Finite Set DSP, with Applications to DNA Sequences

Ronald K. Pearson, Gregory E. Gonye
Department of Pathology, Anatomy,
and Cell Biology
Thomas Jefferson University
1020 Locust St., Philadelphia, PA, USA
{pearson,ggonye}@mail.dbi.tju.edu

and

Moncef Gabbouj
Institute of Signal Processing
Tampere University of Technology
Tampere, Finland
moncef.gabbouj@tut.fi

Abstract

Regular substructures in DNA sequences are important in a number of biological problems including promoter analysis, the detection of recurring anomalies in tumor cells, and the study of certain genetic diseases like fragile-X mental retardation. This paper considers signal processing problems relevant to the analysis of regular or semi-regular structure in DNA sequences that must address the fundamental issue of working with unordered, finite value sets.

1 Introduction

Motivated in part by an interest in the analysis of promoter regions in DNA sequences, this paper considers a number of signal processing issues that arise when dealing with sequences that take values in a finite set Σ. Since nonparametric spectrum estimation is an extremely useful tool for the exploratory analysis of real-valued data sequences, the primary focus of this paper is on some extensions of classical spectrum estimation procedures to Σ-valued sequences. One approach would be to simply label the categories with real index values, effectively converting these sequences into real-valued sequences, permitting the use of well-developed standard methods. It is known, however, that the results obtained in this way generally exhibit an undesirable dependence on the details of the labelling considered (Bloch and Arce, 2002; Buchner and Janjarasjitt, 2003; Johnson and Wang, 1999), in part because any mapping from Σ into the real numbers induces an ordering on the elements of Σ, which may be highly unnatural. To avoid order-induced artifacts, here we adopt the *nominal variable model* from cluster analysis (Gordon, 1999, p. 18) that imposes no additional structure on the set Σ. As a consequence, it is not possible to define linear operations on Σ, so Σ-valued signal processing is inherently nonlinear.

2 BT spectrum estimation

For a stationary sequence $\{x_k\}$ of real-valued random variables, the power spectral density $S_{xx}(f)$ is defined as the discrete Fourier transform of the autocorrelation function:

$$S_{xx}(f) = \sum_{k=-\infty}^{\infty} R_{xx}(k) e^{-i2\pi kfT}. \qquad (1)$$

In this definition, it is assumed that $\{x_k\}$ is a uniformly sampled time-series with intersample spacing T, and $R_{xx}(k)$ is the autocorrelation function:

$$\begin{aligned} R_{xx}(k) &= E\{(x_j - E\{x_j\})(x_{j+|k|} - E\{x_j\})\} \\ &= \rho(x_j, x_{j+|k|})\sigma^2, \qquad (2) \end{aligned}$$

where $\rho(x_j, x_{j+k})$ denotes the correlation coefficient between x_j and x_{j+k} and σ^2 is the variance of the sequence $\{x_k\}$. One way of converting this definition into a computational procedure is to consider the Blackman-Tukey estimator (Kay, 1988, p. 77):

$$\hat{S}_{xx}(f) = \sum_{k=-M}^{M} w_k \hat{R}_{xx}(k) e^{-i2\pi kfT}, \qquad (3)$$

where the real numbers $\{w_k\}$ define a *lag window*, included to manage the bias-variance tradeoff inherent in spectrum estimation (Kay, 1988; Priestley, 1981).

To adapt this formulation to Σ-valued sequences, it is only necessary to specify a useful autocorrelation estimator $\hat{R}_{xx}(k)$. In cluster analysis, the correlation coefficient ρ between two real-valued data vectors provides the basis for a useful *dissimilarity measure* between vectors (Kaufman and Rousseeuw, 1990, p. 19):

$$d_{xy} = \frac{1-\rho_{xy}}{2}. \tag{4}$$

Since it is possible to define dissimilarity measures for sequences taking values in a finite set Σ, we reverse the relation defined in Eq. (4) to obtain the desired autocorrelation measure:

$$\hat{R}_{xx}(k) = \rho(x_j, x_{j+k}) = 1 - 2d(x_j, x_{j+k}). \tag{5}$$

Here, we consider the following dissimilarity measure between subsequences of fixed length K:

$$d(x_j, x_{j+k}) = \frac{1}{K}\sum_{i=0}^{K-1} \delta(x_{i+j}, x_{i+j+k}), \tag{6}$$

where $\delta(x_i, x_j) = 0$ if $x_i = x_j$ and 1 otherwise. Finally, since the autocorrelation estimates $\hat{R}_{xx}(k)$ obtained from Eqs. (5) and (6) generally exhibit a nonzero mean $\bar{R}$, a large zero-frequency peak and its associated side-lobes appears in the estimated power spectrum. These features obscure the spectral characteristics of interest, so we remove them by replacing $\hat{R}_{xx}(k)$ with $\hat{R}_{xx}(k) - \bar{R}$ in Eq. (3).

3 Validation results

Fig. 1 summarizes two results obtained using the spectrum estimation procedure just described, applied to a perfectly periodic sequence of length $L = 100$, corresponding to 20 repetitions of the subsequence GGCTG. The higher-amplitude (solid) curves in these plots correspond to the spectrum estimates obtained using two different lag windows, and the lower-amplitude (dotted) curves correspond to the permutation-based validation results discussed below. The left-hand plot was obtained using the rectangular window $w_k = 1$ for $-25 \leq k \leq 25$ and the right-hand plot shows the results obtained using the triangular Bartlett window defined on the same support set (Priestley, 1981, p. 439). As expected, the Bartlett window reduces variability at the expense of increased bias, which appears here in the form of reduced intensity spectral peaks. In both cases, a clear peak is evident at the fundamental frequency $f = 0.2$, corresponding to the period $P = 5$ of the repetitive sequence, and a weaker peak appears at the second harmonic, $f = 0.4$. The key points here are first, that much of the machinery of classical spectral analysis (e.g., classical lag windows) can be applied to the finite-set formulation proposed here and second, that this procedure gives the correct results in the case of simple periodic sequences.

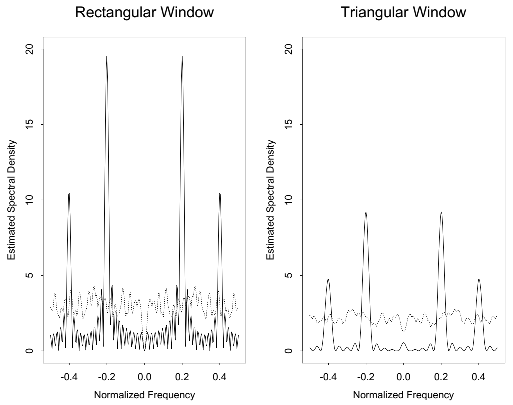

Figure 1: Estimated spectrum and permutation limits using rectangular (left) plot and Bartlett (right plot) lag windows

The lower-amplitude (dotted) curves in Fig. 1 were obtained by applying a variant of the computational negative controls (CNC) strategy proposed by Pearson et al. (2004) for cluster analysis. Specifically, after the spectrum estimate was obtained from the original data sequence, the same spectrum estimation procedure was applied to each of 50 random permutations of this data sequence. These results provide a useful frame of reference since the permutations destroy any regular sequential structure present in the original data sequence, giving essentially 50 white noise sequences with the same distribution of values as this original sequence. Since these sequences should exhibit constant power spectra, only those features in the original spectrum that significantly exceed these randomized spectra should be regarded as significant. The lower curve in Fig. 1 represents the maximum value obtained, at each frequency f, from

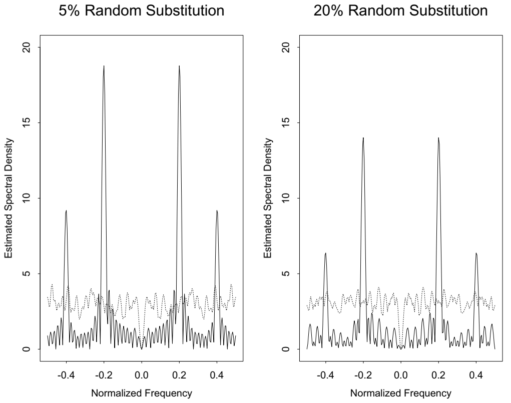

Figure 2: Effects of 5% (left plot) and 20% (right plot) random substitutions on spectrum estimates

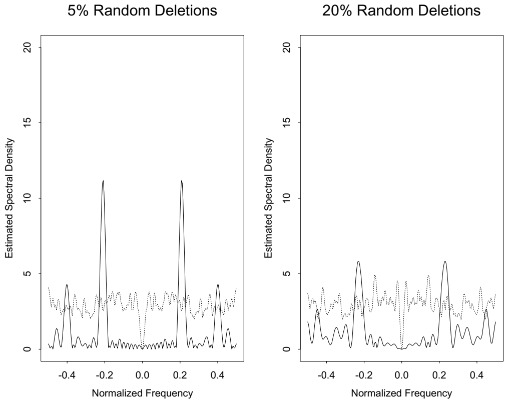

Figure 3: Effects of 5% (left plot) and 20% (right plot) random deletions on spectrum estimates

these 50 randomized results. In Fig. 1, both of the plots show clear evidence of the periodic structure present in the sequence since both the fundamental and the second harmonic peaks clearly exceed this lower reference curve.

For comparison, Fig. 2 shows the results obtained for two contaminated versions of the periodic sequence considered above. Specifically, the left-hand plot shows the results obtained for this periodic sequence contaminated with 5% random substitutions, while the right-hand plot shows the results obtained with 20% random substitutions. The rectangular lag window was used in both cases since it gives intense spectral peaks that are easily distinguished from the randomized reference results. Motivation for considering this contamination model is two-fold: first, it is directly relevant to the study of regular structures in DNA sequence data and second, it is analogous to the problem of outliers that causes significant bias in real-valued dynamic data characterizations (Pearson, 2001). In particular, it is known that outliers "raise the noise floor" in real-valued spectrum estimation, obscuring high-frequency details (Martin and Thomson, 1982). Here, however, the fact that all data values must belong to the small set Σ bounds the magnitude of possible outliers and appears to substantially reduce their severity. In particular, although comparison of Figs. 1 and 2 shows clearly that increased contamination levels cause degradation of the spectral peaks, both the fundamental and the second harmonic peaks remain significant relative to the CNC baseline even with 20% contamination.

Two other important phenomena in DNA sequence characterization are random insertions and random deletions, collectively known as *indels.* Fig. 3 shows the results obtained for the periodic sequence considered above, but with 5% and 20% random deletions in the left-hand and right-hand plots, respectively. Comparing these results with the corresponding random substitution results, it is clear that random deletions pose a much more serious problem for spectrum estimation on finite sets than random substitutions do. Results obtained for comparable levels of random insertions (not shown) are almost identical to those shown here for random deletions.

4 Chromosome 22 results

Although space limitations do not permit a detailed discussion, Fig. 4 shows the results obtained with the spectrum estimator described here for a sequence of 100 bases extracted from human chromosome 22 (July 2003 assembly, from http://genome.ucsc.edu). This chromosome is approximately $50,000,000$ bases long and the sequence considered here corresponds to bases $40,052,600$ through $40,052,699$, selected

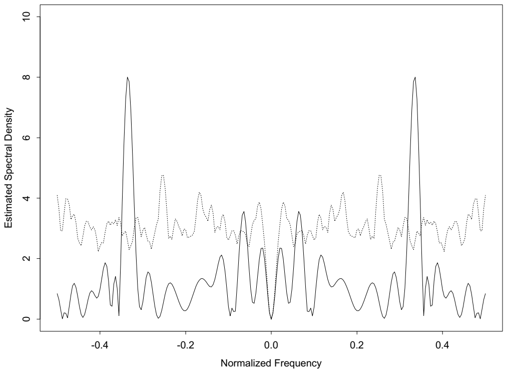

Figure 4: Spectrum estimate from a 100 base subsequence from human chromosome 22

because expert annotation indicates the presence of an approximately periodic repetition of the sequence GGA in the middle of this subsequence. The presence of a peak at $f \simeq 0.33$ that is significant relative to the CNC background spectrum in Fig. 4 is consistent with this characterization.

5 Summary and extensions

This paper has presented a brief, preliminary summary of a spectrum estimation procedure for nominal data sequences, for which ideas of linearity are inherently inapplicable. The approach proposed here computes autocorrelations from dissimilarity measures that can be computed for nominal data sequences without the need for encodings that may induce spurious orderings on the data values. We are currently exploring extensions of this idea that use alternative dissimilarity measures, including those that allow specific mismatches and those appropriate to detection of more complex structures like complimented palindromes (Gusfield, 1997, p. 139). In addition, it is clear from the results presented here that outliers, which pose a sigifncant problem for real-valued spectrum estimation (Martin and Thomson, 1982; Pearson, 2001), are not especially serious here. In particular, finite-set outliers correspond to the random substitutions considered in Sec. 3 where it was seen that the biologically-motivated problem of random insertions and/or deletions (indels) is much more serious. Consequently, we are exploring spectrum estimation and filtering ideas based on variable data windows that can potentially address these issues.

References

Bloch, K. and Arce, G. (2002). Analyzing protein sequences using signal analysis techniques. In Zhang, W. and Shmulevich, I., editors, *Computational and Statistical Approaches to Genomics*, pages 113–124. Kluwer.

Buchner, M. and Janjarasjitt, S. (2003). Detection and visualization of tandem repeats in dna sequences. *IEEE Trans. Signal Proc.*, 51:2280–2287.

Gordon, A. (1999). *Classification.* Chapman and Hall, 2nd edition.

Gusfield, D. (1997). *Algorithms on Strings, Trees, and Sequences.* Cambridge University Press.

Johnson, D. and Wang, W. (1999). Symbolic signal processing. In *Proc. Int. Conf. Acoustics Speech Signal Processing*, volume 3, pages 1361–1364.

Kaufman, L. and Rousseeuw, P. (1990). *Finding Groups in Data.* Wiley.

Kay, S. (1988). *Modern Spectral Estimation: Theory and Application.* Prentice-Hall.

Martin, R. and Thomson, D. (1982). Robust-resistant spectrum estimation. *Proc. IEEE*, 70:1097–1114.

Pearson, R. (2001). Outliers in process modeling and identification. *IEEE Trans. Control System Technology*, 10:55–63.

Pearson, R., Zylkin, T., Schwaber, J., and Gonye, G. (2004). Quantitative evaluation of clustering results using computational negative controls. In *Proc. 2004 SIAM Intl. Conf. Data Mining*, Lake Buena Vista, FL. to appear.

Priestley, M. (1981). *Spectral Analysis and Time Series.* Academic Press.