EUSIPCO 2013 1569741807

LIKELIHOOD-BASED ESTIMATION OF PERIODICITIES IN SYMBOLIC SEQUENCES

Stefan Ingi Adalbjörnsson, Johan Swärd, and Andreas Jakobsson

Dept. of Mathematical Statistics, Lund University, Sweden

ABSTRACT

In this work, we propose a method of estimating periodicities in symbolic sequences, allowing for arbitrary, finite, symbol sets. Different from other common approaches, that often map the symbolic sequence to a numerical representation, we here exploit a likelihood-based formulation to represent the periodic behavior of the sequence. The performance of the proposed method is illustrated on both simulated and real DNA data, showing a notable performance gain as compared to other common estimators.

Index Terms— Periodicity, Symbolic sequences, Spectral estimation, Data analysis, DNA

1. INTRODUCTION

Symbolic data sequences consisting of elements from a finite set or alphabet, often not exhibiting any natural ordering, is a common occurrence in a wide range of fields, including text indicators, genomic data, and different forms of categorical time series analysis (see, e.g., [1]). Commonly, one has an interest in determining periodicities in such sequences, for instance in order to determine the latent periodicities in DNA sequences, which have been shown to be correlated with various forms of functional roles being of importance in DNA analysis (see, e.g., [2–8]). Due to the lack of algebraic structure in symbolic sequences, traditional spectral estimation algorithms are not well suited to determine such periodicities, as these generally exploits the natural ordering among the symbols. To alleviate this problem, several forms of mappings from symbols to numerical representations have been considered in the literature, for instance using PAM- or QPSK-based mappings, minimum entropy mapping, mapping equivalences, transformations, or maximum likelihood formulations of the cyclostationary properties of the periodicities (again, see, e.g., [2–8]). Generally, these methods have relatively high computational complexity and/or suffers from difficulties of expanding the methods to larger alphabets. After such a mapping, the periodicities are commonly determined using a periodogram estimate, although such an approach will suffer from the well-known high variability and/or poor resolution of the method [9]. Recently, we proposed a symbolic periodicity estimator taking the naturally occurring harmonic structure into account, using a MUSIC-like formulation to estimate these [10]. Exploiting this additional structure was there shown to offer preferable performance, especially for the detection of longer periodicities. In this work, we instead examine the problem using a probabilistic model for the symbolic sequence, thereby allowing for a likelihood-based hypothesis testing formulation. As the resulting estimates can be calculated using analytical expressions, the computational complexity is linear in the length of the symbolic sequence and the number of symbols. The performance of the proposed periodicity estimator is illustrated using both simulated sequences and real DNA measurements, showing a remarkable performance gain as compared to earlier methods, in particular for sequences containing more than one periodicity per symbol.

This work was supported in part by the Swedish Research Council and Carl Trygger's foundation.

2. LIKELIHOOD-BASED SYMBOLIC ESTIMATION

Consider a symbolic sequence, s_k, for $k = 1, \ldots, N$, formed from a set, or alphabet, $\mathcal{A}$, having a finite cardinality $|\mathcal{A}| = B$. Assuming that the symbols in the sequence are independent and identically distributed, such that

$$p_j \triangleq \text{Prob}(s_k = \mathcal{A}_j) \tag{1}$$

for $k = 1, \ldots, N$ and $j = 1, \ldots, B$, with non-negative probabilities, p_k, summing to unity over the B symbols, and with $\mathcal{A}_j$ denoting the jth symbol in $\mathcal{A}$, implying that the probability mass function (PMF) is given as

$$\begin{aligned} p_0(\mathbf{x}_N|\mathbf{p}_B) &\triangleq \text{Prob}(\mathbf{s}_N = \mathbf{x}_N) && (2)\\ &= \prod_{j=1}^{N}\prod_{\ell=1}^{B} p_\ell^{[x_j=\mathcal{A}_\ell]} = \prod_{k=1}^{B} p_k^{G_k} && (3)\end{aligned}$$

where $[\cdot]$ denotes the Iversons bracket, which equals one if the statement inside the brackets is true and zero otherwise, and with $\mathbf{p}_B$ and $\mathbf{s}_N$ denoting the vector of probabilities and the symbolic sequence, respectively, i.e.,

$$\begin{aligned} \mathbf{p}_B &= \begin{bmatrix} p_1 & \ldots & p_B \end{bmatrix}^T && (4)\\ \mathbf{s}_N &= \begin{bmatrix} s_1 & \ldots & s_N \end{bmatrix}^T && (5)\end{aligned}$$

with $(\cdot)^T$ denoting the transpose, and where $\mathbf{x}_N$ is a symbolic sequence with elements x_ℓ, for $\ell = 1, \ldots, N$, with each of

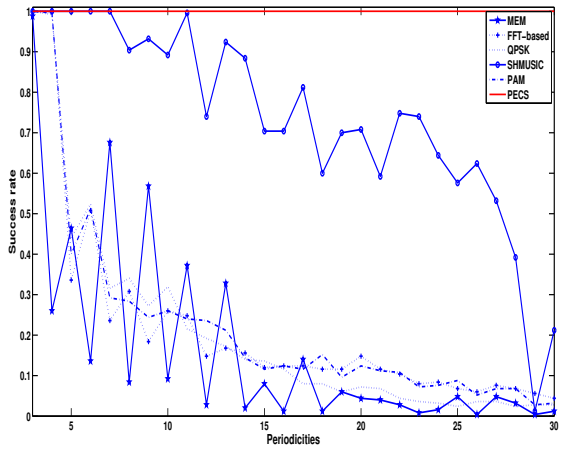

Fig. 1. Rate of success in estimating perfect periods.

the symbols appearing G_k times, for $k = 1, \ldots, B$. As a result, the PMF is a function depending only on the number of times each symbol appears, and on the probability given to each symbol, and not on any vector or numerical value representing the symbols as such. In general, the probabilities, p_k, are unknown and need to be estimated from the observed sequence. This can be done using the maximum likelihood (ML) estimate formed as

$$p_j^* = \frac{G_j}{N} \tag{6}$$

for $j = 1, \ldots, B$, which is an unbiased and asymptotically efficient estimate [11, p. 475]. Furthermore, note that a symbol $\alpha \in \mathcal{A}$, occurring with periodicity m, i.e., with the symbol appearing at every mth index in the sequence, implies that all elements of the sequence should be equal to the symbol α in one of the m possible (disjoint) index sets

$$I_{m,\ell} = \left\{ \ell, \ell + m, \ldots, \ell + \left\lfloor \frac{N - \ell}{m} \right\rfloor m \right\} \tag{7}$$

for all offsets $\ell \in \{1, \ldots, m\}$, where $\lfloor \cdot \rfloor$ denotes the rounding down operation. This means that if a periodicity m is present in a sequence, the sequence is clearly also periodic for every mr, for all natural numbers r; to avoid ambiguity, we here refer to the period as the lowest possible such periodicity. Considering a sequence, $\mathbf{s}_N$, with a periodicity m in the symbol α, with offset n, this implies that all the symbols in the sequence at index p, for $p \in I_{m,n}$, will equal α. Thus, it is a deterministic and not a statistical problem to determine if such a (perfect) periodicity is present; to do so, it is sufficient to determine if any of the m symbolic subsets, $\{s_k\}$, for $k \in I_{m,n}$, is such that it is formed from only a single symbol, α. Such a test may be formed in $N + m$ operations. However, many forms of symbolic sequences, such as, for example, DNA sequences, contain also non-perfect periodicities, such that the sequence may contain the periodicity over only a limited interval, and/or with some of the periodically occurring symbols being replaced by some other symbols, which may occur, for example, due to the presence of measurement noise, coding errors, or some, perhaps unknown, functional equivalence between symbols. In such cases, the PMF for a symbolic sequence with a given periodicity m and offset n will instead be formed from one categorical distribution for indexes $I_{m,n}$ as well as from another categorical distribution for all index in the complementary set, $I_{m,n}^c$, i.e., for all those indices not in $I_{m,n}$. Thus, in this case, the PMF is given as

$$\begin{aligned} p_1(\mathbf{x}_N | \mathbf{p}_B, \tilde{\mathbf{p}}_B) &\triangleq \text{Prob}(\mathbf{s}_N = \mathbf{x}_N) && (8) \\ &= \prod_{j \in I_{m,n}} \prod_{\ell=1}^{B} \tilde{p}_\ell^{[x_j = \mathcal{A}_\ell]} \prod_{j \in I_{m,n}^c} \prod_{\ell=1}^{B} p_\ell^{[x_j = \mathcal{A}_\ell]} \\ &= \prod_{\ell=1}^{B} \tilde{p}_\ell^{\tilde{G}_\ell} \prod_{k=1}^{B} p_k^{G_k} && (9) \end{aligned}$$

where $\tilde{\mathbf{p}}_B$ is a parameter vector containing the probabilities $\tilde{p}_k$, for $k = 1, \ldots, B$, denoting the probability of a symbol, $\mathcal{A}_k$, occurring in the index set $I_{m,n}^c$, and with G_k and $\tilde{G}_k$ denoting the number of times the symbol $\mathcal{A}_k$ occurs in the set $I_{m,n}$ and its complement, respectively. The corresponding ML estimates are found as

$$p_j^* = \frac{G_j}{|I_{m,n}|} \tag{10}$$

$$\tilde{p}_j^* = \frac{\tilde{G}_j}{|I_{m,n}^c|} \tag{11}$$

for $j = 1, \ldots, B$. As a result, one may form a test to determine the hypothesis that a given sequence has a different distribution for the indexes corresponding $I_{m,\ell}$, i.e., the PMF is formed using (9), against the null hypothesis, i.e., that the entire sequence has the same categorical distribution, such that the PMF follows (3), i.e.,

$$\text{H}_0 : \mathbf{p}_B = \tilde{\mathbf{p}}_B \tag{12}$$

$$\text{H}_1 : \mathbf{p}_B \neq \tilde{\mathbf{p}}_B \tag{13}$$

Such a test may be formed as the likelihood ratio test [12]

$$\lambda_{m,\ell}(\mathbf{x}_N) = \frac{p_0(\mathbf{x}_N | \mathbf{p}_B, \text{H}_0)}{p_1(\mathbf{x} | \mathbf{p}_B, \tilde{\mathbf{p}}_B, \text{H}_1)} \tag{14}$$

where the probabilities are determined using (6) under H_0, and using (10) and (11) under H_1. Then, if H_0 is true, one can show that [12, p. 489], as $N \to \infty$

$$-2 \log(\lambda_{m,\ell}(\mathbf{x}_N)) \xrightarrow{d} \chi_{B-1}^2 \tag{15}$$

where $\xrightarrow{d}$ denotes convergence in distribution and χ_k^2 denotes the chi-squared distribution with k degrees of freedom.

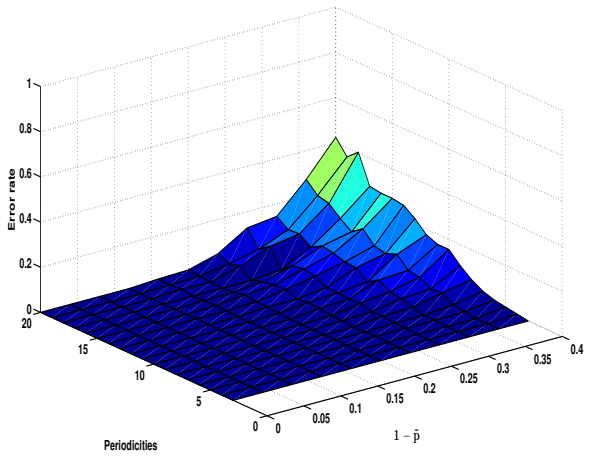

Fig. 2. The error rate of finding the periodicity as a function of $1-\tilde{p}$, and the periodicity for the proposed method.

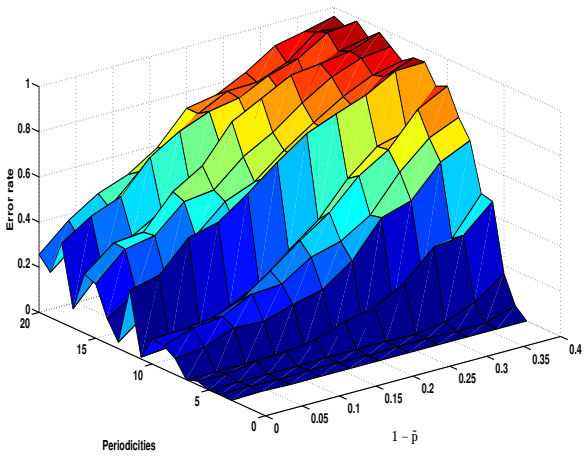

Fig. 3. The error rate of finding the periodicity as a function of the negative probability, $1-\tilde{p}$, and the periodicity for the SPE algorithm.

Should the main interest be in detecting periodicities for a particular symbol, say $\mathcal{A}_j$, then the test should be restricted to test if $p_j \neq \tilde{p}_j$, with all other probabilities being treated as nuisance parameters; for this case, the likelihood ratio test is reformulated such that it only depends on the probability p_j, for the symbol of interest, and on $1-p_j$, for all the other symbols. Then, the ratio test in (14) equals the ratio of PMFs associated with the corresponding two symbols, i.e., the Bernoulli distribution. As a result, the framework allows for flexibility in what is deemed a periodicity, e.g., one might test for a high probability of a certain symbol appearing, or even if the symbols appears with low probability. Both of these ideas will be explored further in the following, where we outline some possible algorithms for estimating periodicities for some commonly occurring situations, namely, estimation of an unknown periodicity, detection of an unknown periodicity, and, finally, estimation of multiple periodicities. In the first case, one may choose the largest likelihood ratio for every symbol, and for every combination of $m \in \{1,\ldots,m_{max}\}$ and $\ell \in \{1,\ldots,m\}$, i.e., select the desired periodicity as the period corresponding to

$$\arg\max_{m,\ell,i} \quad \lambda_{m,\ell}(f_i(\mathbf{x}_N)) \tag{16}$$

where $f_i(\cdot)$ maps the sequence to the appropriate Bernoulli sequence, as described above, and with m_{max} denoting the maximally considered periodicity. The computational cost for such a test is NBm_{max}. To also allow for cases when no periodicity is present, one needs to formulate some lower limit for the likelihood ratio, below which no periodicity is deemed present. A simple way to do so is to exploit, as shown in (15), that, asymptotically, the likelihood ratio for each of the tests will be χ^2 distributed with one degree of freedom, thus since $m_{max}^2/2$ tests are formed in order to compute (16), and if assuming that these are independent, the lower limit may be approximated forming it as some quantile of the random variable

$$\psi = max\left(z_1,\ldots,z_{m_{max}^2/2}\right) \tag{17}$$

where each z_k is χ^2 distributed. The resulting bound may thus easily be formed using Monte Carlo simulations. In the case when multiple periodicities may be present for each symbol, one can extend the proposed estimation procedure using a step-wise approach. Without loss of generality, to simplify the presentation, we here only consider the case of binary symbols, but note that the derivation for $B>2$ follows similarly. For binary symbols, the initial step of the above detailed algorithm will yield an index set $I^{(1)}_{m_1,\ell_1}$, where m_1 and ℓ_1 denote the periodicity and phase, respectively, found in the maximization of (16). Then, in order to determine the next periodicity, the H_0 distribution is formed from (9), using the found index set $I^{(2)}_{m,n} = I_{m_1,\ell_1}$, whereafter the second phase, m_2, and periodicity, ℓ_2 may be determined using (16). This procedure can then be repeated, in the kth step forming the H_0 distribution from (9), using the index set

$$I^{(k)}_{m,n} = I^{(k-1)}_{m,n} \cup I_{m_k,\ell_k} \tag{18}$$

To ensure that at each step one adds to the H_0 hypothesis only sets which are likely to have the symbol appearing, we restrict the H_1 hypothesis such that, the likelihood is maximized over $\tilde{p} \in [0.5, 1]$. The resulting algorithm, here termed the *P*eriodic *E*stimation of *C*ategorical *S*equences (PECS) estimator, is outlined in Algorithm 1 below, with each step in the iteration requiring about $m_{max}N$ operations.

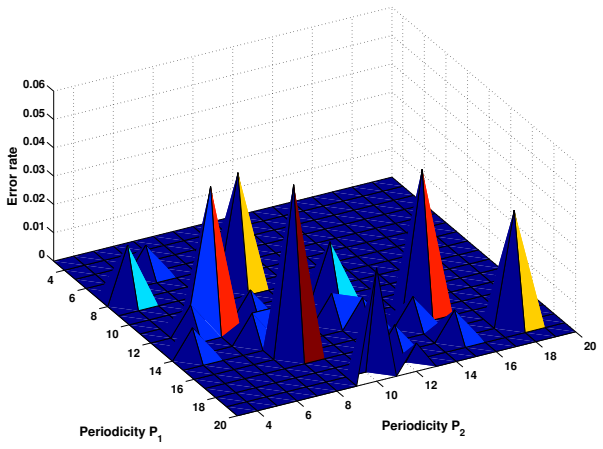

Fig. 4. The error rate of finding two periodicities in the same symbol. The error rate is greatest when two periodicities with the same length are present in the sequence.

3. NUMERICAL RESULTS

We proceed to examine the performance of the proposed likelihood-based estimator using simulated DNA sequences, binary sequences, and measured DNA data. For DNA sequences, only $B = 4$ different symbols are present, namely A, C, G, and T. Initially, we examine a simulated DNA sequence containing one perfect periodicity. Figure 1 illustrates the rate of successfully determining this periodicity as a function of the length of the periodicity, comparing the proposed PECS estimator with the MEM [8], PAM [5], QSPK [3], and SPE [10] estimators, as well as with a Fourier-based estimator detailed in [10]. Here, and in the following, the success rate has been determined using 250 Monte-Carlo simulations using $N = 1000$ equiprobable symbols, with the sought periodicity being inserted appropriately. As is clear from from the figure, the proposed PECS estimator succeeds in successfully determining all the considered periodicities, whereas all the other methods will lose performance notably as the length of the periodicity grows. Proceeding to examine also non-perfect periodicities, we vary $\tilde{p}$ for the index set corresponding to the generated periodicity, with $p = 1/4$ on the complement set. Figures 2 and 3 show the resulting success rate for the PECS and SPE estimators as a function of the periodicity and the probability $\tilde{p}$, again clearly illustrating how PECS outperform SPE (and, similarly, all the other mentioned estimators) for all periodicities and $\tilde{p}$. Next, we investigate how well PECS is able to resolves two periodicities in a binary sequence. In this case, some care needs to be taken when setting up the simulations, as when generating two periodicities, these may overlap or combine to create a new periodicity, e.g., if generating two periodicities of period

Algorithm 1 Periodic Estimation of Categorical Sequences

1: Given a binary sequence, $\mathbf{x}$ of length N
2: $I_{m_0,n_0} = \{\emptyset\}$
3: **for** k = 1 **to** $\max_{iteration}$ **do**
4: $\quad \{m_k, \ell_k\} = \arg\max_{m,\ell} \quad \lambda_{m,\ell}(\mathbf{x}_N)$
5: $\quad I_{m_k,n_k} = I_{m_{k-1},n_{k-1}} \cup \tilde{I}_{m_k,\ell_k}$
6: $\quad H_0$ distribution is replaced with (9) and I_{m_k,n_k}
7: **end for**

six, these may be placed such that they instead form just a single periodicity with period three. Similarly, two periodicities with period four and twelve may cause the resulting sequence to have only a single periodicity of four. In order to avoid ambiguities in the resulting performance measure, the test data has been generated such that it avoids such problems. Figure 4 illustrates the success rate of determining both periodicities correctly, as a function of the length of the two periodicities, with $N = 500$ and again using $\tilde{p} = 3/4$ and $p = 1/4$. As is clear from the figure, even when the sequence contains two periodicities of lengths up to 20, when most of the other discussed estimators completely fail to find even a single perfect periodicities, PECS has only 0.06 as its maximum error rate. Finally, we examine the performance of the PECS estimator on measured genomic data, in the form of the gene C. elegans F56F11.4 [13]. Since genomic data is generally not stationary, the estimate has been formed using a sliding window with length $N = 360$. The results yielded by PECS are shown in Figure 5, where the periodicities with a likelihood ratio greater than the 95% quantile of the maximum of $m^2_{max}/2 = 450$ χ^2 distributed random variables, are shown for each symbol. Figure 6 shows the corresponding $\tilde{p}$. In earlier work, such as [8] and [10], a period of three was found at around index 7000. Such a period of three was also found when using PECS, but when looking at the corresponding $\tilde{p}$, one may note that this periodicity is actually constituted by the lack of the symbol C, i.e., this period is detected since the symbols A, G, and T are alternating in a non-periodic fashion, and since C is always absent at these indexes, this apparently causes the Fourier based methods to indicate a periodicity of three. If one is not interested in finding these sorts of periodicities, one may restrict $\tilde{p}$ to be in $[1/2, 1]$, in the same manner as mentioned above. This will ensure that PECS only finds periodicities that are made up by an increased probability in the presence of a symbol.

4. ACKNOWLEDGEMENT

The authors would like to thank Prof. Lorenzo Galleani and Dr. Roberto Garello at Politecnico di Torino, Italy, for providing us with the their implementation of MEM-algorithm detailed in [8].

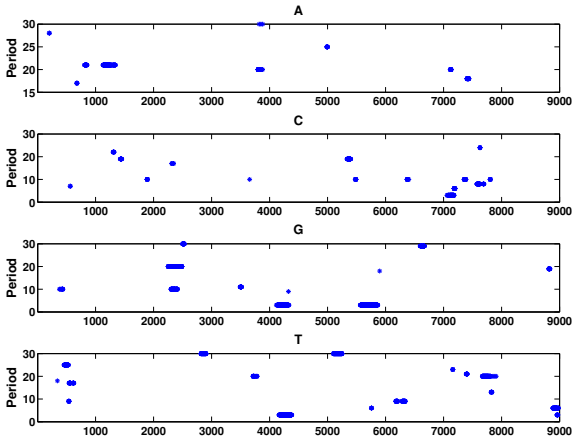

Fig. 5. The periodicities of each symbol in the gene C.elegans F56F11.4 computed using a sliding window.

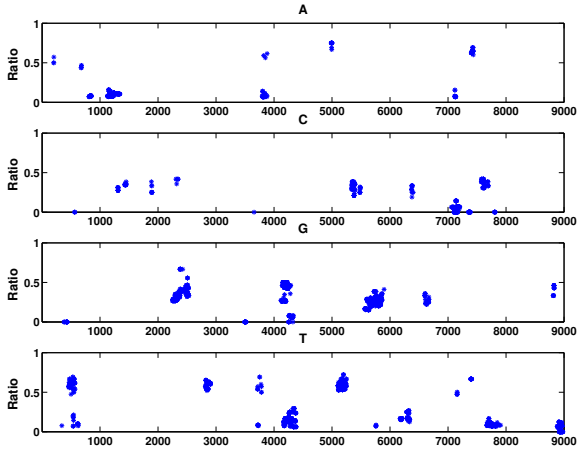

Fig. 6. The corresponding $\tilde{p}$ for each periodicity shown in Figure 5, with some of the found periodicities have a $\tilde{p}$ equal to zero, indicating that the periodicity is based on the absence of that symbol.

5. CONCLUSION

In this work, we have presented a likelihood-based symbolic periodicity estimation technique which has been shown to offer high quality estimates of symbolic periodicities as well as being able to locate periodic changes in the distribution function. The estimates also offers additional insight for symbolic sequences, as might be observed when there is a higher or lower probability of a certain symbol at some periodic indices, as compared to the rest of the sequence. For the case of DNA data, we have shown that previously proposed methods, mapping the symbolic vector to a numerical representation, have a lower success rate of finding periodicities, as well as show artifacts in their frequency estimates, likely as a result of the heuristic symbolic mappings.

6. REFERENCES

[1] Alan Agresti, *Categorical Data Analysis*, John Wiley & Sons, second edition, 2007.

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

[3] D. Anastassiou, "Genomic Signal Processing," *IEEE Signal Processing Magazine*, vol. 18, no. 4, pp. 8–20, July 2001.

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

[5] G. L. Rosen, *Signal Processing for Biologically-Inspired Gradient Source Localization and DNA Sequence Analysis,*, Ph.D. thesis, Georgia Institute of Technology, 2006.

[6] R. Arora, W. A. Sethares, and J. A. Bucklew, "Latent Periodicities in Genome Sequences," *IEEE J. Sel. Topics in Signal Processing*, vol. 2, no. 3, pp. 332–342, June 2008.

[7] L. Wang and D. Schonfeld, "Mapping Equivalence for Symbolic Sequences: Theory and Applications," *IEEE Trans. Signal Process.*, vol. 57, no. 12, pp. 4895–4905, Dec. 2009.

[8] L. Galleani and R. Garello, "The Minimum Entropy Mapping Spectrum of a DNA Sequence," *IEEE Trans. Inf. Theory*, vol. 56, no. 2, pp. 771–783, Feb. 2010.

[9] P. Stoica and R. Moses, *Spectral Analysis of Signals*, Prentice Hall, Upper Saddle River, N.J., 2005.

[10] J. Swärd and A. Jakobsson, "Subspace-based estimation of symbolic periodicities," in *38th International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, May 26-31 2013, accepted.

[11] E. L. Lehmann and G. Casella, *Theory of Point Estimation (Springer Texts in Statistics)*, Springer, 2nd edition, 1998.

[12] G. Casella and R. Berger, *Statistical Inference*, Duxbury, 2nd edition, 2002.

[13] National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/nuccore/FO081497.1.