DECOMPOSITION OF SYMBOLIC SEQUENCES VIA STATISTICAL PERIODICITY

Raman Arora and William Sethares

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

ABSTRACT

Periodicities in symbolic sequences such as DNA are investigated by decomposing a maximum likelihood estimate of the probability mass function (pmf). A linear vector space on the collection of periodic pmfs is defined and the operations on the space are interpreted directly in terms of combining the draws from multiple-urn carousels. The internal structure of the decomposed sequences mimics the structure of the integers and a uniqueness theorem shows how pq-periodic symbolic sequences can be decomposed into p and q-periodic sequences when p and q are relatively prime.

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

I. INTRODUCTION

SYMBOLIC sequences are time series defined on a finite set with no algebra - the only mathematical structure is membership in the set [1]. For example, DNA sequences consist of elements A, G, C, and T, and there is no way to "add" or combine these terms. Locating (hidden) periodicities within these sequences is important because of thier correlation with genetic anomalies [2]. There are several different kinds of periodicities that may occur including homologous, eroded, and latent periodicities [3]. Homologous periodicities occur when short fragments of DNA are repeated in tandem to give periodic sequences. Imperfect or eroded periodicities [4] contain sequences of similar elements that may occur in strands of DNA due to changes in (or erosion of) nucleotides. Latent periodicities [5], [6] occur when the repeating unit is not a fixed sequence but may change in a pattern: for instance, a sequence in which the nth element is always either A or G.

Most current approaches to detecting periodicities transform the symbolic sequences into a numerical sequence [6], [7], [8]; these techniques are primarily aimed at the detection of homological periodicities. A general approach to the detection of these three classes of periodicities was presented in [9] using a maximum likelihood formulation. In this approach, each element in the DNA sequence is assumed to be generated from an information source with an underlying probability mass function (pmf). The number of sources defines the period and the symbols are drawn from these sources in a cyclic manner. Thus, statistical periodicities in the symbols are represented by repetitions of the pmfs. This can be pictured as in Fig. 1. A rotating carousel (labeled A) contains N_A urns, each with its own distribution of balls (which are labeled A, G, C, or T). At each timestep, a ball is drawn from the urn and the carousel rotates one position. The output of the process is not periodic; rather, the distribution from which the symbols are chosen is periodic. This is called *statistical periodicity* or strict sense *cyclostationarity* [10].

This paper defines the random variables on a finite set of symbols (an alphabet) and these symbols are not mapped to numerical values. This avoids imposing an arbitrary mathematical structure and implies that there is no algebra on the sequences. Instead, it is possible to define compositions of probability measures associated with the sequences. This allows a description of multiple periodic sequences analogous to the addition of periodic numerical sequences [11]. For example, one way to compose probability measures is to form a Bernoulli mixture of two biased coins with probability of Heads p and q respectively. At each step, a coin is picked randomly with probability $1/2$ and flipped. The observations can be modeled as Bernoulli random

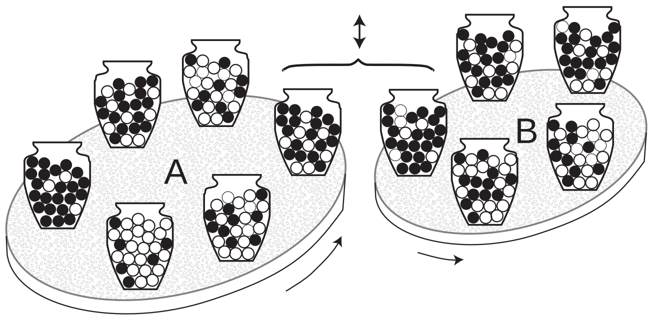

Fig. 1. Each time a ball is removed from one of the N_A urns (indicated by the arrow), platform A rotates, bringing a new urn into position. Similarly, carousel B contains N_B urns, each with its own collection of balls. Draws are made by combining draws from the two aligned urns and results in a $N_A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

variable on the set $S = \{H, T\}$ with parameter $\frac{1}{2}(p+q)$. This composition of probability measures arises naturally due to the underlying experiment and the composition does not imply any algebraic structure on the set S.

This paper investigates a method of composition that has a nice interpretation in terms of erosion or mutation in genes. The corresponding experiment is illustrated in Fig. 1 where two rotating carousels A and B contain 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 both the drawn elements have the same label, the output assumes that label. If the draws give balls with different labels, they are returned to the urns. This continues until an identical pair is drawn. The urns then rotate and the process repeats. Sect. II shows that this method of composition gives a rich mathematical structure in which to study statistical periodicities with multiple hidden periodicities. Thus the figure shows how two cyclostationary sequences with periods N_A and N_B may combine to form a new sequence with period $N_A N_B$.

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

II. Periodic Subspaces

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

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

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

where j ranges from 1 to M. Let $\mathcal{P}_p = \{S \in \mathbb{X} : S \text{ is } p\text{-statistically periodic}\}$. Then $\mathcal{P} = \bigcup_{p \in \mathbb{Z}^+} \mathcal{P}_p$ is the set of all statistically periodic sequences of random variables on the alphabet $\mathcal{A}$. Note that each $X \in P_p$ can be uniquely identified by an $M \times p$ column stochastic matrix $\mathbf{Q}^X$. Therefore $\mathcal{P}_p$ can also be identified as the set of all $M \times p$ column stochastic matrices. With a slight abuse of notation, the elements of $\mathcal{P}_p$ may be referred to as a random symbolic sequence X or as the corresponding pmf $\mathbf{Q}^X$. The law of composition on the pmfs of the random symbolic sequences that captures the experiment in Fig. 1 defines

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

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

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

Again, this is a slight abuse of notation since the binary operation is defined on the matrices $\mathbf{Q}^X, \mathbf{Q}^Y$ but is expressed in terms of the symbolic sequences X, Y. Recall that there exists no algebraic structure on the set $\mathcal{A}$ and consequently it makes no sense to directly combine realizations of symbolic sequences.

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.

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

$$\mathbf{q}_n^Z = \frac{1}{C} \begin{bmatrix} \mathbf{q}_{\hat{n}_p}^X(1)\mathbf{q}_{\hat{n}_q}^Y(1) \\ \vdots \\ \mathbf{q}_{\hat{n}_p}^X(M)\mathbf{q}_{\hat{n}_q}^Y(M) \end{bmatrix} \tag{4}$$

where $C = \sum_{j=1}^M \mathbf{q}_{\hat{n}_p}^X(j)\mathbf{q}_{\hat{n}_q}^Y(j)$ is the normalization factor. If $X = Y$, then $Z \in \mathcal{P}_p$ with

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

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

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

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

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

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

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 statistical periodicities p, q and r respectively. Let $V = X \oplus (Y \oplus Z)$ and $W = (X \oplus Y) \oplus Z$. Then

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

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

Corollary 1: $(\mathcal{P}, \oplus, \circ)$ 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 $\mathcal{P}$.

Note that a vector $X \in \mathcal{P}_p$ can be written as $X = [X_1, \dots, X_p]'$ where each $X_i \in \mathcal{P}_1$.

III. DECOMPOSING PERIODICITIES

A fundamental problem in symbolic signal processing [6] is identifying the periodic structure of the symbolic sources. Given a realization of the symbolic sequence, the maximum likelihood estimates of the statistical periodicity and the corresponding pmf were derived in [9]. This section investigates the problem of decomposing the discovered symbolic source into various smaller components.

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

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

Lemma 2: 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 $[\underbrace{-1 \dots -1}_{p}\underbrace{1 \dots 1}_{q}]$

Lemma 2 shows that if $Z \in \mathcal{P}_{pq}$ can be decomposed into $Z = X \oplus Y$ for some $X \in \mathcal{P}_p$ and $Y \in \mathcal{P}_q$, then it can also be decomposed as

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

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

IV. REFERENCES

[1] Wei Wang and Don H Johnson, "Computing linear transforms of symbolic signals," *IEEE Transactions 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] M B Chaley, E V Korotkov, and K G Skryabin, "Method revealing latent periodicity of the nucleotide sequences modified for a case of small samples," *DNA Research*, vol. 6, pp. 357 – 363, Feb. 1999.

[4] E V Korotkov and D A Phoenix, "Latent periodicity of DNA sequences of many genes," in *Proceedings of Pacific Symposium on Biocomputing 97*, R. B. Altman, A. K. Dunker, L. Hunter, and T. Klein, Eds., Singapore-New-Jersey-London, 1997, pp. 222–229, Word Scientific Press.

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

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

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

[8] 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.

[9] R. Arora and W. A. Sethares, "Detection of periodicities in gene sequences: a maximum likelihood approach," in *Proc. of the Fifth IEEE International Workshop on Genomic Signal Processing and Statistics*, Tuusula, Finland, Jun 2007.

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

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