DETECTION OF PERIODICITIES IN GENE SEQUENCES: A MAXIMUM LIKELIHOOD APPROACH

Raman Arora and William A. Sethares

Department of Electrical and Computer Engineering, University of Wisconsin-Madison
1415 Engineering Drive, Madison WI 53706, ramanarora@wisc.edu, sethares@ece.wisc.edu

ABSTRACT

A novel approach is presented to the detection of homological, eroded and latent periodicities in DNA sequences. Each symbol in a DNA sequence is assumed to be generated from an information source with an underlying probability mass function (pmf) in a cyclic manner. The number of sources can be interpreted as the periodicity of the sequence. The maximum likelihood estimates are developed for the pmfs of the information sources as well as the period of the DNA sequence. The statistical model can also be utilized for building probabilistic representations of RNA families.

I. INTRODUCTION

The structural features of DNA sequences have biological implications [1]. One such structural feature is symbolic periodicity. The periodicities in gene sequences have been linked with evolution of genome and protein strucuture [1], [2], [3]. The DNA sequences exhibit homologous, eroded and latent periodicities. Homologous periodicity occurs when short fragments of DNA are repeated in tandem to give periodic sequences [4]. Most current approaches for finding periodicities transform the symbolic DNA sequence to a numerical sequence [5], [6], [7]; these techniques are primarily aimed at the detection of homological periodicities.

Some researchers have also explored the detection of imperfect or eroded periodicities which model a sequence of similar units repeated but with some changes. Therefore, the homology between repeated units in an eroded sequence is not perfect [4]. The imperfect periodicity may occur in strands of DNA due to changes or erosion of nucleotides.

The periodicity in DNA sequences may also be modeled as latent periodicity [4], for instance an observed period of nucleotides 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. The hidden periodicities may not be found efficiently by algorithms developed for finding homological and eroded periodicities [2]. The latent periodicity detection was studied in [7], [8] and latent periodicities of some human genes were reported.

This paper presents a novel approach to finding latent periodicities in DNA sequences that parallels the extraction of beat information from low level audio features in [9]. Each symbol of the sequence is assumed to be generated by an information source with some underlying probability mass function and the sequence is generated by drawing symbols from these sources in a cyclic manner. The number of sources is equal to the latent period in the sequence and the latent periodicity is same as the statistical periodicity or cyclostationarity. The paper presents maximum likelihood estimates of the pmfs and the period. The symbolic sequences are not transformed into numerical sequences and the method presented here is capable of finding all three kinds of periodicities: homological, eroded and latent.

The problem of detecting latent periodicities in symbolic sequences is formulated mathematically in the next section. The maximum likelihood estimate of the period is developed in section III and results are discussed in section IV. A short discussion on building probabilistic representations for non-coding RNAs is presented in section V.

II. STATISTICAL PERIODICITY

The statistical periodicity model that is employed here to discover possibly hidden periodicities in gene sequences does not assume that the sequence itself is periodic. Instead it is assumed that there is a periodicity in underlying statistical distributions which is locked to a known periodic grid.

A given DNA sequence $\mathcal{D} = [D_1, \ldots, D_N]$ can be denoted by the mapping $\mathcal{D} : \mathbb{N} \rightarrow \mathcal{S}$, from the natural numbers to the alphabet $\mathcal{S} = \{\mathrm{A}, \mathrm{G}, \mathrm{C}, \mathrm{T}\}$. Assume that the statistical periodicity of the sequence $\mathcal{D}$ is T. This implies there are T information sources (or random variables) denoted as $X_1, \ldots, X_T$. The random variable X_i takes values on the alphabet $\mathcal{S}$ according to an associated probability mass function P_i; it generates the j^{th} symbol in $\mathcal{S}$ with probability $P_i(j) = \mathcal{P}(X_i = \mathcal{S}_j)$ for $j = 1, \ldots, |\mathcal{S}|$ where $|\mathcal{S}|$ is the cardinality of the alphabet (which is four for the DNA sequences).

The number of complete statistical periods in $\mathcal{D}$ are $M = \lfloor N/T \rfloor$. Define $\hat{i} = (i \bmod T)$. Then for $1 \leq i \leq N$, the symbol D_i, i.e. the i^{th} symbol in the sequence $\mathcal{D}$, is generated by the random variable $X_{\hat{i}}$. The random variables $X_{\hat{i}}$, $\hat{i} = 1, \ldots, T$ are assumed to be independent. The *structural parameters*, $P_1, \ldots, P_T$, and the *timing parameter* T are unknown. Define $\Theta = [T, P_1, \ldots, P_T]$. The search space for parameter T is

1-4244-0999-3/07/$25.00

the set $B = \{1, \ldots, N_0\}$ for some $N_0 < N$ and for the pmfs $[P_1, \ldots, \mathcal{P}_T]$ the search space is the subset $\mathcal{Q} \subseteq [0,1]^{|S| \times T}$ of column stochastic matrices (for $P \in \mathcal{Q}$, $P_{ji} \in [0,1]$ and $\sum_{j=1}^{|S|} P_{ji} = 1$ for $i = 1, \ldots, T$). Let $\wp = B \times \mathcal{Q}$ denote the search space for the parameter Θ. Given the data, the maximum aposteriori (MAP) estimate of parameter Θ is

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

By Bayes rule the posterior probability is

$$\mathcal{P}(\Theta|\mathcal{D}) = \frac{\mathcal{P}(\mathcal{D}|\Theta)\mathcal{P}(\Theta)}{\mathcal{P}(\mathcal{D})}, \tag{1}$$

where, by independence of X_i's,

$$\mathcal{P}(\mathcal{D}|\Theta) = \prod_{i=1}^{N} \mathcal{P}(X_{\hat{i}} = D_i|\Theta) \tag{2}$$

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

$$\hat{\Theta} = \arg\max_{\Theta \in \wp} \mathcal{P}(\mathcal{D}|\Theta). \tag{3}$$

In words, the MAP estimate is same as the maximum likelihood estimate.

III. THE MAXIMUM LIKELIHOOD ESTIMATE

This section develops the maximum likelihood estimate (MLE) for the unknown parameter Θ. The data-sequence $\mathcal{D} = [D_1, \ldots, D_N]$ is represented by a sequence of vectors $\mathcal{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_N]$ where each $\mathbf{w}_i$ is an $|S| \times 1$ vector with $\mathbf{w}_{ji} = 1 \iff D_i = \mathcal{S}_j$. So, if the i^{th} symbol in the sequence $\mathcal{D}$ is C, i.e. the third symbol of the alphabet $\mathcal{S}$, then the i^{th} vector $\mathbf{w}_i$ in the sequence $\mathcal{W}$ is $[\,0\;0\;1\;0\,]'$. Also define a $|S| \times T$ stochastic matrix $\mathbb{A}$ with entries $\mathbb{A}_{ji} = \mathcal{P}(X_i = \mathcal{S}_j)$. 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 $\mathcal{S}$. Write the unknown parameter $\Theta = [\mathbb{A}, T]$. This notation simplifies the derivation of the MLE by noting that

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

The likelihood can therefore be written as

$$\begin{aligned}
\mathcal{P}(\mathcal{W}|\mathbb{A}, T) &= \prod_{i=1}^{N} \mathcal{P}(X_{\hat{i}} = D_i|\mathbb{A}, T) \\
&= \prod_{i=1}^{N} \prod_{j=1}^{|S|} \left(\mathbb{A}_{j\hat{i}}\right)^{\mathbf{w}_{ji}} \\
&= \prod_{k=1}^{M} \prod_{\hat{i}=1}^{T} \prod_{j=1}^{|S|} \left(\mathbb{A}_{j\hat{i}}\right)^{\mathbf{w}_{ji^{(k)}}} \times \\
&\qquad \prod_{\hat{i}=1}^{N-MT} \prod_{j=1}^{|S|} \left(\mathbb{A}_{j\hat{i}}\right)^{\mathbf{w}_{ji^{(M+1)}}}
\end{aligned} \tag{4}$$

where $i^{(k)} = (k-1)T + \hat{i}$. Note that the first term on the right hand side of (4) captures the observations in M complete periods (given the period T) while the second product captures the observation over the last incomplete cycle. The log-likelihood is

$$\begin{aligned}
\log \mathcal{P}(\mathcal{W}|\mathbb{A}, T) &= \sum_{k=1}^{M} \sum_{\hat{i}=1}^{T} \sum_{j=1}^{|S|} \mathbf{w}_{ji^{(k)}} \log\left(\mathbb{A}_{j\hat{i}}\right) + \\
&\qquad \sum_{\hat{i}=1}^{N-MT} \sum_{j=1}^{|S|} \mathbf{w}_{ji^{(M+1)}} \log\left(\mathbb{A}_{j\hat{i}}\right)
\end{aligned} \tag{5}$$

For a fixed T, the MLE for $\mathbb{A}$ is

$$\hat{\mathbb{A}}^T = \arg\max_{\mathbb{A} \in \mathcal{Q}} \log \mathcal{P}(\mathcal{W}|\mathbb{A}, T). \tag{6}$$

The log-likelihood in (5) is a concave function of variables $\mathbb{A}_{j\hat{i}}$. Also, note that these variables satisfy the constraint: $\sum_{j=1}^{|S|} \mathbb{A}_{j\hat{i}} = 1$ for $\hat{i} = 1, \ldots, T$. Constrained optimization using Lagrange multipliers gives the $(j, \hat{i})^{th}$ element of the matrix $\hat{\mathbb{A}}^T$ as

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

for $j = 1, \ldots, |S|$. The estimates of the parameters $\mathbb{A}$ can then be used to determine the MLE for the period T,

$$T^* = \arg\max_{T \in B} \log \mathcal{P}(\mathcal{W}|\mathbb{A}^T, T). \tag{8}$$

IV. RESULTS

The method in the previous section was applied to a variety of simulated symbolic sequences and to chromosome XVI of $\mathcal{S}$. *cerevisiae* in order to detect periodicities. A homological symbolic sequence from the set $\mathcal{S} = \{\mathrm{A}, \mathrm{G}, \mathrm{C}, \mathrm{T}\}$ with period $T = 7$ was generated. The sequence was eroded by changing the symbols at randomly chosen points in the sequence. The algorithm was tested with various degrees of erosion. The plots in Fig. 1 strongly support a statistical periodicity of 7 even with 85% erosion. The noise floor in the plots increases (i.e. the heights of the peaks decrease) with increased erosion. Note that a T-periodic sequence also shows kT-periodicity for any positive integer k.

Figure 2(a) shows the results with latent periodicity of simulated symbolic sequence where a single period is (A/C)(T/G)(T/A)(G/T)(C/G/A)(G/A). This was generated by six information sources, $X_1, \ldots, X_6$ with X_1 generating A or C each with equal probability and X_5 generating A,G or C each with probability $1/3$. Applying the technique of Sect. III to this sequence results in a plot showing strong six-periodic behaviour. In contrast, when a random sequence is used (i.e. when each source generates all symbols with equal frequency), Fig. 2(b) shows that no significant periodicities are detected.

The algorithm was also tested with the protein coding region of chromosome XVI of $\mathcal{S}$. *cerevisiae* (GenBank accession number NC 001148). The 2160 base-pair(bp)

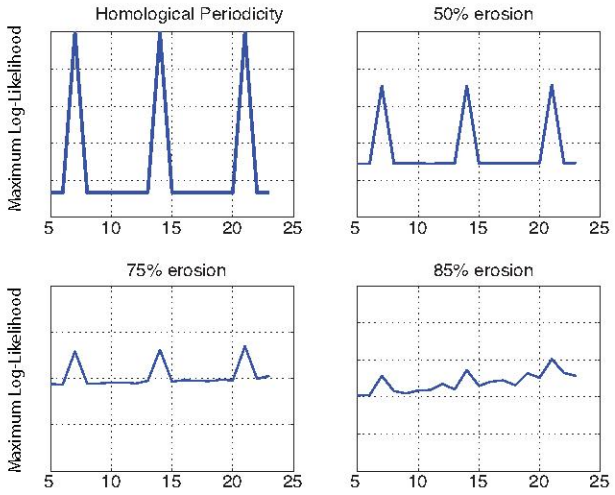

Figure 1. Maximum log-likelihood of data plotted against Period for a simulated symbolic sequence of length 6400 symbols with period 7: (a) Homological periodic sequence (b) 50% eroded sequence (c) 75% eroded sequence (d) 85% eroded sequence.

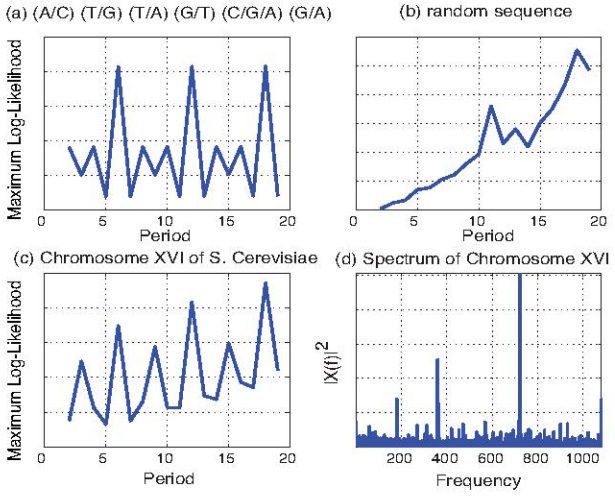

Figure 2. (a) Maximum Log-likelihood of data plotted against period for a simulated symbolic sequence of length 6400 symbols with latent periodicity 7, (b) maximum log-likelihood versus period for completely random symbolic sequence (c) maximum log-likelihood plotted against period for protein coding region of chromosome XVI of *S. cerevisiae* (d) the magnitude of DFT of numerical sequence derived from the sequence in part(c).

long sequence (from bp 85 - 2244) shows a latent periodicity of period three as plotted in Fig. 2(c). The period-3 behaviour of protein coding genes is expected since amino acids are coded by trinucleotide units called *codons* [7], [10]. For comparison, the symbolic sequence is transformed into a numerical sequence as in [7] and the magnitude of the 2160-point DFT is plotted in Fig. 2. The peaks at $f_1 = 720$, $f_2 = 360$ and $f_3 = 180$ correspond to $3, 6$ and 12-periodic behaviour respectively.

V. IDENTIFYING NON-CODING RNAS

The central dogma of molecular biology states that DNA is transcribed to RNA and then translated to proteins. The genetic information therefore flows from DNA to protein through the RNA. However, besides playing the role of a passive intermediary messenger (mRNA), RNAs have been known to play important non-coding function in the process of translation (tRNA, rRNA) [11]. Since these RNAs are not translated into proteins, these are called non-coding RNAs (ncRNAs) or RNA genes. Originally, such RNA genes were considered rare but in the last decade many new RNA genes have been found and have been shown to play diverse roles: chromosome replication, protein degradation and translocation, regulating gene expression and many more. Thus RNA genes may play a much more significant role than previously thought. 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 [10].

The statistical model presented in this paper for finding periodicities in symbolic sequences can be utilized for building probabilistic representations of RNA families. The 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 the 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 a *hairpin* structure as shown in Fig. 3(a); the consecutive base pairs that bond together get stacked onto each other to form a *stem* while the unpaired bases form a *loop*.

Typical methods employed for identification of DNA gene sequences and proteins do not perform as well in the 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 [10]. Therefore, in the identification of ncRNAs 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 [12].

The RNA sequences preserve the secondary structure when undergoing erosion or mutation by compensatory mutation as shown in Fig. 4. This causes 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 a sequence of ncRNA molecules, tandem repeats of which

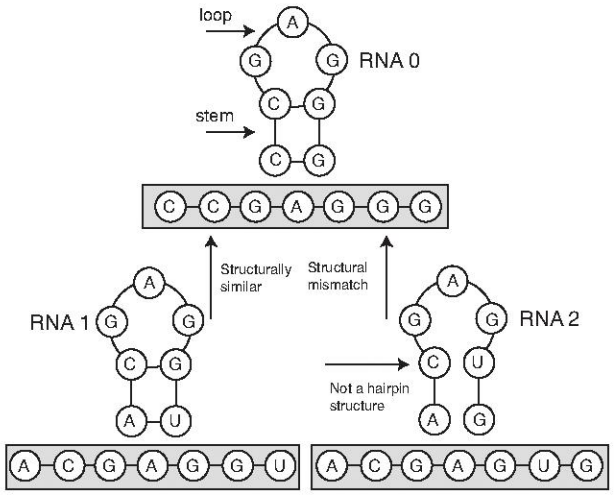

Figure 3. (a) RNA0 has hairpin secondary structure. (b) RNA1 is similar in structure to RNA0. It differs at two positions in the primary sequence from RNA0. (c) RNA2 structure is not hairpin, it has a structural mismatch with RNA0. RNA2 also differs at two position in the primary sequence from RNA0 but it must be scored lower in similarity to RNA0 as compared to RNA1.

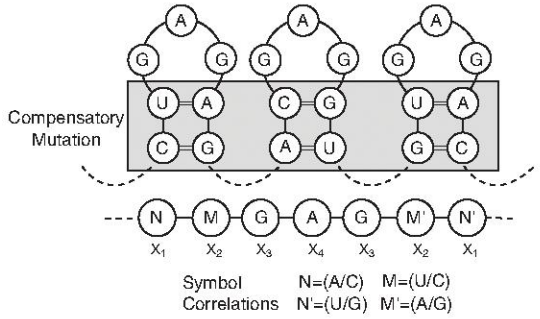

Figure 4. A given base pair in a ncRNA molecule undergoes compensatory mutation i.e if one of the nucleotides in a base-pair mutates, the other nucleotide also changes to complementary nucleotide. So there is a strong correlation between the base positions indicated by N and N′ and base positions indicated by M and M′.

have undergone random *compensatory mutations* as shown in Fig. 4. According to the statistical periodicity model presented in this paper, the sources generating the symbols that do not bond (nucleotides in the loop of a hairpin ncRNA) have a point-mass pmf. On the other hand, the sources corresponding to a bonded base pair have *conjugate distributions* (in Fig. 4 pmfs of N and N′ agree on complementary bases). These conjugate distributions capture the distant base-pair correlations. The sequence of sources corresponding to ncRNA molecules in Fig. 4 after identifying the bonded nucleotides with conjugate sources is $X_1, X_2, X_3, X_4, X_3, X_2', X_1'$. The statistical periodicity model presented here is therefore capable of describing primary as well as secondary structural similarities.

VI. ACKNOWLEDGEMENTS

The authors would like to thank Charu Gera for various helpful discussions on molecular biology.

VII. REFERENCES

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

[2] 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.

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

[4] 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. 153–163, Feb. 1999.

[5] V. R. Chechetkin, L. A. Knizhnikova, and A. Y. 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.

[6] 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.

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

[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] W. A. Sethares, R. D. Morris, and J. C. Sethares, "Beat tracking of audio signals using low level audio features," *IEEE Transactions On Speech and Audio Processing*, vol. 13, no. 2, pp. 275–285, March 2005.

[10] B. J. Yoon and R. 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.

[11] M. S. Waterman, *Introduction to Computational Biology: Maps, sequences and genomes*, Chapman and Hall/CRC, first edition, 1995.

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