DETECTION OF PERIODICITIES IN GENE SEQUENCES: A MAXIMUM LIKELIHOOD APPROACH

Raman Arora and William A. Sethares

University of Wisconsin
Department of Electrical and Computer Engineering
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 then 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 presented can also be utilized for building probabilistic representations of RNA families.

1. INTRODUCTION

The structural features of DNA sequences have biological implications [1]. One such structural feature is symbolic periodicity. Finding periodicities in DNA sequences is important to the determination and understanding of the structure of DNA sequences in various genomes [1, 2, 3]. 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 detection of imperfect or eroded periodicities which model a sequence of similar units repeated but with some changes. In other words 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 me 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 [8, 7] 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. The number of sources is equal to the detected period in the sequence and the symbols are assumed to be drawn from these sources in a cyclic manner. The latent periodicity is then interpreted as statistical periodicity. The paper presents maximum likelihood estimates of the pmfs and the period. Note that the symbolic sequence is not transformed into a numerical sequence and the method presented here is capable of finding all three kinds of periodicities - homological, eroded and latent.

2. 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 gene sequence $\mathcal{D} = [D_1, \ldots, D_N]$ can be denoted by the mapping $\mathcal{D} : \mathbb{N} \rightarrow \mathcal{S}$, where $\mathcal{S}$ is the alphabet $\{A, G, C, T\}$. Assume that the statistical periodicity of the gene sequence $\mathcal{D}$ is $\mathcal{T}$. This implies there are $\mathcal{T}$ information sources (or random variables) denoted as $X_1, \ldots, X_{\mathcal{T}}$. 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 $P_i(j) = \mathcal{P}(X_i = \mathcal{S}_j),\ j = 1, \ldots, |\mathcal{S}|$. Note that $|\mathcal{S}|$ denotes the cardinality of the alphabet $\mathcal{S}$ which for the gene sequencing problem is four.

The number of statistical periods in $\mathcal{D}$ are $M = \lfloor N/\mathcal{T} \rfloor$. Define $\hat{i} = (i \mod \mathcal{T})$. Then for $1 \leq i \leq N$, the symbol D_i, i.e. the i^{th} symbol in the gene sequence $\mathcal{D}$, is generated by the source (random variable) $X_{\hat{i}}$. The *structural parameters*, $P_1, \ldots, P_{\mathcal{T}}$, and the *timing parameter* $\mathcal{T}$ are unknown. Define $\Theta = [\mathcal{T}, P_1, \ldots, P_{\mathcal{T}}]$. The maximum aposteriori (MAP) estimate of Θ is given as

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

By Bayes rule the posterior probability is given as

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

where

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

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

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

i.e. the MAP is same as the maximum likelihood estimate.

3. THE MAXIMUM LIKELIHOOD ESTIMATE

In this section, the maximum likelihood estimate (MLE) is developed 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 $|S| \times \mathcal{T}$ stochastic matrix A with entries $A_{ij} = \mathcal{P}(X_i = \mathcal{S}_j)$. The columns of the matrix A denote the pmfs of the information sources; the entry A_{ij} denotes the probability that the i^{th} source generates the j^{th} symbol of the alphabet $\mathcal{S}$.

This notation simplifies the derivation of the MLE. Note that

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

and the unknown parameter $\Theta = [A, \mathcal{T}]$. The likelihood can therefore be written as

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

where $i^{(k)} = (k-1)M + \hat{i}$. The log-likelihood is given as

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

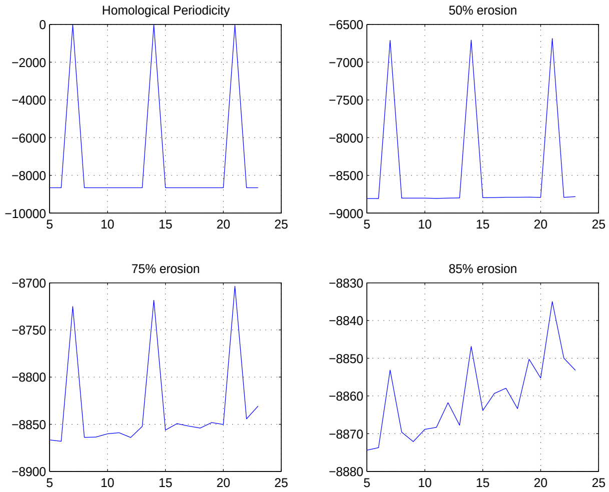

Figure 1. Maximum log-likelihood of data plotted against Period for a simulated symbolic sequence of length 6400 symbols with period 7: (a) Homological periodic sequence (b) 50% eroded sequence (c) 75% eroded sequence (d) 85% eroded sequence.

For a fixed $\mathcal{T}$ the maximum likelihood estimate for A is denoted

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

From equation (5), the $(i,j)^{th}$ element of the matrix $A^{\mathcal{T}}$ is given as

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

for $j = 1, \ldots, \mathcal{T}$.

The maximum likelihood estimate for A is plugged-in to determine the maximum likelihood estimate for the period $\mathcal{T}$,

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

4. RESULTS

The method for detecting periodicities in symbolic sequences was applied to simulated symbolic sequences and chromosome XVI of $\mathcal{S}$. *cerevisiae*. A homological symbolic sequence from the set $\mathcal{S} = A, G, C, T$ with period 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 figure 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 decreases) with the degree of erosion. Note that a $\mathcal{T}$-periodic sequence also shows $p\mathcal{T}$-periodicity for any positive integer p.

Figure 2(a) shows the results with latent periodicity of simulated symbolic sequence where a single period is

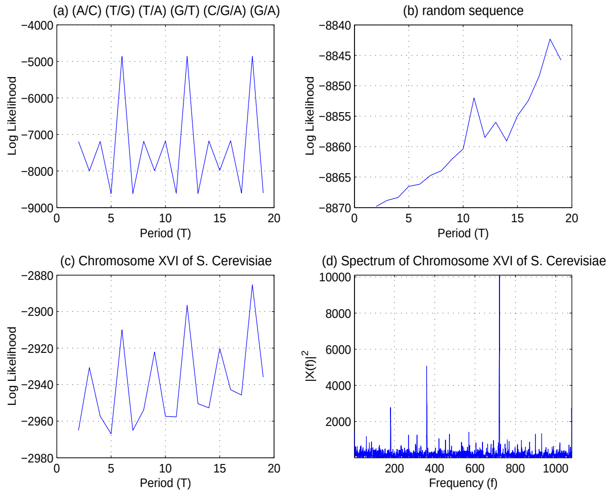

Figure 2. (a) Log-likelihood of data plotted against period for a simulated symbolic sequence of length 6400 symbols with period latent periodicity 7: (b) log-likelihood versus period for completely random symbolic sequence (c) log-likelihood plotted against period for protein coding region of chromosome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

.

(A/C) (T/G)(T/A) (G/T) (C/G/A) (G/A). The sequence is assumed to be generated by six information sources, say $X_1, \ldots, X_6$ with X_1 generating A or C each with equal probability; similarly X_5 generates A,G or C each with probability $1/3$. The plot shows strong six-periodic behaviour. Figure 2(b) shows results with completely random sequence, i.e. each source generating each symbol with equal probability.

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

5. IDENTIFYING NON-CODING RNAS

The central dogma of molecular biology states that DNA is transcribed to RNA and then translated to a protein. 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) also known as RNA genes.

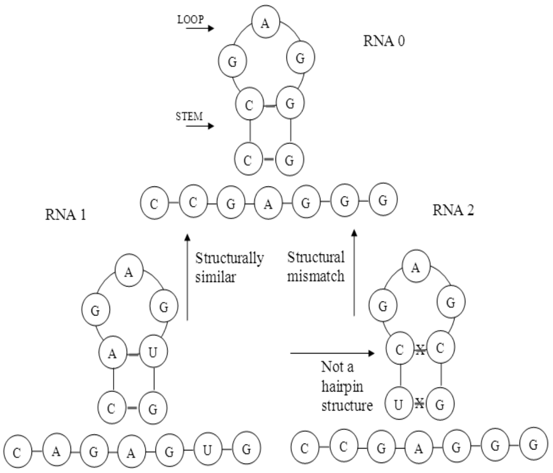

Figure 3. (a) RNA0 has hairpin secondary structure. (b) RNA1 is similar in structure to RNA0. It differs at two position in the primary sequence from RNA0. (c) RNA2 structure is not hairpin, it has a structural mismatch with RNA0. RNA2 also differs at two position in the primary sequence from RNA0 but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

.

The 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 another nucleotide uracil (U) as the base complementary to adenine (A). So, RNA is reprented by the string of nucleotides (or 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 structures called secondary and tertiary respectively. A typical secondary structure is *hairpin* structure as shown in figure 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*.

The methods employed for identification of DNA gene sequences and proteins do not perform well at identification of ncRNAs because many functional ncRNAs pre-

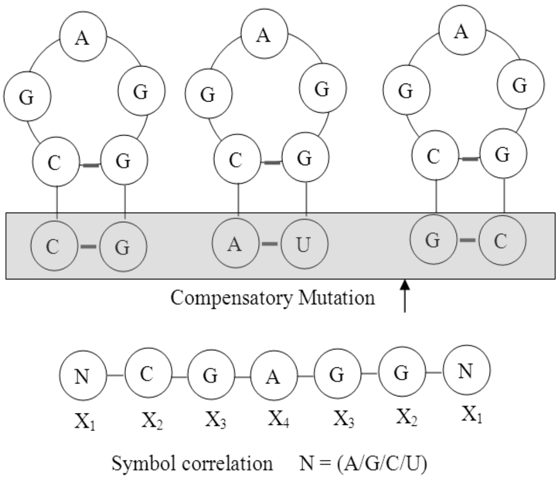

Figure 4. A given base pair in a ncRNA molecule undergoes compensatory mutation i.e if one of the nucleotides in a base-pair mutates, the other nucleotide also changes to complementary nucleotide. So there is a strong correlation between the two base positions indicated by N.

serve their secondary structures more than they preserve their primary sequences [10] and these techniques have been based on finding structural features (like periodicities) in (primary) sequences. Therefore, in identification of ncRNAs there is need for techniques that evaluate similarity between RNA molecules and sequences based on secondary structures also. Such techniques have 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 figure 4. This causes a strong pairwise correlations between distant bases in the primary RNA sequence. Unlike the techniques employed for DNA identification, the approach presented here can describe such pairwise correlations. Consider a sequence of ncRNA molecules, tandem repeats of which have undergone random mutations as shown in figure 4. According to the statistical 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 very *similar* pmfs (in the sense of Kullback-Leibler divergence [13]). If the information sources with similar pmfs are identified as the same the sequence of sources form a *palindrome*; the sequence of sources corresponding to ncRNA molecule in figure 4 after identifying the bonded nucleotides with the same source is X_1, X_2, X_3, X_4, X_3, X_2, X_1. The statistical model presented here is therefore capable of describing structural similarities too.

6. 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.

[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley-Interscience, second edition, 2006.