K)}\right|_{\mathrm{tot}} > \delta_{\mathrm{Th}} \qquad (16)$$

where $|\mathbf{A}-\mathbf{B}|_{\mathrm{tot}} = \sum_{i,j}(a_{ij}-b_{ij})^2$ is the total deviation between matrices $\mathbf{A}$ and $\mathbf{B}$, δ_{Th} is a threshold and T is the number of successive windows over which the test is conducted. The test statistic $\mathbf{\Theta}_t^{(K)}$ for period K is the minimum total deviation between ML estimates for the pmfs in window t and previous T windows. $\mathbf{\Theta}_t^{(K)}$ is plotted in Fig. 8 for the simulated latent periodic sequence used in Fig. 6. The jump in $\mathbf{\Theta}_t^{(6)}$ at $t = 9$ corresponds to the change-point at bp number $N_w + 8\times(N_w - N_c) = 1950$, giving better resolution. The resolution can be further improved upon by decreasing N_c, keeping N_w constant. Note that $\mathbf{\Theta}_t^{(6)}$ is consistently large over transition regions with lobe-width equal to N_w.

IV. Multiple Periodicities

Multiple latent periodicities in symbolic sequences provide evidence of mutations and can help reconstruct the evolution history. In numerical sequences, if multiple periodicities result from addition (composition) of several sequences with different periods, then Periodicity Transforms [30] allow decompositions into likely constituent components. To develop a similar decomposition for symbolic sequences, the evolution and composition mechanisms need to be understood. This section provides a mathematical framework to investigate multiple periodicities in symbolic sequences. The mathematical structure of the periodic subspaces is studied first, and the resulting algebraic properties allow a decomposition of multiple periodicities.

A. Periodic Subspaces

Given a finite alphabet $\mathcal{A}$ of size L, let $\mathcal{P}_p$ be the collection of cyclostationary sources on $\mathcal{A}$ with period p. Then $\mathcal{P} = \bigcup_{p>0}\mathcal{P}_p$ is the set of all cyclostationary sources on $\mathcal{A}$ where p ranges over all positive integers. The set $\mathcal{P}_p$ is identified with the set of $L\times p$ column stochastic matrices. An element $S\in\mathcal{P}_p$ is a cyclostationary source described by an $L\times p$ column-stochastic matrix $\mathbf{Q}^S$ the i^{th} column of which, denoted $\mathbf{q}_i^S$, gives the pmf of S_{np+i} for all $n\in\mathbb{Z}_{\geq 0}$, i.e.

$$P(S_{np+i}=a_j) = P(S_i=a_j) = \mathbf{Q}_{ji}^S \equiv \mathbf{q}_i^S(j) \qquad (17)$$

where $j = 1,\cdots,L$. The following law of composition on the probabilistic sources follows the two carousel model of Fig. 1 in analogy with the DNA replication process. Define

$$\begin{aligned}\oplus : \mathcal{P}\times\mathcal{P} &\to \mathcal{P}\\ (X,Y) &\mapsto Z\end{aligned} \qquad (18)$$

on $\mathcal{P}$ as follows. Let $X,Y\in\mathcal{P}$ be sources with periodicities p and q respectively. Then $Z = X\oplus Y$ is the probabilistic source such that for all $a\in\mathcal{A}$

$$P(Z_n=a) = P\left(X_{\lfloor n\rfloor_p}=a, Y_{\lfloor n\rfloor_q}=a | X_{\lfloor n\rfloor_p} = Y_{\lfloor n\rfloor_q}\right). \qquad (19)$$

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 $\lfloor m\rfloor_p = \lfloor n\rfloor_p$ and $\lfloor m\rfloor_q = \lfloor n\rfloor_q$. Thus for all $a\in\mathcal{A}$, $P(Z_m=a) = P(X_{\lfloor n\rfloor_p}=a, Y_{\lfloor n\rfloor_q}=a | X_{\lfloor n\rfloor_p}=Y_{\lfloor n\rfloor_q}) = P(Z_n=a)$. ■

Corollary 1: Let $X,Y\in\mathcal{P}_p$. Then $X\oplus Y$ is p-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 (19), the n^{th} column of the $L\times pq$ matrix $\mathbf{Q}^Z$ is

$$\mathbf{q}_n^Z = \frac{1}{C}\begin{bmatrix}\mathbf{q}_{\lfloor n\rfloor_p}^X(1)\mathbf{q}_{\lfloor n\rfloor_q}^Y(1)\\ \vdots \\ \mathbf{q}_{\lfloor n\rfloor_p}^X(L)\mathbf{q}_{\lfloor n\rfloor_q}^Y(L)\end{bmatrix} \qquad (20)$$

where $C = \sum_{j=1}^{L}\mathbf{q}_{\lfloor n\rfloor_p}^X(j)\mathbf{q}_{\lfloor n\rfloor_q}^Y(j)$ is the normalization factor.

Example 1: Consider an example of composition of two cyclostationary sources with statistical periods 2 and 3. Equation (20) gives

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

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

The first source in the sequence X acts like the identity and the last source of the sequence Y acts like an infinity of the binary operation. The dominant periods of X and Y are $D_X^* = [N\ A]$ and $D_Y^* = [T\ T\ A]$ respectively, where N denotes (A/G/C/T).■

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

$$\mathbf{q}_n^Z(k) = \left(\mathbf{q}_n^X(k)\right)^2 / \sum_{j=1}^{L} \left(\mathbf{q}_n^X(j)\right)^2 \tag{21}$$

for $k = 1, \ldots, L$ and $n = 1, \ldots, p$. The operation of composing a probabilistic source 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

$$\begin{aligned} \circ : \mathbb{R} \times \mathcal{P} &\to \mathcal{P} \\ (r, X) &\mapsto Z \end{aligned} \tag{22}$$

so that $Z = r \circ X$ is the information source with

$$P(Z_n = a) = \frac{P(X_n = a)^r}{\sum_{b\in\mathcal{A}} P(X_n = b)^r} \tag{23}$$

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 source with X_i distributed as $\mathbf{q}_i^X = [1/2\ \ 1/4\ \ 1/4\ \ 0]'$. If $Y = 2 \circ X$ then Y_i is distributed as $\mathbf{q}_i^Y = [2/3\ \ 1/6\ \ 1/6\ \ 0]'$.

We now state the first of our main results of the section which follows simply from the definitions of binary composition and scalar multiplication.

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

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 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 shown at the bottom of the page, for $j = 1, \ldots, L$ and $i = 1, \ldots, pqr$. The unique identity element, denoted E, is the stationary or 1-periodic random sequence such that $P(E = a_j) = 1/L$ 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. ■

It is a consequence of the theorem above that the collection of cyclostationary sources is closed under the binary law defined in (18). The periodic structure of a cyclostationary source is thus preserved under composition and the resulting probabilistic source exhibits cyclostationarities of the components, which can be identified from the periodicity analysis. Combined with the scalar multiplication, a richer structure is found on the periodic subspaces.

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 $\alpha \in \mathbb{R}$, $X \in \mathcal{P}_p, Y \in \mathcal{P}_q$, let $V = \alpha \circ (X \oplus Y)$ and $W = (\alpha \circ X) \oplus (\alpha \circ Y)$. Then $(j, i)^{th}$ entry of $\mathbf{Q}^V$ is given as

$$\mathbf{Q}_{ji}^V = \frac{\left(\mathbf{Q}_{j\lfloor i\rfloor_{pq}}^{X\oplus Y}\right)^\alpha}{\sum_{j'}\left(\mathbf{Q}_{j'\lfloor i\rfloor_{pq}}^{X\oplus Y}\right)^\alpha} = \frac{\left(\mathbf{Q}_{j\lfloor i\rfloor_p}^X \mathbf{Q}_{j\lfloor i\rfloor_q}^Y\right)^\alpha}{\sum_{j'}\left(\mathbf{Q}_{j'\lfloor i\rfloor_p}^X \mathbf{Q}_{j'\lfloor i\rfloor_q}^X\right)^\alpha} = \mathbf{Q}_{ji}^W.$$

Similarly, it is easy to check that scalar multiplication distributes over scalar addition: 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: let $V = \alpha \circ (\beta \circ X), W = (\alpha\beta) \circ X$. Then

$$\mathbf{Q}_{ji}^V = \frac{\left(\mathbf{Q}_{ji}^{\beta\circ X}\right)^\alpha}{\sum_{j'}\left(\mathbf{Q}_{j'i}^{\beta\circ X}\right)^\alpha} = \frac{\left(\left(\mathbf{Q}_{ji}^X\right)^\beta\right)^\alpha}{\sum_{j'}\left(\left(\mathbf{Q}_{j'i}^X\right)^\beta\right)^\alpha} = \mathbf{Q}_{ji}^W.$$

■

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

The significance of Theorem 2 is that it allows for varying degrees of constituent periodicities. A symbolic source may exhibit a much stronger p-period than q-period. In such cases the scalar multiplier captures the relative weight of each component. The periodic subspaces are also closed under scalar multiplication and hence behave much like real-valued signal spaces.

B. Decomposing Multiple Periodicities

This section investigates the problem of decomposing the discovered probabilistic source that exhibits multiple periodicities into various smaller components. Multiple latent periodicities have been observed in various DNA sequences. The high-sulphur wool matrix protein B2A from sheep (SHPWMPBB at NCBI [31]) exhibits multiple latent periodicities

$$\frac{\mathbf{Q}_{j\lfloor i\rfloor_p}^X \left(\mathbf{Q}_{j\lfloor i\rfloor_q}^Y \mathbf{Q}_{j\lfloor i\rfloor_r}^Z\right)}{\sum_{j'} \mathbf{Q}_{j'\lfloor i\rfloor_p}^X \left(\mathbf{Q}_{j'\lfloor i\rfloor_q}^Y \mathbf{Q}_{j'\lfloor i\rfloor_r}^Z\right)} = \frac{\left(\mathbf{Q}_{j\lfloor i\rfloor_p}^X \mathbf{Q}_{j\lfloor i\rfloor_q}^Y\right) \mathbf{Q}_{j\lfloor i\rfloor_r}^Z}{\sum_{j'} \left(\mathbf{Q}_{j'\lfloor i\rfloor_p}^X \mathbf{Q}_{j'\lfloor i\rfloor_q}^Y\right) \mathbf{Q}_{j'\lfloor i\rfloor_r}^Z} = \mathbf{Q}_{ji}^W$$

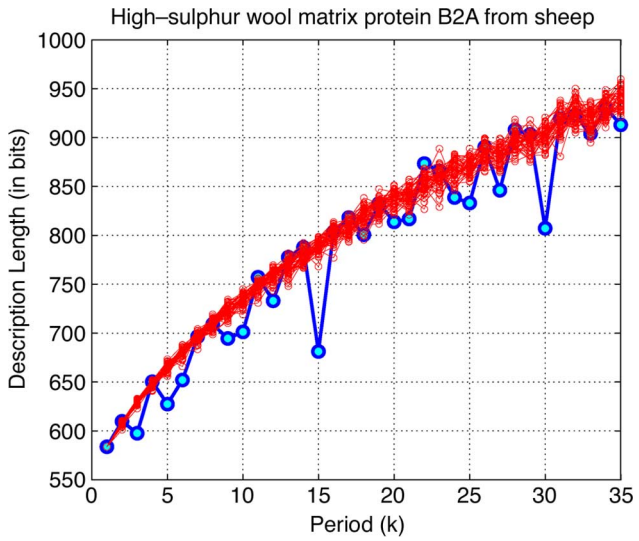


Fig. 9. Description length (in bits) plotted against the period for high-sulphur wool matrix protein B2A from sheep (bp:273–561). The DNA sequence exhibits multiple latent periodicities with period 3 and 5.

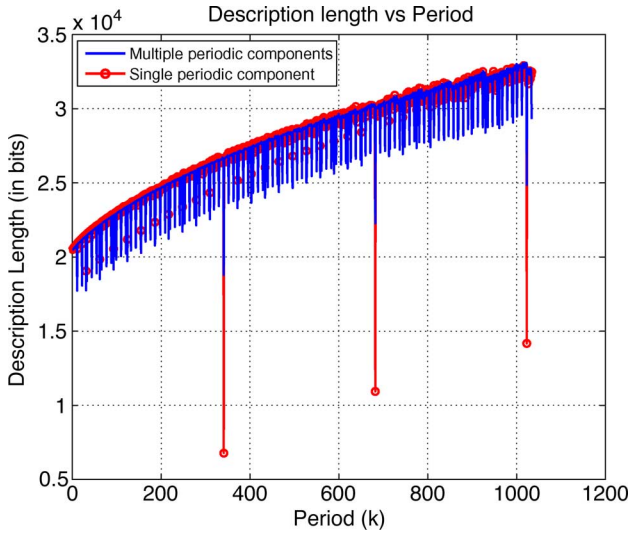


Fig. 10. Description length (in bits) plotted against the period for two cyclostationary sequences both with period 341. The dotted plot (in red) corresponds to the sequence comprising of a single cyclostationary source with period 341 while the other is composed of two cyclostationary sources with period 11 and 31.

with period 3 and 5. The description length (in bits) is plotted against the period for the base pairs 273–561 in Fig. 9. The statistical significant periods seen are 3 and 5 as well as the subharmonics 6,9,12 and 10,15,30. The dominant period is found to be [CTGCCGGCCGGCCTG]. Several other instances of multiple periodicities were discovered using the penalized ML estimator. In the T-cell receptor alpha-chain gene of *Fugu rubripes* (Japanese pufferfish, accession no. AF110525 [31]) the latent periodicity with length equal to 59 bases was observed in the protein coding region (bp:13628–14594). In *Deinococcus radiodurans* gene for *c-di-GMP phosphodiesterase* (from sequence AE000513 [31]) latent periodicity equal to 120 bases was observed from base pairs 3108 to 3963 and in *Methylobacterium extorquens* methanol oxidation gene *mxaE* (from sequence AF017434 [31]) latent periodicity equal to 126 bases was observed from base pairs 165–1010. However, it should be remarked that not all sequences with composite latent periods exhibit multiple periodicities. The minimum description length is plotted in Fig. 10 for two sequences with periodicity of 341. One of the sequences exhibits strong 11-periodic and 31-periodic behavior as well, thus admitting an exact decomposition. It is evident from the plot that the other sequence is not composed from smaller sources but generated from a single cyclostationary source with period 341.

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

$$\begin{bmatrix} Z_1 \\ \vdots \\ Z_{pq} \end{bmatrix}_{pq\times 1} = \underbrace{\begin{bmatrix} I_p & I_q \\ \vdots & \vdots \\ I_p & I_q \end{bmatrix}}_{\mathbf{T}_{pq\times(p+q)}} \circ \begin{bmatrix} X_1 \\ \vdots \\ X_p \\ Y_1 \\ \vdots \\ Y_q \end{bmatrix}_{(p+q)\times 1} . \quad (24)$$

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 $[\underbrace{-1 \ldots -1}_{p}\underbrace{1 \ldots 1}_{q}]'$.

Proof: See Appendix. ■

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

$$(X \oplus \delta_p) \oplus (Y \ominus \delta_q) = Z \quad (25)$$

where $Y \ominus \delta_q = Y \oplus (-1 \circ \delta_q)$ and $\delta_r = [\overbrace{\delta, \ldots, \delta}^{r \text{ times}}]$ 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 and Y unique only 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 $\delta = [2/10 \ \ 3/10 \ \ 3/10 \ \ 2/10]'$ and $-1 \circ \delta = [(3/10 \ \ 2/10 \ \ 2/10 \ \ 3/10]'$. The dominant periods of X' and Y' are $D^*_{X'} = [(G/C)A]$ and $D^*_{Y'} = [T \ T \ A]$ respectively. ■

Comparing the dominant periods in examples 1 and 3 shows that there is more than one decomposition, in terms of dominant periods, of the cyclostationary source Z. This is a consequence

of Theorem 3. A decomposition that is biologically correct may be discovered by generating the class of all possible decompositions. Two possible decompositions of the latent period [CTGCCGGCCGGCCTG] for wool matrix protein B2A (SHPWMPBB) were found to be [GGT, CG(G/C)CG] and [GCT, CGTCG]. The latter seems biologically correct since the triplet (GCT) in the coding regions is considered to be the dominating pattern in ancient codons, given the variants GCN, TCT, CCT, ACT, GAT and GGT which code for the amino acids Ala, Ser, Pro, Thr, Asp and Gly respectively (see genetic code [9]), are considered to be the earliest codons [10]. The triplet also results, by the process of transcription, in the pattern $(\mathrm{GCU})_n$ in mRNA which serves for maintaining a correct reading frame during translation by making the in-frame binding energetically favorable [10]. The decomposition above is achieved by a simple algorithm, briefly outlined next.

Consider decomposition of an r-periodic probabilistic source Z into p and q-periodic probabilistic sources X and Y respectively, where $r = pq$ and p, q are coprime. Assume that the minimum description length is attained at period equal to r and the periods p and q are statistically significant (relative to CNC variants). The objective is to determine $\mathbf{Q}^X, \mathbf{Q}^Y, \mathbf{Q}^Z$ such that $Z = X \oplus Y$. A good estimate of $\mathbf{Q}^Z$ is $\mathbf{Q}^{(r)}_{\mathrm{ML}}$ whereas $\mathbf{Q}^{(p)}_{\mathrm{ML}}$ and $\mathbf{Q}^{(q)}_{\mathrm{ML}}$ only provide initial starting points for $\mathbf{Q}^X$ and $\mathbf{Q}^Y$ in an iterative procedure. At each iteration, the probabilistic source that has smaller description length ($\mathbf{Q}^X$ or $\mathbf{Q}^Y$) is kept fixed while the parameters of other are adapted so as to minimize the total deviation between $\mathbf{Q}^X \oplus \mathbf{Q}^Y$ and $\mathbf{Q}^{(r)}_{\mathrm{ML}}$. The process is repeated until the total deviation is within a specified tolerance. The convergence of this adaptive technique can be established by appealing to the topological properties of the periodic subspaces and the continuity of the law of composition.

V. Discussion

Various regions of DNA sequences exhibit characteristic statistical periodicities. Mapping this behavior to structural and functional roles is an important aspect of genomic signal processing. The investigation of multiple periodicities in gene sequences and their decomposition into smaller periodic components may be useful as a way to understand the underlying generative mechanism. 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 here is to model the symbolic sequence as a nonstationary random process on a finite alphabet and to study the (de)composition of the distributions under a composition rule inspired by the biological model for DNA 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 a single realization of data. This sequential 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 [32].

Appendix

Proof of Theorem 3: Without loss of generality assume that $p \leq q$. Then $\mathbf{T}_j$, the j^{th} column of matrix $\mathbf{T}$, is of the form

$$\left[\underbrace{\mathbf{e}'_{p,j} \ldots \mathbf{e}'_{p,j}}_{q \text{ copies}}\right]' \text{ if } j \leq p \text{ and } \left[\underbrace{\mathbf{e}'_{q,j-p} \ldots \mathbf{e}'_{q,j-p}}_{p \text{ copies}}\right]' \text{ if } j > p$$

where, $\mathbf{e}_{p,j}$ is a $p \times 1$ vector such that the j^{th} entry is one and rest are zero. Note that

$$\sum_{j=1}^{p} \mathbf{T}_j = \mathbf{1}_{pq} \text{ and } \sum_{j=p+1}^{q} \mathbf{T}_j = \mathbf{1}_{pq}, \tag{26}$$

where $\mathbf{1}_{pq}$ is a $pq \times 1$ vector of all ones. Let $\mathbf{u} = [\underbrace{-1 \ldots -1}_{p} \underbrace{1 \ldots 1}_{q}]'$. Then

$$\mathbf{T}\mathbf{u} = \sum_{j=1}^{p} -\mathbf{T}_j + \sum_{j=p+1}^{q} \mathbf{T}_j = -\mathbf{1}_{pq} + \mathbf{1}_{pq} = \mathbf{0}. \tag{27}$$

Therefore, $\mathbf{T}$ is not full-rank and $\mathbf{u}$ is in the null-space of $\mathbf{T}$. Now we show that any collection of $p+q-1$ columns of $\mathbf{T}$ is linearly independent. Consider the following $pq \times (p+q-1)$ matrix

$$\mathbf{T}' = [\mathbf{T}_1 \ldots \mathbf{T}_{k-1}\, \mathbf{T}_{k+1} \ldots \mathbf{T}_{p+q}] \tag{28}$$

consisting of all but the k^{th} column of $\mathbf{T}$. The j^{th} row of $\mathbf{T}$ has unity at two locations: $\lfloor j \rfloor_p$ and $p + \lfloor j \rfloor_q$. Define

$$J = \{j \in \{1, \ldots, pq\} | \lfloor j \rfloor_p = k \text{ or } p + \lfloor j \rfloor_q = k\}.$$

The first condition fails if $k > p$ and second fails otherwise. Without loss of generality assume that $k \leq p$. Then $J = \{k, k+p, \ldots, k+(q-1)p\} = \{k+mp | m = 0, \ldots, q-1\}$. For any $i \in J$, the i^{th} row of $\mathbf{T}'$ has a single nonzero entry, $\mathbf{T}^*_{i,p+1+\lfloor i-1 \rfloor_q}$, and for any nonzero vector $\mathbf{v} = [\mathbf{v}_1 \ldots \mathbf{v}_{k-1} \mathbf{v}_{k+1} \ldots \mathbf{v}_{p+q}]$ in $\mathbb{R}^{p+q-1}$,

$$[\mathbf{T}'\mathbf{v}]_j = \begin{cases} \mathbf{v}_{p+1+\lfloor j-1 \rfloor_q}, & j \in J \\ \mathbf{v}_{1+\lfloor j-1 \rfloor_p} + \mathbf{v}_{p+1+\lfloor j-1 \rfloor_q}, & j \in \{1, \ldots, pq\} \setminus J. \end{cases}$$

Let $j_1, j_2 \in J$ such that $j_1 \neq j_2; j_1 = k+mp$ and $j_2 = k+np$ for some $n \neq m$. Then $\lfloor j_1 - 1 \rfloor_q = \lfloor j_2 - 1 \rfloor_q$ if and only if q divides $j_1 - j_2$, i.e., q divides $(n-m)p$. But p and q are co-prime and therefore all $j \in J$ are distinct so that $\{\lfloor j - 1 \rfloor_q : j \in J\} = \{0, 1, \ldots, q-1\}$. Thus $\{[\mathbf{T}'\mathbf{v}]_j : j \in J\} = \{\mathbf{v}_{p+1+\lfloor j-1 \rfloor_q} : j \in J\} = \{\mathbf{v}_{p+1}, \ldots, \mathbf{v}_{p+q}\}$. And $\mathbf{T}'\mathbf{v} = 0$ if and only if $\mathbf{v}_{p+1} = \ldots = \mathbf{v}_{p+q} = 0$ which implies $\{[\mathbf{T}'\mathbf{v}]_j : j \in J^c\} = \{\mathbf{v}_{1+\lfloor j-1 \rfloor_p : j \in J^c}\} = \{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$. Again $\mathbf{T}'\mathbf{v} = 0$ implies $\mathbf{v}_1 = \ldots = \mathbf{v}_p = 0$. This contradicts that $\mathbf{v}$ is nonzero. Therefore the columns of $\mathbf{T}'$ are linearly independent and $\mathbf{T}$ has rank $p+q-1$. The null space of $\mathbf{T}$ is one-dimensional and spanned by $\mathbf{u}$. ■

REFERENCES

[1] W. Wang and D. H. Johnson, "Computing linear transforms of symbolic signals," *IEEE Trans. Signal Process.*, vol. 50, no. 3, pp. 628–634, Mar. 2002.

[2] E. V. Korotkov and N. Kudryaschov, "Latent periodicity of many genes," *Genome Inform.*, 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, Mar. 1993.

[4] C. M. Hearne, S. Ghosh, and J. A. Todd, "Microsatellites for linkage analysis of genetic traits," *Trends Genetics*, vol. 8, p. 288, 1992.

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

[6] 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 Res.*, vol. 6, no. 3, pp. 153–163, 1999.

[7] E. V. Korotkov and D. A. Phoenix, "Latent periodicity of DNA sequences of many genes," in *Proc. Pacific Symp. Biocomputing*, 1997, pp. 222–229.

[8] E. V. Korotkov and M. A. Korotova, "Latent periodicity of DNA sequences of some human genes," *DNA Seq.*, vol. 5, p. 353, 1995.

[9] D. Anastassiou, "Genomic signal processing," *IEEE Signal Process. Mag.*, vol. 18, no. 4, pp. 8–20, Jul. 2001.

[10] E. N. Trifonov, "3-, 10.5-, 200- and 400-base periodicities in genome sequences," in *Phys. A: Statist. Theoretical Physics*. New York: Elsevier, 1998, vol. 249, pp. 511–516.

[11] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharaya, and R. Ramaswamy, "Prediction of probable genes by Fourier analysis of genomic sequences," *Bioinformatics*, vol. 13, no. 3, pp. 263–270, 1997.

[12] V. R. Chechetkin, L. A. Knizhnikova, and A. Y. Turygin, "Three-quasiperiodicity, mutual correlations, ordering and long-range modulations in genomic nucleotide sequences for viruses," *J. Biomolec. Struct. Dynam.*, vol. 12, pp. 271–299, 1994.

[13] E. A. Cheever, D. B. Searls, W. Karunaratne, and G. C. Overton, "Using signal processing techniques for DNA sequence comparison," in *Proc. 1989 Fifteenth Annu. Northeast Bioengineering Conf.*, Mar. 1989, pp. 173–174.

[14] R. Gupta, D. Sarthi, A. Mittal, and K. Singh, "Exactly periodic subspace decomposition based approach for identifying tandem repeats in DNA sequences," in *Proc. of the 14th Eur. Signal Processing Conf. (EUSIPCO)*, Florence, Italy, Sep. 2006.

[15] P. D. Cristea, "Genetic signal representation and analysis," in *Proc. SPIE Conf., Int. Biomedical Optics Symp. (BIOS02)*, 2002, pp. 77–84.

[16] M. Buchner and S. Janjarasjitt, "Detection and visualization of tandem repeats in DNA sequences," *IEEE Trans. Signal Process.*, vol. 51, no. 9, pp. 2280–2287, Sep. 2003.

[17] G. L. Rosen, "Signal Processing for Biologically-Inspired Gradient Source Localization and DNA Sequence Analysis," Ph.D. dissertation, Georgia Inst. Technol., Atlanta, 2006.

[18] M. Akhtar, J. Epps, and E. Ambikairajah, "On DNA numerical representations for period-3 based exon prediction," in *GENSIPS (Genomic Signal Processing and Statistics)*, Tuusula, Finland, Jun. 2007.

[19] W. A. Gardner, A. Napolitano, and L. Paura, "Cyclostationarity: Half a century of research," *Signal Process.*, vol. 86, no. 4, pp. 639–697, 2006.

[20] R. H. Burdon, *Genes and the Environment*. Phialdelphia?, PA: Taylor & Francis, 1999.

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

[22] R. Arora and W. A. Sethares, "Decomposing statistical periodicities," in *IEEE Workshop on Statistical Signal Processing (SSP)*, Madison, WI, Aug. 2007.

[23] R. Arora, W. A. Sethares, and J. Bucklew, "Localizing time-varying periodicities in symbolic sequences," in *Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2008.

[24] P. Grunwald, I. J. Myung, and M. Pitt, *Advances in Minimum Description Length: Theory and Applications*. Cambridge, MA: MIT Press, Apr. 2005.

[25] R. Pearson, T. Zylkin, J. Schwaber, and G. Gonye, "Quantitative evaluation of clustering results using computational negative controls," in *Proc. SIAM Int. Conf. Data Mining*, Lake Buena Vista, FL, 2004, pp. 188–199.

[26] UCSC Gene Sorter [Online]. Available: http://genome.ucsc.edu/

[27] B. J. Yoon and P. P. Vaidyanathan, "Computational identification and analysis of noncoding RNAs—Unearthing the buried treasures in the genome," *IEEE Signal Process. Mag.*, vol. 24, no. 1, pp. 64–74, Jan. 2007.

[28] R. Hall and L. Stern, "A rapid method for illustrating features in both coding and non-coding regions of a genome," *Bioinformatics*, vol. 20, no. 6, pp. 982–983, 2004.

[29] E. Santo and N. Dimitrova, "Improvement of spectral analysis as a genomic analysis title," in *GENSIPS*, Tuusula, Finland, June 2007.

[30] W. A. Sethares and T. W. Staley, "Periodicity transforms," *IEEE Trans. Signal Process.*, vol. 47, no. 11, pp. 2953–2964, Nov. 1999.

[31] National center for biotechnology information [Online]. Available: http://www.ncbi.nlm.nih.gov/

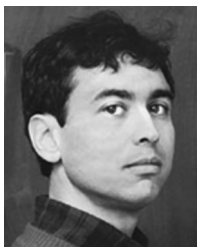
[32] W. A. Gardner, *Cyclostationarity in Communications and Signal Processing*. New York: IEEE Press, 1994.

Raman Arora (S'02) received the B.Engg. degree in electronics and communication from Delhi University, India, in 2001 and the M.S. degree in electrical and computer engineering from University of Wisconsin, Madison in 2005. He is currently pursuing the Ph.D. degree at the University of Wisconsin, Madison.

From 2001 to 2003, he was with the DSP Group of Hughes Software Systems. His research interests include spectral methods and algebraic techniques in signal processing with applications to acoustics, bioinformatics and vision.

William A. Sethares (M'87) received the B.A. degree in mathematics from Brandeis University, Waltham, MA and the M.S. and Ph.D. degrees in electrical engineering from Cornell University, Ithaca, NY.

He has worked at the Raytheon Company as a Systems Engineer and is currently Professor in the Department of Electrical and Computer Engineering at the University of Wisconsin, Madison. He has held visiting positions at the Australian National University, at CCMIX, Paris, France, at the Technical Institute in Gdansk, Poland, and at the NASA Ames Research Center, Mountain View, CA. His research interests include adaptation and learning in signal processing, communications, and acoustics. He is the author of *Tuning, Timbre, Spectrum, Scale* (Springer, now in its second edition), coauthor of *Telecommunication Breakdown: Concepts of Communications Transmitted via Software Radio* (Prentice-Hall, 2004) and has recently completed *Rhythm and Transforms* (Springer, 2007).



James A. Bucklew (SM'02) received the Ph.D. degree in electrical engineering from Purdue University, West Lafayette, IN, in 1979.

He is currently a Professor in the Department of Electrical and Computer Engineering and the Department of Mathematics at the University of Wisconsin, Madison. His research interests are in the application of probability and statistics to signal processing and communication problems. He has published over 100 articles in these fields. He is the author of the books *Large Deviation Techniques in Decision, Simulation, and Estimation* (Wiley-Interscience, 1990) and *Introduction to Rare Event Simulation* (Springer-Verlag, 2004).

Dr. Bucklew has served as an Associate Editor (1990–1992) for the IEEE TRANSACTIONS ON INFORMATION THEORY and as Associate Editor (1997–1999) for the IEEE TRANSACTIONS ON SIGNAL PROCESSING.