# Latent Periodicities in Genome Sequences

Raman Arora, William A. Sethares and James A. Bucklew

**Abstract**

A novel approach is presented to the detection of periodicities in DNA sequences. A DNA sequence can be modelled as a nonstationary stochastic process that exhibits various statistical periodicities in 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 sequences 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 the point mutations in gene sequences. This composition is shown to give a rich mathematical structure on the collection of cyclostationary sources and allows a uniqueness theorem for the decomposition of statistically periodic symbolic sources.

**Index Terms**

Symbolic periodicity, symbolic sequences, genomic signal processing, gene replication, cyclostationarity.

## I. INTRODUCTION

SYMBOLIC sequences consist of strings of elements drawn from a finite set, 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, locating hidden periodicities in DNA sequences is important since repetitions in DNA have been shown to be correlated with several structural and functional roles [2]. For example, a base (symbol) periodicity of 21 is associated with $\alpha$-helical formation for synthesized protein molecules [2] and a base periodicity of three

The authors are with the Department of Electrical and Computer Engineering, University of Wisconsin-Madison, Madison, WI 53706-1691 USA. 608-262-5669 ramanarora@wisc.edu, sethares@ece.wisc.edu

is identified with protein coding region 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 evolution 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 are replaced or altered (including 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

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

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 often provide insights into the nature of early version of the sequences. For instance, in mRNA, the latent period (G)(C)(U) is believed to be sequence fossil of ancient codons which dominated the earliest stages of evolution [10]. Of course, this taxonomy of periodicities applies to any symbolic sequence.

Symbolic random variables take values on a set called the *alphabet* and its elements are called *symbols*. Most current approaches to detecting periodicities transform the symbolic sequences into numerical sequences and compute Fourier transform [11], [9], [12], [13] or perform exact 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 (T$=0$, C$=1$, A$=2$, G$=3$) suggested in [15] puts a total order on the set; the complex representation (A$=1+j$, G$=-1+j$, C$=-1-j$, 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 (T=0, C=1, A=2, 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 third 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, some of which were reported in [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], [16], [14].

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) and the sequence is generated by drawing symbols 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 Figure 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 periodicity* or *strict sense cyclostationarity* [19]. The number of sources is equal to the latent period in the sequence. The cyclic model is justified by observing that it captures all three notions of periodicities in symbolic sequences: tandem repeats result in information sources with trivial zero-one pmfs while the eroded and latent periodicities correspond to pmfs that allow for flipping of symbols.

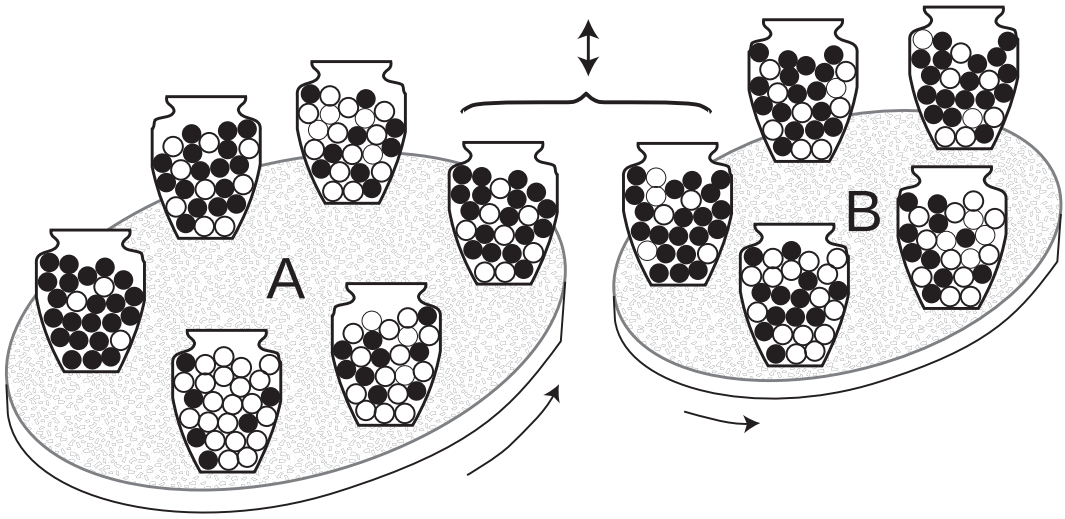


Fig. 1. Each time a ball is removed from one of the $N_A$ urns (indicated by the arrow), platform A rotates, bringing a new urn into position. Similarly, carousel B contains $N_B$ urns, each with its own collection of balls. The urns are the information sources and the cyclostationary sequences generated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

In DNA sequences, multiple periodicities have also been observed [7]. For example, latent periodicities of 120

and 126 base-pairs were reported in various genes in [2]. Such longer periods that are multiples of 3 tend to occur in coding regions. As noted by Korotkov et. al. [7], these periodicities can be related to evolutionary origins via multiple duplications. This paper creates a framework for studying multiple periodicities in symbolic random sequences 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 $\beta$ 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 $\beta$ 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 finite number of sequences. However, the operation is not associative and the order in which the sequences are composed is crucial.

This paper presents a (different) method of composition in analogy with the DNA replication process. The corresponding physical experiment is illustrated in Figure 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 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 and 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 following: the former defines an event as complementary base pairs attached to the two strands of new DNA sequence; 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 the original sequence and the second sequence of complementary nucleotides generated

by the polymerase. The mutations in the latter sequence manifest itself by altering the statistical periodicity profile of the sequence. This method of composition defines a rich mathematical structure (as detailed in Section IV) in which to study statistical periodicities with multiple hidden periodicities. In particular, the binary law is associative. This makes the extension to a finite number of sequences obvious and the order of composition irrelevant.

The paper is organized as follows. The problem of detecting latent periodicities in general symbolic sequences is formulated mathematically in the next section. The maximum likelihood estimate of the dominant period is developed in Section II-A and the estimates are improved by incorporating a complexity term derived from the minimum description length (MDL) principle likelihood function in Section II-B. The model is then applied to both simulated sequences and to DNA sequence data in SectionIII-A. The application of the method developed to finding genes in DNA sequences and building probabilistic representations for non-coding RNAs is presented in section III. Section IV presents the mathematical structures needed to make sense of multiple simultaneous periodicities in symbolic sequences. The corresponding inverse problem, how a cyclostationary symbolic sequence can be decomposed into constituent cyclostationary subsequences, is also addressed. While the DNA sequences provides motivation for this work, the underlying mathematics is general enough to easily include any symbolic set with any (finite) number of elements. Some parts of this paper were previously presented in [21] and [22].

## II. Statistical Periodicity

A given symbolic sequence $D = D_1 D_2 \ldots$ can be denoted by the mapping $D : \mathbb{N} \to \mathcal{X}$, from the natural numbers to an alphabet $\mathcal{X}$. For DNA sequences, $\mathcal{X} = \{\mathrm{A}, \mathrm{G}, \mathrm{C}, \mathrm{T}\}$ where the symbols denote nucleotides Adenine, Guanine, Cytosine and Thymine respectively. Let $P$ denote a probability distribution on $\mathcal{X}$ and let $X$ denote the corresponding random variable (or information source). Let $\mathcal{X}^n$ denote the $n$-fold cartesian product of $\mathcal{X}$ and $x^n \in \mathcal{X}^n$ denote a random 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 $\mathcal{X}^1, \mathcal{X}^2, \ldots$ such that for all $n$, and for all $x^n \in \mathcal{X}^n$, $P^{(n)}(x^n) = \sum_{y \in \mathcal{X}} P^{(n+1)}(x^n, y)$.

If a symbolic sequence $D$ is generated by repeatedly picking subsequences from a probabilistic source $P^{(\mathcal{T})}$ and concatenating, the statistical periodicity of $D$ is $\mathcal{T}$. In other words, the sequence $D$ is generated by $\mathcal{T}$ information sources denoted as $X_1, \ldots, X_{\mathcal{T}}$, in a cyclic fashion. The random variable $X_i$ takes values on the alphabet $\mathcal{X}$ according to an associated probability mass function $P_i$; it generates the $j^{th}$ symbol in $\mathcal{X}$ with probability $P_i(j) =$

$\mathcal{P}(X_i = \mathcal{X}_j)$ for $j = 1, \ldots, |\mathcal{X}|$ where $|\mathcal{X}|$ is the cardinality of the alphabet (which is four for the DNA sequences). The *dominant period* of a $\mathcal{T}$-periodic cyclostationary sequence is defined to be the symbolic sequence $D^* = [D_1^*, \ldots, D_{\mathcal{T}}^*]$ of length $\mathcal{T}$ such that the $k^{th}$ symbol in every period is more likely to be $D_k^*$ than any other symbol from the alphabet. Mathematically, $D_k^* = \arg\max_{j \in \mathcal{X}} P_i(j)$. If $D_k^*$ is not unique then the following notation is adopted: the dominant period [A(G/C)(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.

The number of complete statistical periods in $D$ are $M = \lfloor N/\mathcal{T} \rfloor$, where $\lfloor x \rfloor$ denotes the largest integer that is smaller than or equal to $x$. Define $\widehat{i}_{\mathcal{T}} = 1 + ((i-1) \mod \mathcal{T})$, where $(x \mod y)$ denotes the remainder after division of $x$ by $y$. Then for $1 \leq i \leq N$, the symbol $D_i$, i.e. the $i^{th}$ symbol in the sequence $D$, is generated by the random variable $X_{\widehat{i}_{\mathcal{T}}}$. The random variables $X_{\widehat{i}_{\mathcal{T}}}$ for $\widehat{i}_{\mathcal{T}} = 1, \ldots, \mathcal{T}$ are assumed to be independent. The parameters $P_1, \ldots, P_{\mathcal{T}}$, and $\mathcal{T}$ are unknown. Define $\Theta = \{\mathcal{T}, [P_1, \ldots, P_{\mathcal{T}}]\}$. The search space for parameter $\mathcal{T}$ is the set $B = \{1, \ldots, N_0\}$, for some $N_0 < N$ and for the pmfs $[P_1, \ldots, P_{\mathcal{T}}]$ the search space is the subset $\mathcal{Q} \subseteq [0,1]^{|\mathcal{X}| \times \mathcal{T}}$ of column stochastic matrices (for $P \in \mathcal{Q}$, $P_{ji} \in [0,1]$ and $\sum_{j=1}^{|\mathcal{X}|} P_{ji} = 1$ for $i = 1, \ldots, \mathcal{T}$). Let $\wp = B \times \mathcal{Q}$ denote the search space for the parameter $\Theta$. Given the data, the maximum aposteriori (MAP) estimate of parameter $\Theta$ is

$$\Theta_{\text{MAP}} = \arg\max_{\Theta \in \wp} \mathcal{P}(\Theta|D).$$

By Bayes rule the posterior probability is

$$\mathcal{P}(\Theta|D) = \frac{\mathcal{P}(D|\Theta)\mathcal{P}(\Theta)}{\mathcal{P}(D)},$$

where, by independence of $X_i$'s,

$$\mathcal{P}(D|\Theta) \;=\; \prod_{i=1}^{N} \mathcal{P}(X_{\widehat{i}_{\mathcal{T}}} = D_i|\Theta)$$

is the likelihood. Note that the probability $\mathcal{P}(D) = \int_{-\infty}^{\infty} \mathcal{P}(D|\Theta)\mathcal{P}(\Theta) d\Theta$ is a constant and thus, assuming a uniform prior on $\Theta$,

$$\Theta_{\text{MAP}} = \arg\max_{\Theta \in \wp} \mathcal{P}(D|\Theta) = \Theta_{\text{ML}}.$$

In words, the MAP estimate is same as the maximum likelihood estimate under the uniform prior assumption. The maximum likelihood estimates (MLE) for the unknown parameters are developed in the next section. However, as seen from the experimental results on simulated sequences and real gene data, 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. The estimator is not ad-hoc; it is derived using the refined minimum description length (MDL) principles. The penalization then corresponds to assuming the universal prior on the parameters and refined MDL estimator is essentially the MAP estimator with respect to the universal prior.

### *A. The Maximum Likelihood Estimate*

The derivation of the MLE is greatly simplified by adopting the following notation. Represent the data-sequence $D = [D_1, \ldots, D_N]$ by a sequence of vectors $\mathcal{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_N]$ where each $\mathbf{w}_i$ is a $|\mathcal{X}| \times 1$ vector with

$$\mathbf{w}_{ji} = \begin{cases} 1 & D_i = \mathcal{X}_j \\ 0 & \text{otherwise} \end{cases}. \tag{2}$$

For DNA sequences, if the $i^{th}$ symbol in the sequence $D$ is C, i.e. the third symbol of the alphabet $\mathcal{X}$, then the $i^{th}$ vector $\mathbf{w}_i$ in the sequence $\mathcal{W}$ is $[\,0\;0\;1\;0\,]'$. Also define a $|\mathcal{X}|\times\mathcal{T}$ stochastic matrix $\mathbb{A}$ with entries $\mathbb{A}_{ji} = \mathcal{P}(X_i = \mathcal{X}_j)$. The columns of the matrix $\mathbb{A}$ denote the pmfs of the information sources; the entry $\mathbb{A}_{ji}$ denotes the probability that the $i^{th}$ source generates the $j^{th}$ symbol of the alphabet $\mathcal{X}$. Write the unknown parameter $\Theta = [\mathbb{A}, \mathcal{T}]$. Then

$$\mathcal{P}(X_{\widehat{i}_\mathcal{T}} = D_i | \mathbb{A}, \mathcal{T}) = \prod_{j=1}^{|\mathcal{X}|} \left(\mathbb{A}_{j\widehat{i}_\mathcal{T}}\right)^{\mathbf{w}_{ji}}.$$

The likelihood can therefore be written as

$$\begin{aligned} \mathcal{P}(\mathcal{W}|\mathbb{A}, \mathcal{T}) &= \prod_{i=1}^{N} \mathcal{P}(X_{\widehat{i}_\mathcal{T}} = D_i | \mathbb{A}, \mathcal{T}) \\ &= \prod_{i=1}^{N} \prod_{j=1}^{|\mathcal{X}|} \left(\mathbb{A}_{j\widehat{i}_\mathcal{T}}\right)^{\mathbf{w}_{ji}} \\ &= \prod_{k=1}^{M} \prod_{\widehat{i}_\mathcal{T}=1}^{\mathcal{T}} \prod_{j=1}^{|\mathcal{X}|} \left(\mathbb{A}_{j\widehat{i}_\mathcal{T}}\right)^{\mathbf{w}_{ji^{(k)}}} \times \prod_{\widehat{i}_\mathcal{T}=1}^{N-M\mathcal{T}} \prod_{j=1}^{|\mathcal{X}|} \left(\mathbb{A}_{j\widehat{i}_\mathcal{T}}\right)^{\mathbf{w}_{ji^{(M+1)}}} \end{aligned} \tag{3}$$

where $i^{(k)} = (k-1)\mathcal{T} + \widehat{i}_\mathcal{T}$. Note that the first term on the right hand side of (3) captures the observations in $M$ complete periods (given the period $\mathcal{T}$) while the second product captures the observation over the last incomplete cycle. The corresponding log-likelihood is

$$\begin{aligned} \log \mathcal{P}(\mathcal{W}|\mathbb{A}, \mathcal{T}) &= \sum_{k=1}^{M} \sum_{\widehat{i}_\mathcal{T}=1}^{\mathcal{T}} \sum_{j=1}^{|\mathcal{X}|} \mathbf{w}_{ji^{(k)}} \log \left(\mathbb{A}_{j\widehat{i}_\mathcal{T}}\right) + \\ &\qquad \sum_{\widehat{i}_\mathcal{T}=1}^{N-M\mathcal{T}} \sum_{j=1}^{|\mathcal{X}|} \mathbf{w}_{ji^{(M+1)}} \log \left(\mathbb{A}_{j\widehat{i}_\mathcal{T}}\right) \end{aligned} \tag{4}$$

The MLE for $\mathbb{A}$ is first derived and then substituted in (4) to form the plug-in maximum-likelihood-estimator for $\mathcal{T}$. For a fixed $\mathcal{T}$, the MLE for $\mathbb{A}$ is given as

$$\mathbb{A}_{\text{ML}}^{\mathcal{T}} = \arg\max_{\mathbb{A}\in\mathcal{Q}} \log \mathcal{P}(\mathcal{W}|\mathbb{A}, \mathcal{T}). \tag{5}$$

Equivalently,

$$\mathbb{A}_{\text{ML}}^{\mathcal{T}} = \arg\min_{\mathbb{A}\in\mathcal{Q}} -\log \mathcal{P}(\mathcal{W}|\mathbb{A}, \mathcal{T}). \tag{6}$$

The log-likelihood in (4) is a concave function of variables $\mathbb{A}_{j\widehat{i}_{\mathcal{T}}}$ which also satisfy the constraints: $\sum_{j=1}^{|\mathcal{X}|} \mathbb{A}_{j\widehat{i}_{\mathcal{T}}} = 1$ for $\widehat{i}_{\mathcal{T}} = 1, \ldots, \mathcal{T}$. Constrained optimization using Lagrange multipliers gives the $(j, \widehat{i}_{\mathcal{T}})^{th}$ element of the matrix $\mathbb{A}_{\text{ML}}^{\mathcal{T}}$ as

$$\mathbb{A}_{\text{ML}}^{\mathcal{T}}(j, \widehat{i}_{\mathcal{T}}) = \begin{cases} \frac{1}{M+1} \sum_{k=1}^{M+1} \mathbf{w}_{ji^{(k)}}, & \widehat{i}_{\mathcal{T}} = 1, \ldots, N - M\mathcal{T} \\ \\ \frac{1}{M} \sum_{k=1}^{M} \mathbf{w}_{ji^{(k)}}, & \widehat{i}_{\mathcal{T}} = N - M\mathcal{T}, \ldots, N \end{cases} \tag{7}$$

for $j = 1, \ldots, |\mathcal{X}|$. The MLE for the probability mass functions of the random variables, given the period, is quite intuitive. Given the period is $\mathcal{T}$, it amounts to segmentation of the data sequence into $\mathcal{T}$ non-overlapping subsequences. Then the pmf of the $k^{\text{th}}$ information source is given by the relative frequency of each symbol in the $k^{\text{th}}$ subsequence. For instance, if the hypothesized statistical period in a gene sequence is $3$ then the MLE of the pmf of the $2^{\text{nd}}$ information source is given by the empirical probabilities of nucleotides in the subsequence comprising of every third symbol, starting with the second symbol.

The estimates of the parameter $\mathbb{A}$ can be used to determine the MLE for the period $\mathcal{T}$,

$$\mathcal{T}_{\text{ML}} = \arg\min_{\mathcal{T}\in B} \quad -\log \mathcal{P}(\mathcal{W}|\mathbb{A}_{\text{ML}}^{\mathcal{T}}, \mathcal{T}). \tag{8}$$

This is a simple plug-in estimator where the search is over a collection of models with complexity that is increasing with $\mathcal{T}$. In each model, the best fit for the data is picked - this is the MLE $\mathbb{A}_{\text{ML}}^{k}$, given the period $k$. This set of MLEs, from different models, indexed by $k$, are then compared for the goodness-of-fit, in terms of the likelihood.

### *B. Minimum description length estimator*

The minimum description length (MDL) principle is an important tool for statistical inference. It has been applied successfully to the problem of *model selection* to determine which of the possible explanations of the data is the best given a finite number of observations. The fundamental idea or the intuition behind MDL is that more regular

the data is, the easier it is to compress and thus learn [23]. For instance in a homological sequence, a single period captures the entire data whereas a sequence of coin-tosses is completely random and there may not be any shorter description of the data than the data itself. Most of the 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. If two models fit data equally well, it picks the simpler one - in that sense it is like Occam's Razor.

The key intuition for minimum description length principle is that learning from the data is equivalent to data compression. However, data compression varies with the choice of the description method. Kolmogorov described the complexity of a data sequence as length of the shortest program in a general purpose programming language that generates the sequence and halts. It may seem that Kolmogorov complexity of the data is dependent on the computer language used but a famous result, the *invariance theorem*, states that for long enough sequences, the Kolmogorov complexity with respect to two different programming languages differs only by a constant that does not depend on data. However, Kolmogorov complexity is not computable and MDL procedure based on it becomes arbitrary for small data samples. Thus, much of the focus in MDL is at simpler description methods such that for any data sequence the length of the shortest description is computable. Then, given the data set $D$ and a collection of hypothesis $\mathcal{H}$, the MDL principle for model selection is to pick the hypothesis that compresses the data most with respect to the description method.

Let $D$ denote the data and let $\mathcal{H}^{(1)}, \mathcal{H}^{(2)}, \ldots$ be a list of candidate models or hypotheses, where $\mathcal{H}^{(k)} = \{Q | Q \text{ is an } M \times k \text{ column-stochastic matrix}\}$ for $k = 1, \ldots, N_0$. Define $\mathcal{H} = \cup_{k=1}^{N_0} \mathcal{H}^{(k)}$. Then the best explanation of the data $D$ is the hypothesis $H \in \mathcal{H}$ that minimizes the description length

$$L(D|\mathcal{H}) = L(\mathcal{H}^{(k)}) + L(D|H_{\text{ML}}^{(k)}) \tag{9}$$

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

The second term $L(D|H_{\text{ML}}^{(k)})$ in (9) is the codelength of the data when encoded with the hypothesis $H_{\text{ML}}^{(k)}$. Assuming the hypotheses are probabilistic, the Shannon-Fano code are optimal in terms of the expected codelength. Thus,

$L(D|H^{(k)}_{\text{ML}}) = -\log P(D|H^{(k)}_{\text{ML}})$, where $P(D|H^{(k)}_{\text{ML}})$ is the probability of observing $D$ conditioned on the hypothesis $L(D|H^{(k)}_{\text{ML}})$. The codelength is therefore the negative-log-likelihood of having observed the data $D$. This term is exactly the same as in previous section, with $H^{(k)} = \mathbb{A}^k$.

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 $Q \in \mathcal{H}^{(k)}$ are described by $k' = Mk$ frequencies or probabilities that are determined by the counts in the set $\{0, 1, \ldots, \lceil \frac{N}{k} \rceil\}$, thus taking $k' \log(\lceil \frac{N}{k} \rceil + 1)$ bits. The total codelength for the code is therefore

$$L(H) + L(D|H) = 2\lceil \log k \rceil + 1 + Mk \log \lceil \frac{N}{k} \rceil - \log P(D|H) \tag{10}$$

for $H \in \mathcal{H}^{(k)}$. It is clear from (10) that the MDL principle yields a penalized ML estimate. The code used here is a *universal code* and implies a universal prior on the hypothesis.

## III. Examples and Applications

We discuss some applications of studying cyclostationary structure of symbolic DNA and RNA sequences in this section. Section III-A applies the methods of Section II to both simulated and real gene sequences. The methods are extended to consider spatially varying periodicities in symbolic DNA sequences using a windowed approach in Section III-B, and Section III-C shows how the same ideas can be generalized to analysis of secondary structures in RNA.

### *A. Finding Periodicities in DNA Sequences*

For testing, a homological symbolic sequence from the set $\mathcal{X} = \{\text{A}, \text{G}, \text{C}, \text{T}\}$ with period $\mathcal{T} = 7$ was generated. 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 Figure 2(a). The periodic behaviour is very evident from the plots. Also notable are the sub-harmonics, i.e. the integer multiples of the true period. The plots strongly support a statistical periodicity of 7 even with $60\%$ erosion. The noise floor in the plots increases with erosion and at $75\%$ erosion, the sequence exhibits no repetitive behaviour. The dotted red plot was obtained by a variant of computational negative controls (CNC) strategy proposed in [24] - it corresponds to the negative log-likelihood for various permutations of the original sequence. It provides a good reference for

comparison since a random permutation would destroy any 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 statistically significant.

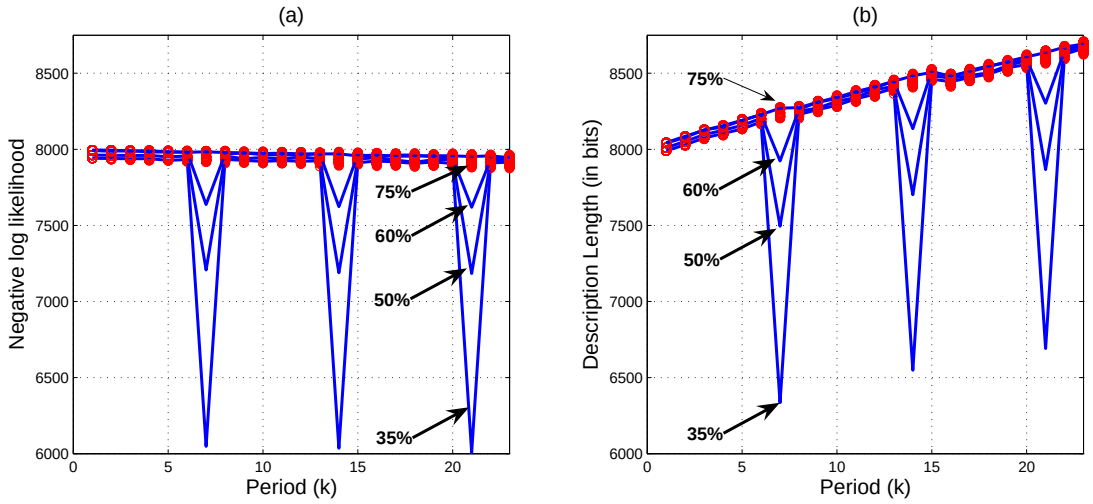


Fig. 2. (a) Negative log-likelihood for the ML estimate plotted against Period for a simulated symbolic sequence of length 4000, with period 7 under $35\%$, $50\%$, $60\%$ and $75\%$ erosion, (b) Description length (in bits) plotted for the ML estimate in $\mathcal{H}^{(k)}$ plotted against $k$ for corresponding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

The algorithm was also tested with the protein coding region of chromosome III of $\mathcal{S}$. *cerevisiae* [25]. The 1629 base-pair (bp) long sequence (bp: 6,571 - 8,199) shows a latent periodicity of period three in Figure 3(a). The period-3 behaviour of protein coding genes is expected since amino acids are coded by trinucleotide units called *codons* [9], [26]. For comparison, the symbolic sequence is transformed into a numerical sequence using the complex mapping developed in [9] for identification of protein coding regions (A = $0.1 + 0.12i$, G = $0.45 - 0.19i$, C = $0$, T = $-0.3 - 0.2i$). The magnitude of the $1629$-point DFT of numerical sequence of poly-nucleotides is plotted against the frequency in Figure 3(b). The peaks at $f_1 = 543$ and $f_2 = 272$ correspond to $3$ and $6$-periodic behaviour respectively; however, some other peaks are simply the artifacts, perhaps of the numerical mapping.

The MLE is compared with the MDL estimator in Figure 2 for simulated sequences and in Figure 4(a),(b) for $191$ base pair long sequence from Chromosome XVI (bp: 521,009 - 521,199) of the $\mathcal{S}$. cerevisiae Genome [25]. The problem of *overfitting* is evident from the negative tilt of 'valleys' in the plots. This behaviour is manifested by equation (8), giving the largest integer multiple of $\mathcal{T} \in B$. However, the MDL estimator resolves the issue by penalizing the models commensurately with complexity.

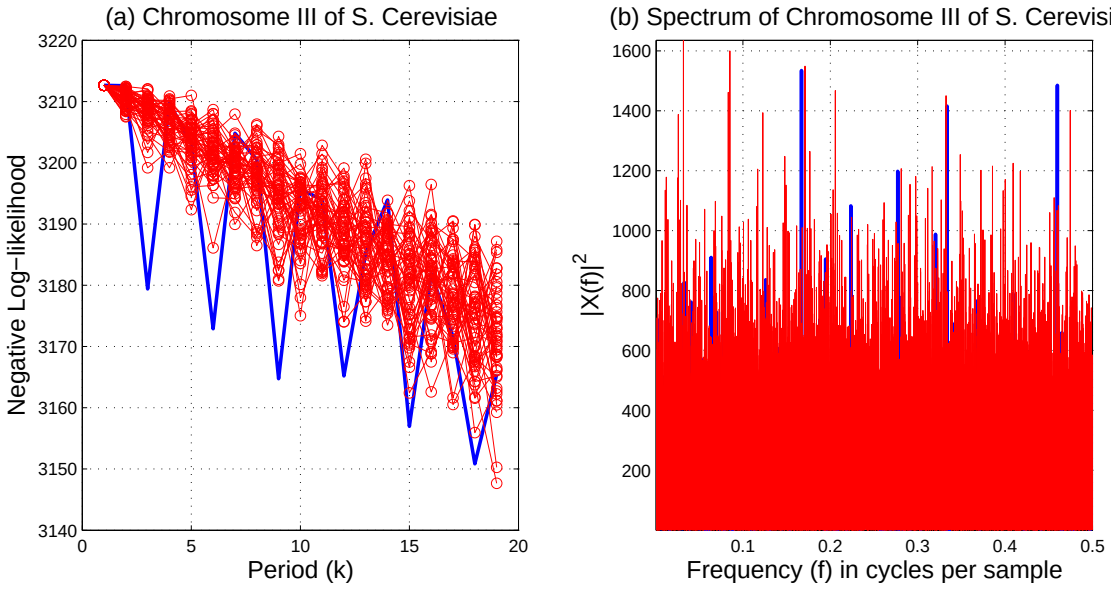


Fig. 3. (a) Negative Log-likelihood for ML estimate plotted against period for the 1629 base-pair long sequence from the protein-coding region of chromosome III (bp: 6,571 - 8,199) of $\mathcal{S}$. *cerevisiae* genome, (b) the magnitude of DFT of numerical sequence derived from the same sequence. The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

Figure 4(c) shows results where the symbol sequence is generated by a latent periodicity where a single period is given by equation (1). The plot reveals a strong six-periodic behaviour 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), Figure 4(d) shows that no significant periodicities are detected, the minimum MDL occurs at a "periodicity" of period one.

Although the method of Anastassiou [9] and other numerical representation techniques combined with Fourier transform perform poorly at severe mutation rates (see Figure 3), their performance in low noise conditions is comparable to the MDL estimator. Figure 5 shows results for 1305 base pair long sequence from Chromosome 20 (bp:22,557,488-22,558,792) of the Human Genome [25]. 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 behaviour 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.

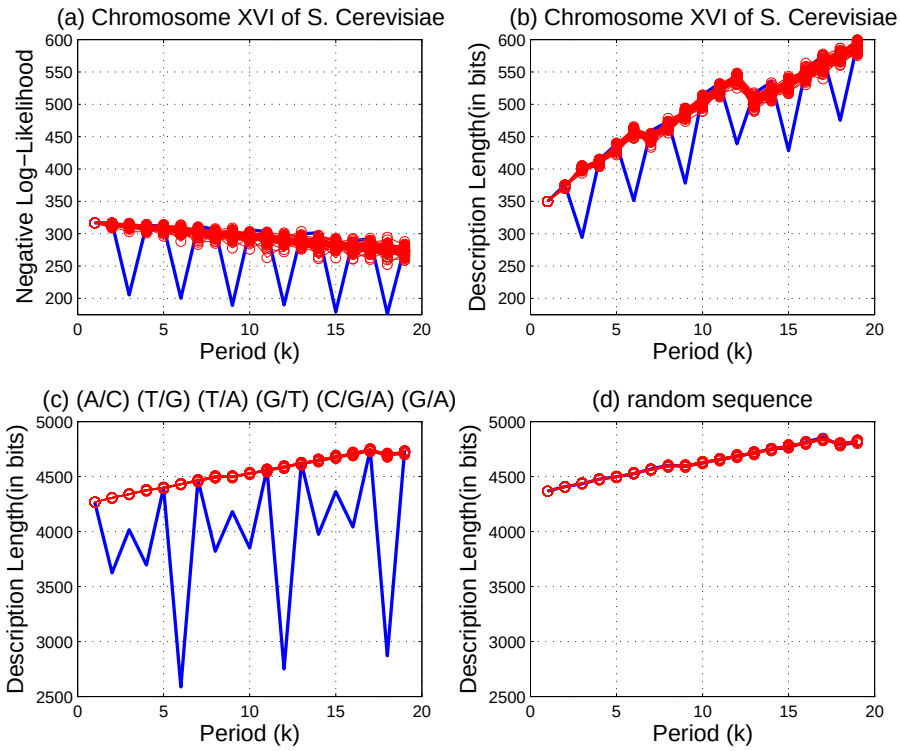


Fig. 4. (a) Negative log-likelihood for ML estimate of the protein coding region of chromosome XVI (bp: 521,009 - 521,199) of $\mathcal{S}$. *cerevisiae* genome. Description length (in bits) plotted for the penalized ML estimate in $\mathcal{H}^{(k)}$ plotted against $k$ for (b) the protein coding region 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 completely random symbolic sequence. The CNC variants are plotted in red.

### *B. Identifying Exons in DNA sequences*

The cyclostationarity profile of DNA sequences varies with location. The coding part of DNA, in particular, displays statistical periodicity with period three. The varying periodicities in DNA can be discovered by using sliding windows and a cumulative sum test is presented in this section to detect the change points. 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 $M < N$ with an overlap of $H$ symbols between successive windows. The method presented here is similar to windowed Fourier transform techniques for generating the spectrogram in [16], [27], [28], except that no numerical mapping is imposed in this paper.

Figure 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 completely random in the middle.

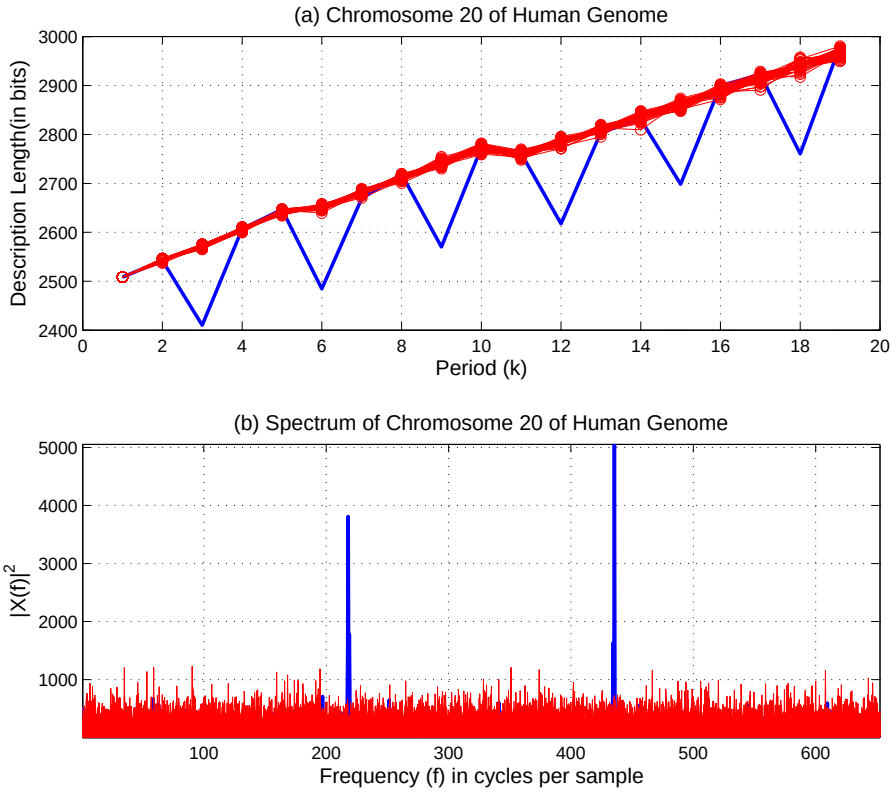


Fig. 5. (a) Description length (in bits) for the ML estimate in $\mathcal{H}^{(k)}$ plotted against $k$ for the protein coding region of chromosome 20 of human genome; (b) The magnitude of DFT of numerical sequence derived from the protein coding region of chromosome 20 of Human genome. The CNC variants are plotted in red.

Thus there are two *change points* in the sequence. The latent period of the periodic 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 behaviour is very evident from the plot as are the sub-harmonics.

The sliding window method was applied to chromosome 20 of the human genome [25]. 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 (*exons*) flanked by non-coding parts (*introns*) on both sides. The contour plot in Figure 7 shows a latent periodicity of period three beginning at sliding window number 60 which corresponds to bp 22,557,427 ($M = 750$, $H = 75$). This period-3 behaviour of protein coding genes is expected since amino acids are coded by trinucleotide units called *codons* [9].

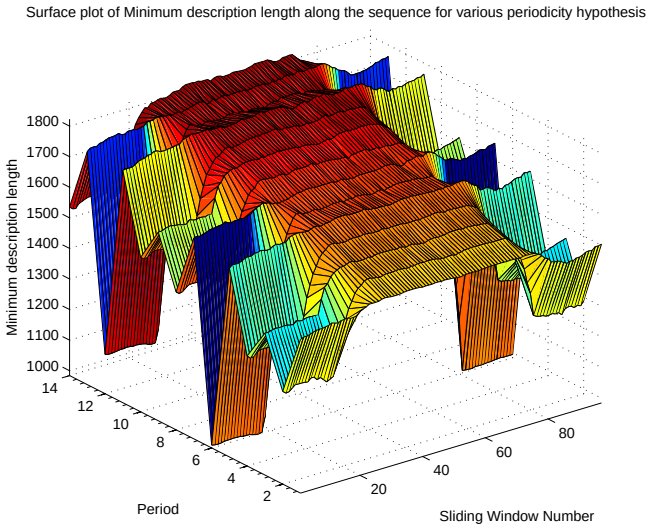


Fig. 6. Description length (in bits) for the ML estimate in $\mathcal{H}^{(k)}$ plotted against period $k$ along the sequence.

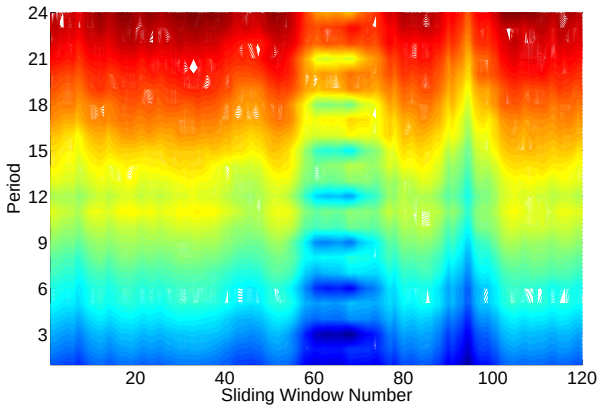


Fig. 7. Contour plot of description length (in bits) for the ML estimate in $\mathcal{H}^{(k)}$ plotted against period $k$ along the sequence.

The window size $M$ determines a 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 neighbourhood. Another problem with poor resolution is detecting two change points that are very close to each other. For instance, if the random part of the sequence in Figure 6 is much smaller than the window size, the change points may go undetected. A multi-resolution multi-scale technique may therefore be preferred where various sizes for the sliding window are used. A coarse search is first performed followed by a fine search in the regions of interest.

Near the change points, the periodicity profile changes, while in other parts the profile remains constant except for some small fluctuations due to the noise in data. Thus a uniformly most powerful (UMP) test may be constructed based on the positive inflection rate over multiple successive windows. If the maximum likelihood period reported is $P$ then the alternate composite hypothesis is that the period is no longer $P$. 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

$$\mathbf{\Theta}_t^{(P)} = \min_{m \in \{1,\ldots,T\}} |\mathbf{Q}_{\text{ML},t}^{(P)} - \mathbf{Q}_{\text{ML},t-m}^{(P)}|_{\text{tot}} > \delta_{\text{Th}} \tag{11}$$

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

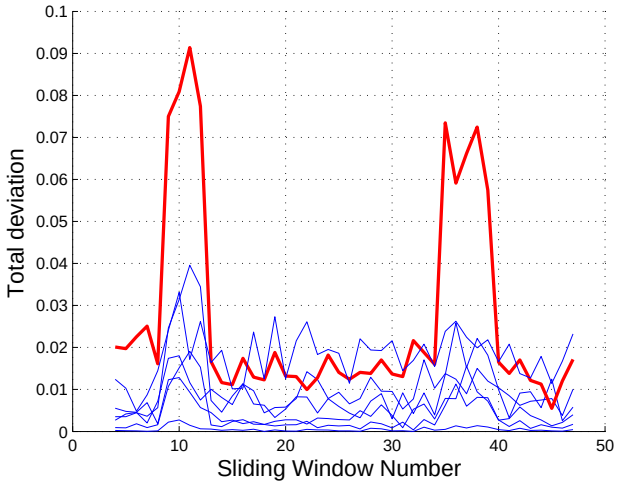


Fig. 8. $\mathbf{\Theta}_t^{(P)}$ plotted for the sequence from Figure 6. $\mathbf{\Theta}_t^{(6)}$ is plotted in red ($M = 750, H = 150, T = 3$).

### *C. RNA structure analysis*

Till recently, RNAs were considered to be passive intermediary messengers (mRNA) of genetic information from DNA to protein via the process of translation. During the last decade, RNAs have been found to play several

important non-coding functions including chromosome replication, protein degradation and translocation, regulating gene expression and many more. Such RNAs are called non-coding RNAs (ncRNAs) or RNA genes. The number of ncRNAs in human genomes is in the order of tens of thousands and considering the vast amount of genomic data there is a need for computational methods for identification of ncRNAs [26].

The statistical model presented in this paper for finding periodicities in symbolic sequences can be utilized for building probabilistic representations of RNA families. RNA has the same primary structure as DNA, consisting of a sugar-phosphate backbone with nucleotides attached to it. However, in RNA the nucleotide Thymine(T) is replaced by Uracil (U) as the base complementary to Adenine (A). So, RNA is represented by a string of bases: A, C, G and U. RNA exists as a single-stranded molecule since the replacement of Thymine by Uracil makes RNA too bulky to form a stable double helix. However, the complementary bases (A and U, G and C) can form a hydrogen bond and such consecutive base pairs cause the RNA to fold onto itself resulting in 2-D and 3-D secondary and tertiary structures. A typical secondary structure is *hairpin* structure as shown in Figure 9(a); the consecutive base pairs that bond together get stacked onto each other to form a *stem* while the unpaired bases form a *loop*.

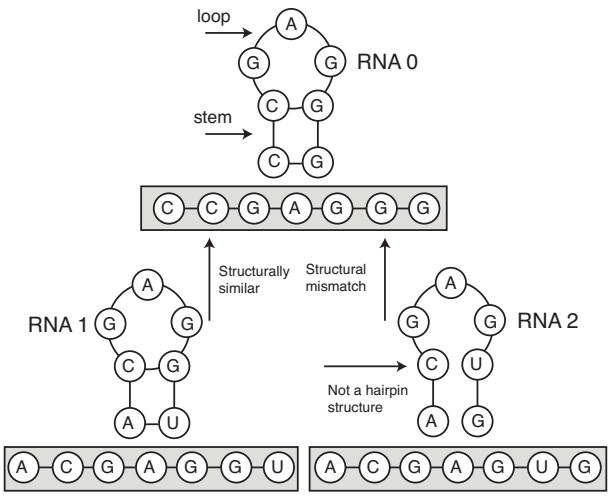


Fig. 9. (a) RNA0 has hairpin secondary structure. (b) RNA1 is similar in structure to RNA0. It differs at two positions in the primary sequence from RNA0. (c) RNA2 structure is not hairpin, it has a structural mismatch with RNA0. RNA2 also differs at two position in the primary sequence from RNA0 but it must be scored lower in similarity to RNA0 as compared to RNA1.

Typical methods employed for identification of DNA gene sequences and proteins do not perform as well in identification of ncRNAs because they are based on finding structural features (like periodicities) in primary sequences whereas most functional ncRNAs preserve their secondary structures more than they preserve their primary

sequences [26] as seen in Figure 9. Therefore, there is need for techniques that also evaluate similarity between secondary structures. Such techniques have been shown to be more effective in comparing and discriminating RNA sequences [29]. We develop signatures for RNA sequences that can discriminate between different secondary structures. These signatures find application in multiple alignment and database search of RNA sequences.

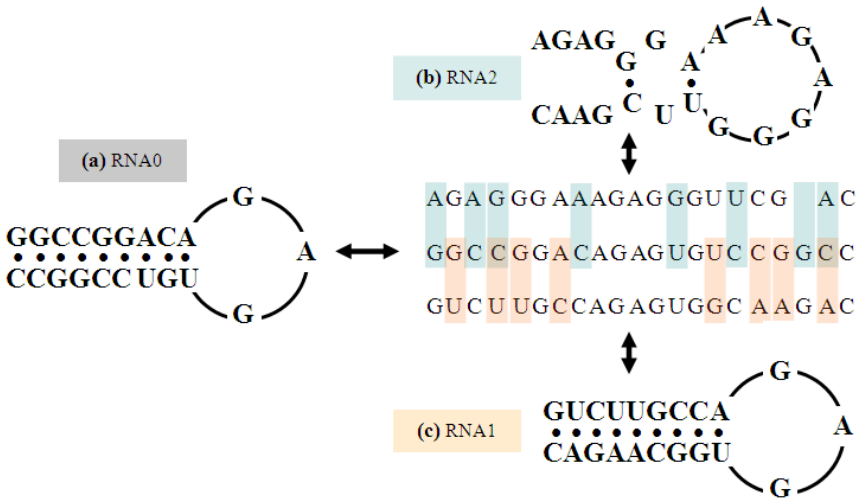


Fig. 10. Comparing the primary sequence and secondary structures of (a) RNA0 with hairpin structure, (b) RNA1 evolved from RNA0 under compensatory mutation and (c) RNA2 that appears to have been evolved from RNA0 but is structurally different.

RNA sequences preserve the secondary structure via compensatory mutations which cause strong pairwise correlations between distant bases in the primary RNA sequence. Unlike the techniques employed for DNA identification in earlier works, the approach presented here can describe such pairwise correlations. Consider three ncRNA sequences shown in Figure 10. In multiple alignments or database searches, the objective often is to determine if the given sequences are homologous. RNA0 and RNA1 have hairpin secondary structure and the two sequences differ at eight base positions. The sequence RNA2 also differs from RNA0 at eight base locations but it has a different secondary structure and must be scored lower in similarity to RNA0 as compared to RNA1. In order to determine structural similarity, two binary symbolic sequences are obtained from the given reference ncRNA sequence. The first sequence is generated by replacing symbols A and G with M and symbols C and U with M′. The second sequence is obtained by replacing symbols A and C with N and symbols G and U with N′. The $k$-periodic source distribution matrices are estimated for the two binary sequences as described in (4) for $1 \leq k \leq N_0$; let $\mathbb{A}^{(k)}$ and $\mathbb{B}^{(k)}$ denote the corresponding matrices. Then the following equation describes a sequence of similarity

scores

$$\Delta^{(k)}_{\text{RNA}} \;=\; -\sum_{i,j=1}^{2} \mathbb{A}^{(k)}(i,j)\log(\mathbb{A}^{(k)}(i,j)) \;\; -\sum_{i,j=1}^{2} \mathbb{B}^{(k)}(i,j)\log(\mathbb{B}^{(k)}(i,j)), \tag{12}$$

for $k = 1, \ldots, N_0$. Various linear combinations of $\Delta^{(k)}_{\text{RNA}}$ yield multi-dimensional signatures for ncRNAs. RNA sequences in Figure 10 give following distribution matrices for the 2-periodic model: for RNA0 and RNA1

$$\mathbb{A}^{(2)} = \begin{bmatrix} 5/11 & 6/10 \\ 6/11 & 4/10 \end{bmatrix}, \;\; \mathbb{B}^{(2)} = \begin{bmatrix} 5/11 & 4/10 \\ 6/11 & 6/10 \end{bmatrix}, \tag{13}$$

and corresponding matrices for RNA2 are

$$\mathbb{A}^{(2)} = \begin{bmatrix} 8/11 & 9/10 \\ 3/11 & 1/10 \end{bmatrix}, \;\; \mathbb{B}^{(2)} = \begin{bmatrix} 8/11 & 2/10 \\ 3/11 & 8/10 \end{bmatrix}. \tag{14}$$

Computing the scores in equation (12) gives $\Delta^{(2)}_{\text{RNA0}} = 3.93$, $\Delta^{(2)}_{\text{RNA1}} = 3.93$ and $\Delta^{(2)}_{\text{RNA1}} = 2.88$. The absolute difference of the scores results in good discrimination of secondary structures even in the face of significant mutations. The quantity $\Delta^{(2)}_{\text{RNA}}$ gives *compensatory-mutation-invariant signature* for some secondary RNA structures: hairpin with odd number of bases in the loop as shown above and for certain pseudoknots as well. Consider the RNA *inhibitor of HIV reverse transcriptase* [30], which has a *pseudoknot* structure, and its possible homologues shown in Figure 11. Computing the secondary-structure similarity score gives $\Delta^{(2)}_{RNA0} = 3.8825 = \Delta^{(2)}_{RNA1}$ and $\Delta^{(2)}_{RNA2} = 2.8912$. In general, however, several linear combinations of $\{\Delta^{(k)}_{\text{RNA}}\}_{k=1}^{N_0}$ should be used to generate a multi-dimensional signature [31].

The statistical periodicity model provides a framework for systematically developing signatures for the varied class of RNA secondary structures. These signatures find application in multiple alignments of instances of similar RNAs from different genomes (for example human, rat, chicken) and in database search of homologues of a given RNA. A family of related RNAs often share a common secondary structure besides similar primary sequence motifs. When searching a sequence database for homologous RNAs, it will be advantageous to combine the structural signatures with the primary sequence similarity scores. For instance, in Figure 10 the RNA of interest is the sequence RNA0 with hairpin structure and conserved loop motif GAG - as seen above the invariant signature based on $\Delta^{(2)}_{\text{RNA}}$ determines the sequence RNA1 to be homologous to RNA0. The RNA signatures are also useful for consensus structure prediction from multiple alignments by the process of comparative RNA sequence analysis [30], [32]. In a structurally correct multiple alignment of RNAs (sequences RNA0 and RNA1 in Figures 10 and 11) the conserved

base pairs (shaded symbols) are revealed by presence of correlated compensatory mutations. The invariants provide a quantitative measure of pairwise sequence covariation.

Many current RNA pattern-matching algorithms are based on RNAMOT [33] and search for deterministic motifs with secondary structure constraints. These methods typically work best for small, well-defined patterns but become increasingly inaccurate with less conserved sequences [30]. Another shortcoming of existing methods is that they need to be carefully customized for each RNA of interest and the context-free-grammar based algorithms are incapable of describing the pseudoknots. The RNA signatures introduced in this section do not present these limitations [31].

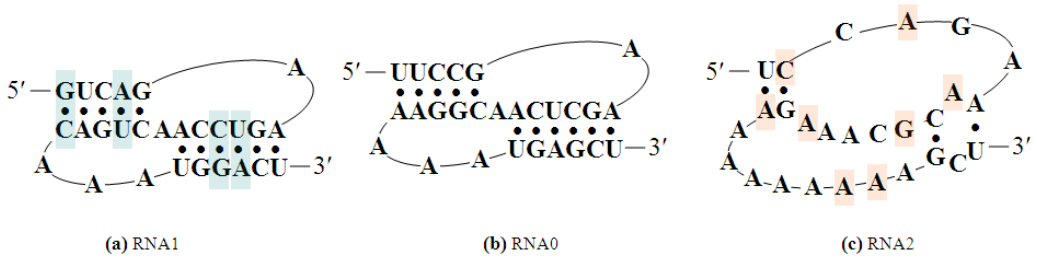


Fig. 11. (a) RNA0 - the RNA inhibitor of HIV reverse transcriptase [30] with pseudoknot structure (b) RNA1 - mutated from RNA0 at shaded base locations (c) RNA2 sequence with a pseudoknot and an internal loop structure. RNA2 is structurally different from RNA1 and RNA0.

## IV. Multiple Periodicities

Multiple latent periodicities in symbolic sequences provide evidence of mutations and can help reconstruct the evolution history just like numerical sequences. In numerical sequences, if multiple periodicities result from addition (composition) of several sequences with different periods, then Periodicity Transforms [34] provide the decomposition 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 that properly defines the notion of multiple periodicities. The mathematical structure of the periodic subspaces is studied first, and the resulting algebraic properties allow a decomposition of multiple periodicities.

### *A. Periodic Subspaces*

Let $\mathcal{X} = \{a_1, \ldots, a_M\}$ be a finite alphabet with cardinality $M$. Let $\mathcal{P}_p$ be the collection of cyclostationary sequences on $\mathcal{X}$ with period $p$. Then $\mathcal{P} = \bigcup_{p>0} \mathcal{P}_p$ is the set of all cyclostationary sequences on $\mathcal{X}$ where $p$ ranges

over all positive integers. The set $\mathcal{P}_p$ can also be identified with the set of $M \times p$ column stochastic matrices. An element $S \in \mathcal{P}_p$ is a sequence of random variables and is described by an $M \times p$ 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 $n \in \mathbb{Z}^+$, i.e.

$$P(S_{np+i} = a_j) = P(S_i = a_j) = \mathbf{Q}_{ji}^S \equiv \mathbf{q}_i^S(j) \tag{15}$$

where $j = 1, \cdots, M$. The following law of composition on the pmfs of the random symbolic sequences follows the double carousel model of Figure 1 in analogy with the gene replication process. Define

$$\begin{aligned} \oplus : \mathcal{P} \times \mathcal{P} &\rightarrow \mathcal{P} \\ (X, Y) &\mapsto Z \end{aligned} \tag{16}$$

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 $a \in \mathcal{X}$

$$P\Big(Z_n = a\Big) = P\Big(X_{\hat{n}_p} = a, Y_{\hat{n}_q} = a \Big| X_{\hat{n}_p} = Y_{\hat{n}_q}\Big). \tag{17}$$

Note that the binary operation is defined on the matrices $\mathbf{Q}^X, \mathbf{Q}^Y$ but 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 $a \in \mathcal{X}$, $P\Big(Z_m = a\Big) = P\Big(X_{\hat{n}_p} = a, Y_{\hat{n}_q} = a \Big| X_{\hat{n}_p} = Y_{\hat{n}_q}\Big) = P\Big(Z_n = a\Big)$.* ■

**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 (17), the $n^{th}$ column of the $M \times pq$ matrix $\mathbf{Q}^Z$ is

$$\mathbf{q}_n^Z = \frac{1}{C} \begin{bmatrix} \mathbf{q}_{\hat{n}_p}^X(1)\mathbf{q}_{\hat{n}_q}^Y(1) \\ \vdots \\ \mathbf{q}_{\hat{n}_p}^X(M)\mathbf{q}_{\hat{n}_q}^Y(M) \end{bmatrix} \tag{18}$$

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

**Example 1.** *Consider an example of composition of two cyclostationary sources with statistical periods* 2 *and* 3*. Eqn. (18) 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 $X$ 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

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

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

$$\begin{aligned} \circ : \mathbb{R} \times \mathcal{P} &\rightarrow \mathcal{P} \\ (r, X) &\mapsto Z \end{aligned} \tag{19}$$

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

$$P\big(Z_n = a\big) = \frac{P(X_n = a)^r}{\sum_{b\in\mathcal{X}} P(X_n = b)^r} \tag{20}$$

for all $a \in \mathcal{X}$ 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 $\mathbf{q}_i^Y = [\frac{2}{3}\ \frac{1}{6}\ \frac{1}{6}\ 0]^T$.*

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 $\oplus : \mathcal{P} \times \mathcal{P} \rightarrow \mathcal{P}$.*

*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 $\mathbf{Q}^V_{ji}$ can be rewritten as*

$$\frac{\mathbf{Q}^X_{ji_p}\left(\mathbf{Q}^Y_{ji_q}\mathbf{Q}^Z_{ji_r}\right)}{\sum_j \mathbf{Q}^X_{ji_p}\left(\mathbf{Q}^Y_{ji_q}\mathbf{Q}^Z_{ji_r}\right)} = \frac{\left(\mathbf{Q}^X_{ji_p}\mathbf{Q}^Y_{ji_q}\right)\mathbf{Q}^Z_{ji_r}}{\sum_j \left(\mathbf{Q}^X_{ji_p}\mathbf{Q}^Y_{ji_q}\right)\mathbf{Q}^Z_{ji_r}} = \mathbf{Q}^W_{ji}$$

*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 $a_j \in \mathcal{X}$. 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 in closed under the binary law defined in (16). The periodic structure of a random sequence is thus preserved under composition and the resulting sequence exhibits periodicities 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 $\alpha \in \mathbb{R}$, $X \in \mathcal{P}_p$ 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 $\mathcal{P}$.*

The significance of Theorem 2 is that it allows for varying degrees of constituent periodicities. A symbolic sequence 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 [35]) exhibits multiple

latent periodicities with period $3$ and $5$. The description length (in bits) is plotted against the period for the base pairs 273-561 in Figure 12. The statistical significant periods seen are $3$ and $5$ as well as the sub-harmonics $6, 9, 12$ and $10, 15, 30$ and 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 [35]) 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 [35]) 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 [35]) 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 period exhibit multiple periodicities. The minimum description length is plotted in Figure 13 for two sequences with periodicity of $341$. One of the sequences exhibits strong $11$-periodic and $31$-periodic behaviour 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 $341$ long probabilistic source.

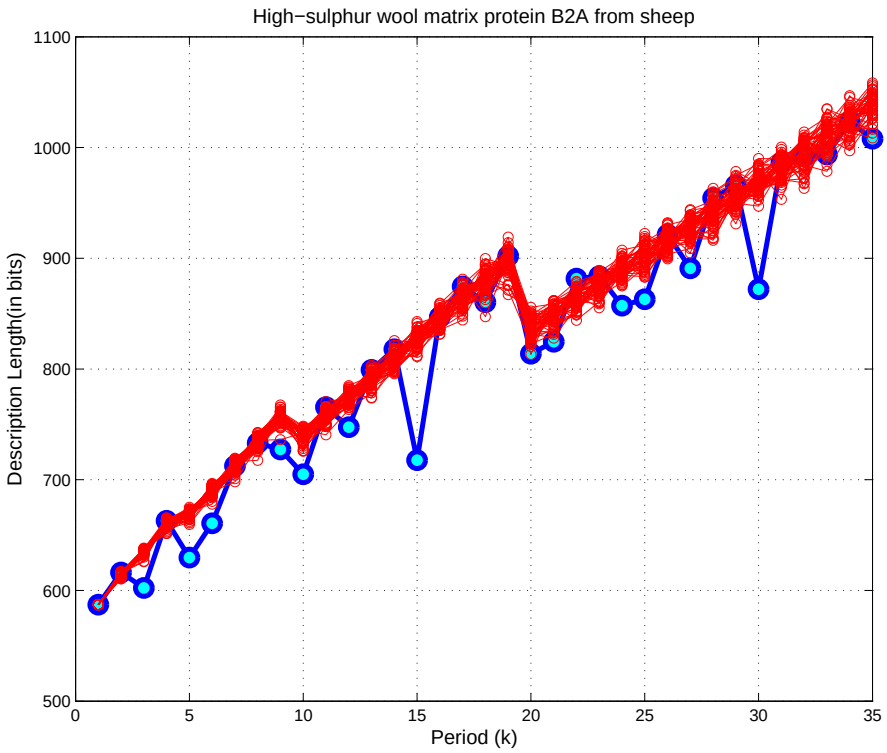


Fig. 12. Description length (in bits) plotted against the period for high-sulphur wool matrix protein B2A from sheep (bp:273-561). The DNA sequence exhibits multiple latent periodicities with period $3$ and $5$.

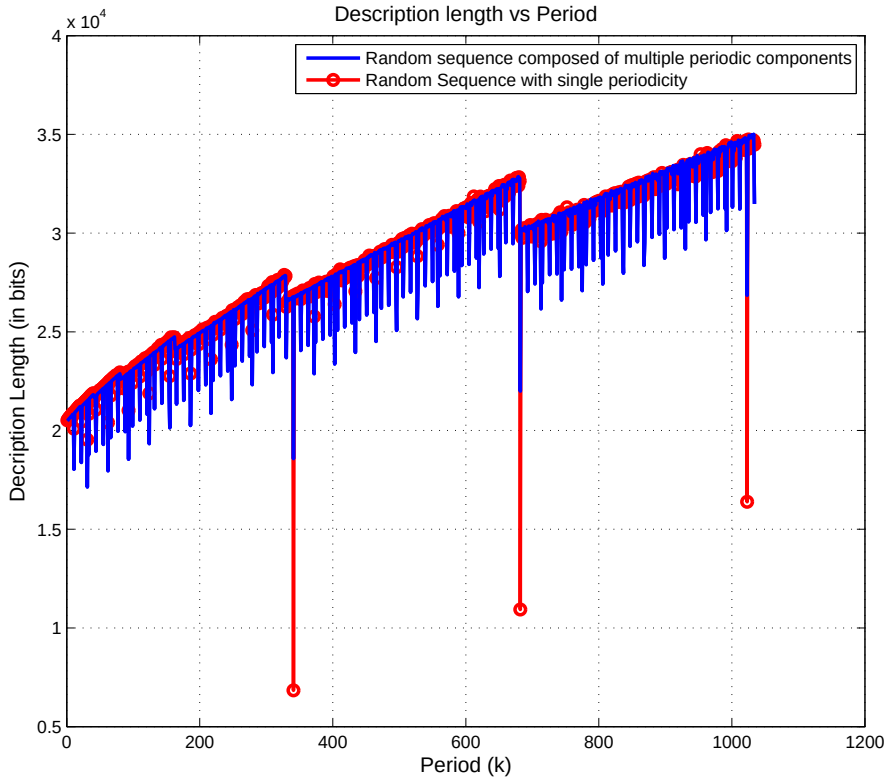


Fig. 13. Description length (in bits) plotted against the period for two cyclostationary sequences both with period 341. The lower curve (in blue) corresponds to the sequence composed of two cyclostationary sources with period 11 and 31.

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 system of equations can be expressed in matrix form as

$$\begin{bmatrix} Z_1 \\ \vdots \\ Z_{pq} \end{bmatrix}_{pq \times 1} = \underbrace{\begin{bmatrix} I_p & I_q \\ \vdots & \vdots \\ I_p & I_q \end{bmatrix}}_{\mathbf{T}_{pq \times (p+q)}} \circ \begin{bmatrix} X_1 \\ \vdots \\ X_p \\ Y_1 \\ \vdots \\ Y_q \end{bmatrix}_{(p+q) \times 1} . \tag{21}$$

**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

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

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 $X, Y$ 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 $\delta = \left[\frac{2}{10}\ \frac{3}{10}\ \frac{3}{10}\ \frac{2}{10}\right]^T$ and $-1 \circ \delta = \left[\frac{3}{10}\ \frac{2}{10}\ \frac{2}{10}\ \frac{3}{10}\right]^T$. The dominant periods of $X'$ and $Y'$ are $D^*_{X'} = [\ (G/C)A\ ]$ and $D^*_{Y'} = [T\ \ T\ A]$ respectively.* ■

On comparing the dominant periods in examples 1 and 3 it is observed that there is more than one decomposition, in terms of latent 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 $(\text{GCU})_n$ 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 $\mathbb{A}_{\text{ML}}^{(r)}$ whereas $\mathbb{A}_{\text{ML}}^{(p)}$ and $\mathbb{A}_{\text{ML}}^{(q)}$ 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 fixed while the parameters of other are adapted so as to minimize the total deviation between $\mathbf{Q}^X \oplus \mathbf{Q}^Y$ and $\mathbb{A}_{\text{ML}}^{(r)}$. 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 parts of DNA sequences exhibit characteristic statistical periodicities. Mapping this behaviour 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 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 a 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 [36].

## VI. 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

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

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,

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

where $\mathbf{1}_{pq}$ is a $pq \times 1$ vector of all ones. Clearly then,

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

Therefore, $\mathbf{T}$ is not full-rank and $\mathbf{w}$ 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

$$\mathbf{T}' = [\mathbf{T}_1 \dots \mathbf{T}_{k-1} \ \mathbf{T}_{k+1} \dots \mathbf{T}_{p+q}]$$

consisting of all but the $k^{th}$ column of $\mathbf{T}$. Note that the $j^{th}$ row of $\mathbf{T}$ has unity at two locations: $\widehat{j}_p$ and $p + \widehat{j}_q$. Define

$$J = \left\{ j \in \{1, \dots, pq\} \ \mid \ \widehat{j}_p = k \text{ or } \ p + \widehat{j}_q = k \right\}.$$

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

$$[\mathbf{T}'\mathbf{v}]_j = \begin{cases} \mathbf{v}_{p+1+\widehat{(j-1)}_q}, & j \in J \\ \mathbf{v}_{1+\widehat{(j-1)}_p} + \mathbf{v}_{p+1+\widehat{(j-1)}_q}, & j \in \{1, \dots, pq\} \backslash J \end{cases}$$

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 $\widehat{(j_1 - 1)}_q = \widehat{(j_2 - 1)}_q$ 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 $\{\widehat{(j-1)}_q \ : \ j \in J\} = \{0, 1, \dots, q-1\}$. Thus $\{[\mathbf{T}'\mathbf{v}]_j \ : \ j \in J\} = \left\{ \mathbf{v}_{p+1+\widehat{(j-1)}_q} \ : \ j \in J \right\} = \{\mathbf{v}_{p+1}, \dots, \mathbf{v}_{p+q}\}$. And $\mathbf{T}'\mathbf{v} = \mathbf{0}$ if and only if $\mathbf{v}_{p+1} = \dots = \mathbf{v}_{p+q} = 0$ which implies $\{[\mathbf{T}'\mathbf{v}]_j \ : \ j \in J^c\} = \left\{ \mathbf{v}_{1+\widehat{(j-1)}_p} \ : \ j \in J^c \right\} = \{\mathbf{v}_1, \dots, \mathbf{v}_p\}$. Again $\mathbf{T}'\mathbf{v} = \mathbf{0}$ implies $\mathbf{v}_1 = \dots = \mathbf{v}_p = 0$. This contradicts that $\mathbf{v}$ is non-zero. 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 $\mathbf{w}$. ■

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

[5] 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.

[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 Research*, 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 *Proceedings of Pacific Symposium on Biocomputing*, 1997, pp. 222–229.

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

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

[10] E. N. Trifonov, "3-, 10.5-, 200- and 400-base periodicities in genome sequences," *Physica A, Elsevier Science*.

[11] S. Tiwari, S. Ramachandran, A. Bhattacharya, Bhattacharya, and R. Ramaswamy, "Prediction of probable genes by fourier analysis of genomic sequences," *Computer Applications in the Biosciences*, vol. 13, no. 3, pp. 263270, 1997.

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

[13] 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.

[14] 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.

[15] P. D. Cristea, "Genetic signal representation and analysis," in *Proceeding SPIE Conference, International Biomedical Optics Symposium (BIOS02)*, 2002, p. 7784.

[16] 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.

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

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

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

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

[21] Raman 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] Raman Arora and W. A. Sethares, "Decomposing statistical periodicities," in *IEEE workshop on Statistical Signal Processing*, Madison, Wisconsin, Aug 2007.

[23] Peter Grunwald, I.J. Myung, and M. Pitt, *Advances in Minimum Description Length: Theory and Applications*, MIT Press, 2005.

[24] Ronald Pearson, T. Zylkin, J. Schwaber, and G. Gonye, "Quantitative evaluation of clustering results using computational negative controls," in *Proc. of the SIAM International conference on Data Mining*, Lake Buena Vista, FL, 2004, pp. 188–199.

[25] UCSC Gene Sorter, [Online] http://genome.ucsc.edu/.

[26] Byung Jun Yoon and P. P. Vaidyanathan, "Computational identification and analysis of noncoding RNAs - unearthing the buried treasures in the genome," *IEEE Signal Processing Magazine*, vol. 24, no. 1, pp. 64–74, Jan 2007.

[27] 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.

[28] Evan Santo and Nevenka Dimitrova, "Improvement of spectral analysis as a genomic analysis title," in *GENSIPS*, Tuusula, Finland, June 2007.

[29] Sean R. Eddy, "Non-coding RNA genes and the modern RNA world," *Nature Reviews Genetics*, vol. 2, no. 12, pp. 919–929, December 2001.

[30] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, first edition, 1998.

[31] Raman Arora, Colin Dewey, and W. A. Sethares, "Building signatures for non-coding rnas," in *Submitted to GENSIPS*, 2008.

[32] C. R. Woese and N. R. Pace, "Probing rna structure, function and history by comparative analysis," *The RNA World*, pp. 971–983, 1993.

[33] D. Gautheret, F. Major, and R. Cedergren, "Pattern searching/allignment with RNA primary and secondary strucutures: an effective descriptor for tRNA," *Computer Applications in the Biosciences*, vol. 6, pp. 325–331, 1990.

[34] W. A. Sethares and Thomas W. Staley, "Periodicity transforms," *IEEE Transactions On Signal Processing*, vol. 47, no. 11, pp. 2953–2964, Nov 1999.

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

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