LOCALIZING TIME-VARYING PERIODICITIES IN SYMBOLIC SEQUENCES

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

University of Wisconsin
Department of Electrical and Computer Engineering
1415 Engineering Drive, Madison, WI-53706

ABSTRACT

A novel approach is presented for the detection and localization of changing periodicities in symbolic sequences. Various symbolic sequences like DNA can be modelled as stochastic processes that exhibit time-varying cyclostationarity. The coding regions of the DNA, for instance, exhibit statistical periodicity with period three. The complexity-regularized maximum-likelihood estimates are developed in this paper for the statistical period of symbolic sequences. The changing periodicities along the sequence are discovered by using sliding windows. A cumulative sum test is also presented to detect the change points. The formulation in this paper avoids any kind of numerical mapping for the symbolic DNA sequences and does not impose any algebraic structure.

Index Terms— Symbolic periodicity, finding exons, cyclostationarity.

I. INTRODUCTION

SYMBOLIC sequences are time series defined on a finite set with no algebra. In DNA sequences, economic indicator data, and other nominal time series, the only mathematical structure is the set membership [1]. An interesting and important behaviour such symbolic sequences may exhibit is *periodicity* and finding such periodicities is fundamental to the understanding and determination of 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 α-helical formation for synthesized protein molecules [2] and a base periodicity of 3 is identified with exons, the protein coding regions of the DNA. Such investigations also find application in diagnosis of genetic disorders (like Huntington's disease [3]), DNA forensics and reconstructing evolution history [4].

Symbolic periodicities can be classified into homologous, eroded and latent. Homologous periodicities occur when short fragments are repeated in tandem. Eroded periodicities [5] result when some of the symbols in a homologous periodic sequence get replaced or altered so that the tandem repeats are imperfect. These may also be observed as *indels* (insertions and deletions) in homologous periodic sequences. Latent periodicities [5] occur when the repeating unit is not fixed but may change in a patterned way: for instance, a sequence in which the nth element is always either A or G. An observed latent period of nucleotides in a DNA sequence may be (A/C)(T/G)(T/A)(G/T)(C/G/A)(G/A), i.e. the first nucleotide of a period may be A or C followed by a T or G and so on.

Symbolic random variables take values on a set called the *alphabet* whose elements are called *symbols*. A symbolic sequence is defined as a sequence of symbolic random variables. Most current approaches for detecting periodicities transform the symbolic sequences into numerical sequences which defines an algebra on the alphabet [6]. This 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 [7] puts a total order on the set; the complex representation (A$=1+j$, G$=-1+j$, C$=-1-j$, T$=1-j$) used in [8], [6] implies that the euclidean distance between A and C is greater than the distance between A and T. Artifacts of such mappings are reported in [9]. A good survey of various numerical representations for DNA sequences is presented in [10]. Most of these techniques are primarily aimed at the detection of homological periodicities [11], [8], [9].

In contrast, the formulation in this paper implies no mathematical structure on the alphabet and presents a general approach to the detection of the three classes of periodicities in a maximum likelihood framework. Each symbol of the sequence is assumed to be generated by an information source with an underlying probability mass function (pmf). 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, referred to as *statistical periodicity* or *strict sense cyclostationarity*. The number of sources is equal to the latent period in the sequence.

The problem of detecting latent periodicities is formulated mathematically in the next section. The maximum likelihood estimates of the period and the distributions of the sources were developed in [12]. The estimates are improved in this

paper by incorporating a complexity term with the likelihood function in Section III. This penalized maximum likelihood estimator is justified by the application of the minimum description length (MDL) principle to the model selection problem. In Section IV the MDL estimates are computed in sliding windows over various simulated and real DNA sequences. The series of estimates characterizes the time-varying behaviour of the sequences.

II. STATISTICAL PERIODICITY

A given symbolic sequence $D = D_1 D_2 \dots D_N$ of length N can be denoted by a mapping from natural numbers to an alphabet $\mathcal{X}$. For DNA sequences $\mathcal{X} = \{\mathrm{A}, \mathrm{G}, \mathrm{C}, \mathrm{T}\}$, where the symbols denote the nucleotides Adenine, Guanine, Cytosine and Thymine respectively. Let X be an $\mathcal{X}$-valued random variable (or information source) with probability distribution P. Let $\mathcal{X}^n$ denote the n-fold cartesian product of $\mathcal{X}$ and $x^n \in \mathcal{X}^n$ denote a sequence of length n. A *probabilistic source* is defined as a sequence of probability distributions $P^{(1)}, P^{(2)}, \dots$ on corresponding sequence of alphabets $\mathcal{X}^1, \mathcal{X}^2, \dots$ 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 repeated concatenation of realizations of a probabilistic source $P^{(\mathcal{T})}$ the statistical period of D is defined to be $\mathcal{T}$. In other words, a $\mathcal{T}$-periodic cyclostationary sequence D is generated by $\mathcal{T}$ information sources, $X_1, \dots, X_{\mathcal{T}}$, 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 $P_i(j) = P(X_i = \mathcal{X}_j)$ for $j = 1, \dots, |\mathcal{X}|$ where $|\mathcal{X}|$ is the cardinality of the alphabet ($|\mathcal{X}| = 4$ for the DNA sequences).

Let $\mathcal{T}$ denote the true period and k denote the hypothesized period. The number of complete statistical periods in an N-symbol long k-periodic cyclostationary sequence D are $M = \lfloor N/k \rfloor$, where $\lfloor x \rfloor$ denotes the largest integer less than or equal to x. Define $\widehat{i}_k = 1 + ((i-1) \mod k)$, where $(x \mod y)$ denotes the remainder after division of x by y. Then for $1 \le i \le N$, the symbol D_i, i.e. the i^{th} symbol in the sequence D, is generated by the random variable $X_{\widehat{i}_k}$. The random variables $X_{\widehat{i}_k}$ for $\widehat{i}_k = 1, \dots, k$ are assumed to be independent. The parameters, period k and pmfs $P_1, \dots, P_k$ of corresponding information sources, are unknown. The search space for parameter k is the set $B = \{1, \dots, N_0\}$, for some $N_0 < N$ and for the pmfs $\mathbf{Q}^{(k)} = [P_1, \dots, P_k]$ the search space is the subset $\mathcal{Q}^{(k)} \subseteq [0,1]^{|\mathcal{X}| \times k}$ of column stochastic matrices (for $\mathbf{Q} \in \mathcal{Q}^{(k)}$, $\mathbf{Q}_{ji} \in [0,1]$ and $\sum_{j=1}^{|\mathcal{X}|} \mathbf{Q}_{ji} = 1$ for $i = 1, \dots, k$). Conditioned on k, the maximum likelihood estimate of $\mathbf{Q}^{(k)}$ is given as

$$\mathbf{Q}^{(k)}_{\mathrm{ML}} = \arg \max_{\mathbf{Q} \in \mathcal{Q}^{(k)}} P(D|\mathbf{Q}). \tag{1}$$

The plug-in MLE for the statistical period can be written as

$$\mathcal{T}_{\mathrm{ML}} = \arg \max_{k \in B} P(D|\mathbf{Q}^{(k)}_{\mathrm{ML}}). \tag{2}$$

The maximum likelihood estimates $\mathbf{Q}^{(k)}_{\mathrm{ML}}$ and $\mathcal{T}_{\mathrm{ML}}$ were developed in [12]. However, as seen from the experimental results on simulated sequences and real gene data the estimates tend to overfit the data. To address the problem of over-fitting, a penalized maximum likelihood estimator is presented in Section III. The estimator is derived using the refined minimum description length (MDL) principle. The penalization corresponds to assuming the universal prior on the parameters and refined MDL estimator is essentially the MAP estimator with respect to the universal prior.

III. PENALIZED MAXIMUM LIKELIHOOD ESTIMATOR

The fundamental idea behind MDL is that more regular the data is, the easier it is to compress and learn [13]. As in the previous section let D denote the data and let $\mathcal{Q}^{(1)}, \mathcal{Q}^{(2)}, \dots, \mathcal{Q}^{(N_0)}$ be the list of class of models or hypotheses. Define $\mathcal{H} = \cup_{k=1}^{N_0} \mathcal{Q}^{(k)}$. Then the best explanation of the data D is the hypothesis $H \in \mathcal{H}$ that minimizes the description length

$$L(D|\mathcal{H}) = L(\mathcal{Q}^{(k)}) + L(D|\mathbf{Q}^{(k)}_{\mathrm{ML}}), \tag{3}$$

where $L(\mathcal{Q}^{(k)})$ is the length, in bits, of the description of the model class $\mathcal{Q}^{(k)}$ and $L(D|\mathbf{Q}^{(k)}_{\mathrm{ML}})$ is the length of the description of the data by the best fitting model in the class $\mathcal{Q}^{(k)}$. The term $L(D|\mathcal{H})$ is sometimes referred to as the *stochastic complexity* of the data given the model whereas $L(\mathcal{Q}^{(k)})$ is called the *parametric complexity*. Clearly, the MDL model selection involves the trade-off between goodness-of-fit and complexity.

The second term $L(D|\mathbf{Q}^{(k)}_{\mathrm{ML}})$ in the two part code, represents the codelength of the data when encoded with the hypothesis $\mathbf{Q}^{(k)}_{\mathrm{ML}}$. Assuming the hypotheses are probabilistic, the Shannon-Fano code is optimal in terms of the expected codelength. Thus, $L(D|\mathbf{Q}^{(k)}_{\mathrm{ML}}) = -\log P(D|\mathbf{Q}^{(k)}_{\mathrm{ML}})$, where $P(D|\mathbf{Q}^{(k)}_{\mathrm{ML}})$ is the probability of observing D given the model $\mathbf{Q}^{(k)}_{\mathrm{ML}}$. The codelength is therefore the negative-log-likelihood of observing the data D. As derived in [12], the $(j, \widehat{i}_k)^{th}$ element of the matrix $\mathbf{Q}^{(k)}_{\mathrm{ML}}$ is given as

$$\mathbf{Q}^{(k)}_{\mathrm{ML}}(j, \widehat{i}_k) = \frac{1}{L} \sum_{l=1}^{L} \mathbf{1}\{D_{(l-1)k + \widehat{i}_k} = \mathcal{X}_j\}, \tag{4}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function, $\widehat{i}_k = 1, \dots, k$ and $L = M$ if $(Mk + \widehat{i}_k) > N$, $L = (M+1)$ otherwise.

The MLE for the probability mass functions of the random variables is quite intuitive. Simply stated, given the period k, it amounts to segmentation of the data sequence into k non-overlapping de-interleaved subsequences. And the pmf of the

m^{th} information source is given by the relative frequencies of each symbol. For instance, if the hypothesized statistical period in a gene sequence is 3 then the MLE of the pmf of the 2^{nd} information source is given by the empirical probabilities of nucleotides in the subsequence comprising of every third symbol, starting with the second.

For the first term in equation (3), following code may be adopted. 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 the binary representation of k. Note that this a prefix code and takes $2\lceil \log k \rceil + 1$ bits. The parameters of $\mathbf{Q} \in \mathcal{Q}^{(k)}$ are described by $k' = k \cdot |\mathcal{X}|$ 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

$$L(H)+L(D|H) = 2\lceil \log k \rceil + 1 + k|\mathcal{X}| \log \lceil \frac{N}{k} \rceil - \log P(D|H) \tag{5}$$

for $H \in \mathcal{H}$. It is clear from the equation above that the MDL principle yields a penalized maximum likelihood estimate. The code used here is a *universal code* and implies a universal prior on the hypothesis.

IV. TIME VARYING PERIODICITIES

The penalized MLE is applied to various simulated symbolic sequences and real gene sequences. In order to detect time-varying periodicities in a sequence of N symbols, the estimates are computed in a sliding window of size $N_w \ll N$ with an overlap of N_c symbols between successive windows. Figure 1 shows results for a simulated 8000-symbols long DNA sequence that has latent periodicity of period 6 for subsequences with indices $1-2000$ and $6001-8000$ and is uniformly random in the middle. Thus there are two *change points* in the sequence. The latent period of the periodic part of the sequence is (A/C)(T/G)(T/A)(G/T)(C/G/A)(G/A), i.e. it is generated by six information sources, $X_1, \ldots, X_6$ with X_1 generating A or C each with equal probability, X_5 generating A, G or C each with probability $1/3$ and so on. 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). Both change points are seen in the surface plot. Also the six-periodic behaviour is evident from the plot as are the subharmonics, the integer multiples of the true period.

The algorithm was also tested with chromosome 20 of the human genome [14]. The 9748 base-pair (bp) long sequence (bp: 22,553,000-22,562,747) contains 1305 bp 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 2 shows a latent periodicity of period three beginning at sliding window number 60 which corresponds to bp number 22,557,427 ($N_w = 750$, $N_c = 675$). The period-3 behaviour of protein coding genes is expected

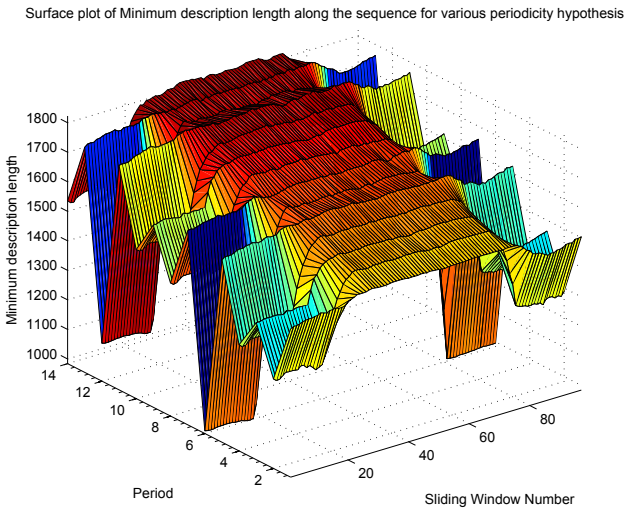

Fig. 1. Description length (in bits) for the ML estimate in $\mathcal{Q}^{(k)}$ plotted against period k along the sequence.

since amino acids are coded by trinucleotide units called *codons* [6].

The window size N_w determines the usual trade-off between the resolution and accuracy of the estimates. The larger the window size, the better the estimates since averaging in the empirical estimator is 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. A 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 1 is much smaller than the window size, the change points may go undetected. A multi-resolution multi-scale

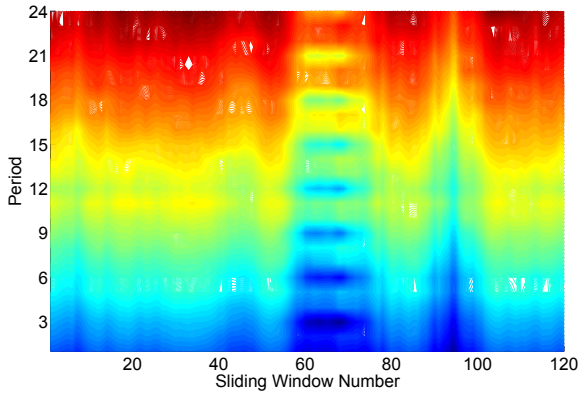

Fig. 2. Contour plot of description length (in bits) for the ML estimate in $\mathcal{Q}^{(k)}$ plotted against period k along the sequence.

technique is therefore 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.

As observed from the plots the periodicity profile of the sequences changes gradually near the change points whereas in other parts the profile remains constant except for small fluctuations due to noisy data. Thus, a statistical test based on the positive inflection rate over multiple successive windows can be constructed. Given $\mathcal{T}_{\text{ML}} = k$, the alternative composite hypothesis is that the period is no longer k. The null hypothesis that there is no change is rejected if

$$\mathbf{\Theta}_t^{(k)} = \min_{m \in \{1,\ldots,T\}} |\mathbf{Q}_{\text{ML},t}^{(k)} - \mathbf{Q}_{\text{ML},t-m}^{(k)}|_{\text{tot}} \; > \; \delta_{\text{Th}} \qquad (6)$$

where $|\mathbf{A} - \mathbf{B}|_{\text{tot}} = \sum_{i,j} (\mathbf{A}(i,j) - \mathbf{B}(i,j))^2$ is the total deviation between matrices $\mathbf{A}$ and $\mathbf{B}$, δ_{Th} is a threshold and T is the number of successive windows over which the test is conducted. The test statistic $\mathbf{\Theta}_t^{(k)}$ for period k is the minimum total deviation between ML estimates for the pmfs in window t and previous T windows. The formulation in (6) is similar to the change-point problem in statistics and the test proposed here is based on the cumulative sum approach. $\mathbf{\Theta}_t^{(k)}$ is plotted in Figure 3 for the simulated latent periodic sequence used in Figure 1. The jump in $\mathbf{\Theta}_t^{(6)}$ at $t = 9$ corresponds to the change-point at bp number $N_w + 8 \times (N_w - N_c) = 1950$, giving much better resolution. The resolution can be further improved upon by increasing N_c, keeping N_w constant. Note that $\mathbf{\Theta}_t^{(6)}$ is consistently large over transition regions with lobe-width equal to N_w.

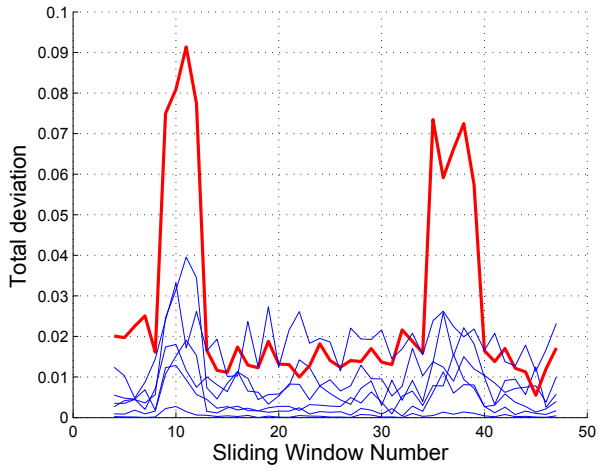

Fig. 3. $\mathbf{\Theta}_t^{(k)}$ plotted for the sequence from Figure 1. $\mathbf{\Theta}_t^{(6)}$ is plotted in red ($N_w = 750$, $N_c = 600$, $T = 3$).

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 is challenging at least in part due to the lack of an algebraic structure. The approach in this paper is to model symbolic sequences as nonstationary random processes on a finite alphabet. The time-varying nature of symbolic sequences is studied and a uniformly most powerful test is constructed for detecting the transition points.

VI. REFERENCES

[1] Wei Wang and Don H. Johnson, "Computing linear transforms of symbolic signals," *IEEE Trans. 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, Mar 1993.

[4] C. M. Hearne, S. Ghosh, and J. A. Todd, "Microsatellites for linkage analysis of genetic traits," *Trends in Genetics*, vol. 8, pp. 288–294, 1992.

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

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

[7] P. D. Cristea, "Genetic signal representation and analysis," in *Proc. SPIE Conf.*, 2002, p. 77 84.

[8] M. Buchner and S. Janjarasjitt, "Detection and visualization of tandem repeats in DNA sequences," *IEEE Trans. Sig. Proc.*, vol. 51, pp. 2280–2287, Sep 2003.

[9] 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} *EUSIPCO*, Sep 2006.

[10] M. Akhtar, J. Epps, and E. Ambikairajah, "On DNA numerical representations for period-3 based exon prediction," in *GENSIPS*, Tuusula, Finland, June 2007.

[11] 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.

[12] Raman Arora and W. A. Sethares, "Detection of periodicities in gene sequences: a maximum likelihood approach," in *GENSIPS*, Tuusula, Finland, June 2007.

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

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