In Silico Biology 6 (2006) 215–222
IOS Press

Are Categorical Periodograms and Indicator Sequences of Genomes Spectrally Equivalent?

Achuthsankar S. Nair[1] and T. Mahalakshmi[2]

Department of Computer Science, University of Kerala, India – 695 581

Edited by H. Michael; received 17 November 2005; revised and accepted 22 March 2006; published 20 May 2006

ABSTRACT: This paper reports a novel symbol-to-signal mapping for DNA sequences, based on the concept of categorical periodograms. A categorical periodogram is a numeric sequence with the n-th element of the sequence indicating the number of occurrences of cycles with period n in it. The period of the cycle is defined as the number of intervening events plus one. Spectral analysis studies have been conducted on Cumulative Categorical Periodogram (CCP) of 10 genes from the data set of Burset and Guigo. It is observed that the spectral signatures in CCP are functionally equivalent to the established $N/3$ peak in the spectrum of indicator sequences of genomes. Being a single sequence compared to four sequences in the case of indicator sequence representation, the method is claimed to be functionally equivalent, but computationally better for identification of gene coding regions in sequences.

KEYWORDS: Digital signature, categorical periodogram, cumulative categorical periodogram, mapping, indicator sequences, genomic signal processing

INTRODUCTION

Being a nascent field, bioinformatics is continuing to apply new as well as traditional tools and techniques from other fields of knowledge to solve its problems. That bioinformatics concerns itself with a number of types of symbol sequences which are being generated in astronomical volumes, has been a major concern in dovetailing the traditional tools and techniques with bioinformatics. Symbol sequences representing genes and proteins are considered as Categorical Time Series in mathematical parlance. While there are specialized methods to handle such data [1,2] there have been major attempts in bioinformatics to transform the categorical series data to numerical data so that they are made amenable to a host of traditional techniques of analysis.

Genomic Signal Processing (GSP) has spearheaded this approach wherein DNA sequences are converted to 4 numerical indicator sequences and the traditional Digital Signal Processing (DSP) are applied [3–6]. One of the major problems in GSP is the mapping technique applied to transform the symbol

[1]Present address: Centre for Bioinformatics, University of Kerala, Thiruvananthapuram, India 695 581. E-mail: sankar.achuth@gmail.com.

[2]Present address: Sree Narayana Institute of Technology, Kollam, India 691 010 and National Institute of Computer Technology, Kollam, India 691 001. E-mail: mlakshmi@sancaharnet.in.

Electronic publication can be found in *In Silico Biol.* **6**, 0019 <http://www.bioinfo.de/isb/2006/06/0019/>, 20 May 2006.

Table 1
Four Indicator sequences corresponding to a small DNA sequence

String	A	G	G	C	C	T	T	A	G	G	C	T	A	A	A
S_A	1	0	0	0	0	0	0	1	0	0	0	0	1	1	1
S_G	0	1	1	0	0	0	0	0	1	1	0	0	0	0	0
S_C	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0
S_T	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0

sequence into numerical sequence [7]. Intuitively, this mapping could be a crucial choice which could reveal or hide the significant signatures in the gene sequences.

The various mapping techniques reported are indicator sequences [8], electron-ion interaction pseudopotential (EIIP) sequences [9], inter nucleotide distance sequence [10], binucleotide distance sequences [11], complex sequences [12], mapping of nucleotides to digits 0, 1, 2 & 3 [13] etc. Of these, indicator sequences are widely used for analyzing sequences for predicting various hotspots in the sequences. Recently a very efficient way to compute the DNA spectrum to detect the period-3 component is proposed in [14] in which a new parameter 'Position Count Function' (PCF) is defined for the indicator sequences. In this paper, a new mapping technique is presented which yields results similar to established results with certain advantages.

INDICATOR SEQUENCE

One of the most important mapping techniques that are being used frequently is indicator sequence mapping. These are sequences containing numbers zero or one to indicate the absence or presence of the symbol in the original sequence. Since nucleotide sequence consists of four symbols – A, G, C, and T – there are four indicator sequences for a given nucleotide sequence. The four sequences represent the frequency content of each nucleotide.

These four sequences are defined as follows. Consider a sequence S of length N consisting of four symbols A, G, C, T. The four indicator sequences are:

$$S_A(n) = 1 \text{ if } S(n) = A \text{ or } 0 \text{ otherwise}$$

$$S_G(n) = 1 \text{ if } S(n) = G \text{ or } 0 \text{ otherwise}$$

$$S_C(n) = 1 \text{ if } S(n) = C \text{ or } 0 \text{ otherwise}$$

$$S_T(n) = 1 \text{ if } S(n) = T \text{ or } 0 \text{ otherwise}$$

For example consider a small sequence AGGCCTTAGGCTAAACT. The corresponding four indicator sequences are given in Table 1.

Four indicator sequences of gene BOVANPA (Bovine gastrin gene of *Bos taurus* – cow) of length 1769 base pairs is given as an example in Fig. 1

In DSP, many tools are available for signal processing like Discrete Fourier Transform (DFT) and Discrete Wavelet Transform (DWT) [9]. A measure of total spectrum content U(k) of the DNA sequence [7] at frequency k is given by the quantity:

$$U(k) = |U_A(k)|^2 + |U_G(k)|^2 + |U_C(k)|^2 + |U_T(k)|^2$$

where $U_A = \text{DFT}(S_A)$, $U_G = \text{DFT}(S_G)$, $U_C = \text{DFT}(S_C)$, $U_T = \text{DFT}(S_T)$.

It has been shown that a protein coding region in DNA typically has a peak at the frequency $k = N/3$, where N is the length of the sequence [15–17]. In the next section a new mapping technique based on Categorical Periodogram is defined.

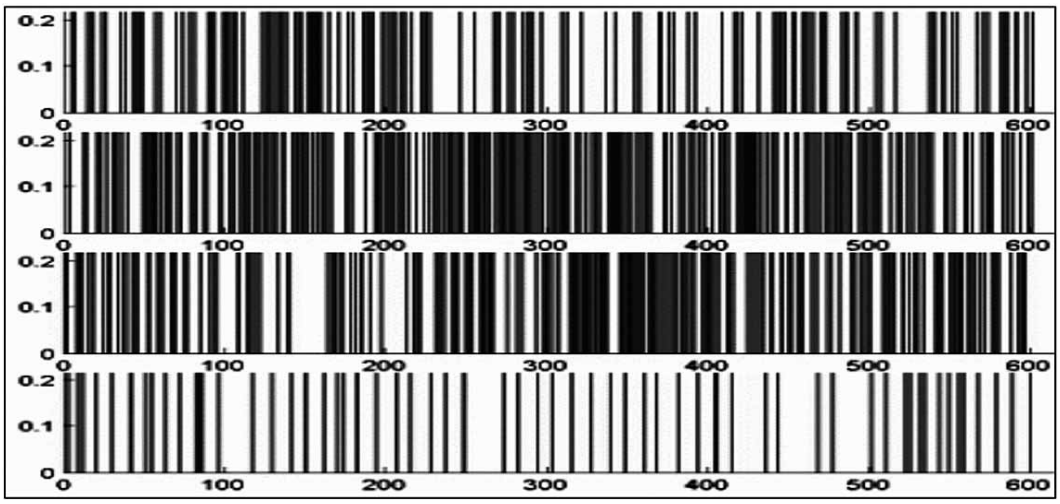


Fig. 1. Indicator sequences corresponding to IA, IG, IC and IT of BOVANPA.

CATEGORICAL PERIODOGRAM SEQUENCE

A categorical element is an element which has the possible measure or assigned values consisting of a discrete set of categories [18]. That is, categorical data are identified by category rather than numerical value [19]. Hence such a data set can be subdivided into finite subsets. DNA sequences can be considered as categorical data which can be subdivided into four subsets depending on the four symbols in it.

A new mapping technique, Cumulative Categorical Periodogram (CCP), is now defined for any categorical sequence. CCP(i), $1 \leqslant i \leqslant N$ for a sequence of length N is the number of occurrences of cycles with period i. In the context of categorical time series, a cycle is said to be achieved when it returns to a previously encountered state and the period of the cycle is defined as the number of intervening events plus one. In other words

$$\mathrm{CCP}(i) = \sum_{j=i+1}^{N} \begin{cases} 1 \text{ if } S(j) = S(j-i) \text{ and for } i < k < j-i, S(k) \neq S(i) \\ 0 \text{ otherwise} \end{cases}$$

Consider a sequence $S =$ ACAAACC. Cumulative Categorical Periodogram of S is calculated as follows.

$$\begin{aligned} \mathrm{CCP}(1) &= \text{ No. of occurrences of cycle with period 1} \\ &= \text{No. of occurrences of cycle with zero intervening event} \\ &= 3 \text{ (see Fig. 2)} \end{aligned}$$

Fig. 2. Computation of CCP(1).

$\mathrm{CCP}(2) =$ No. of occurrence of cycle with period 2

$=$ No. of occurrences of cycle with one intervening event

$= 1$ (see Fig. 3)

Fig. 3. Computation of CCP(2).

$\mathrm{CCP}(3) =$ No. of occurrence of cycle with period 3

$=$ No. of occurrences of cycle with two intervening event

$= 0$

$\mathrm{CCP}(4) =$ No. of occurrence of cycle with period 4

$=$ No. of occurrences of cycle with three intervening events

$= 1$ (see Fig. 4)

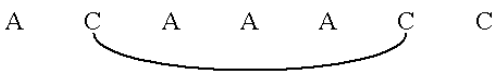


Fig. 4. Computation of CCP(4).

Similarly CCP(5) = CCP(6) = CCP(7) = 0

In short, CCP(k) measures the existence of pairs of identical elements at a distance of k base pairs [20]. In a DNA sequence, as there are 4 categorical elements, cycles of various periods between A & A, C & C, G & G and T & T emerge. In the case of $S =$ ACAAACC, the cycle counts are shown in Table 2.

The authors have investigated the question as to whether the CCP is the sequence auto correlation itself. It is well known that the auto correlation measure of a sequence $\{x_n\}$ of length N is generally described as:

$$y(k) = \sum_{i=0}^{N} x(i) * x(i+k), \quad k = 0, 1, \ldots, N-1$$

If we consider auto correlation of an indicator sequence, as all elements are 1 or 0, let us investigate if they boil down to the same result. To enable a simple comparison, let us consider a sequence segment with a single nucleotide repeating itself across the sequence, such as $S =$ AAAAAAA. The indicator sequences S_G, S_C and S_T will all be [0 0 0 0 0 0 0] and hence we need to consider a single sequence $S_A(k) =$ [1 1 1 1 1 1 1] alone. Let us compare them with the CCP of this sequence which is trivial to calculate, as CCP(k) = [7 0 0 0 0 0 0]. They are obviously different and their auto correlation also can be computed as different.

$$S_A(k) = \textit{[1 1 1 1 1 1 1]}$$

$$\mathrm{CCP}(k) = \textit{[7 0 0 0 0 0 0]}$$

$$S_A(k) * S_A(k) = \textit{[1 2 3 4 5 6 7 8 7 6 5 4 3 2 1]}$$

$$\mathrm{CCP}(k) * \mathrm{CCP}(k) = \textit{[0 0 0 0 0 0 49 0 0 0 0 0 0]}$$

Table 2
Cycle counts and CCP for $S =$ ACAAACC

i(Period)	Cycle count between A ↔ A	Cycle count between C ↔ C	Cycle count between G ↔ G	Cycle count between T ↔ T	Cumulative cycle count
1	2	1	0	0	3
2	1	0	0	0	1
3	0	0	0	0	0
4	0	1	0	0	1
5	0	0	0	0	0
6	0	0	0	0	0
7	0	0	0	0	0

Thus we conclude that CCP is a unique representation of the DNA sequence and hence it is further investigated. The CCPs of a set of genome sequence where studied with the view of extracting useful signatures for characterizing them.

The authors have experimented with various measures of the categorical periodogram of DNA sequences and found that the Fourier analysis gives striking signatures. The total spectral content of a CCP of a DNA sequence is given by the quantity

$$C(k) = \sum_{n=0}^{N-1} \mathrm{CCP}(n) e^{-j2\pi kn/N}, k = 0, 1, 2, \ldots, N-1$$

A study of the plot of the reciprocal of spectrum of CCP of various genes revealed that they indicate presence of protein coding by means of prominent peaks which very closely matched the classical $N/3$ spectral peak in indicator sequences reported in [15–17]. The specific results are given in the next section.

RESULTS

A set of 10 eukaryotic genes was selected at random from the dataset of Burset and Guigo (http://genome.imim.es/databases/genomics96). The exon areas of these genes were combined to form a new sequence equivalent to the spliced mRNA, as is seen done in the earlier works [8,15,16]. Both the indicator sequence mapping and CCP mapping were applied to these sequences. The figures in Table 3 show the spectral plot corresponding to each of these methods. For comparing the results, both the results have been normalized.

Table 4 contains a comparison of the results obtained from the above observations. The last column of the table clearly indicates that the spectral signatures in the two sequences are strikingly close. This leads us to conclude that the CCP mapping is a useful DNA representation in so far as traditional indicator sequence analyses have been useful. Further, in the case of CCP, there is no longer the computational overload of handling 4 subsequences determining their Fourier transforms, as the CCP is a single sequence. The biological significance of the equivalence is, at this point of time, unknown to the authors and requires further research.

It is well known that in certain prokaryotes, the period-3 component is absent in coding regions. To investigate the corresponding situation in the case of CCP, a set of 5 prokaryotic genes was selected at random from NCBI database. To these sequences both indicator sequence mapping and CCP mapping were applied. The figures in Table 5 show the normalized spectral plot corresponding to each of these methods. Table 6 shows the results obtained from these observations.

Table 3
Spectral content of genes obtained from indicator sequence and CCP mapping techniques

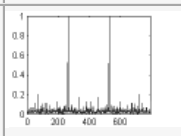
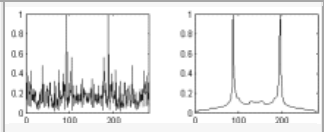
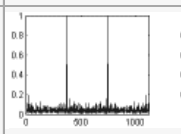
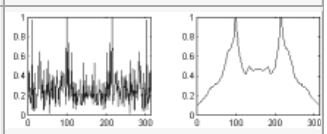
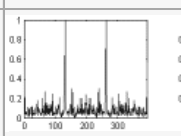
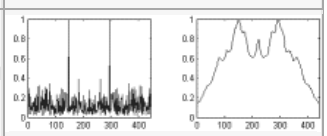
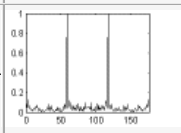
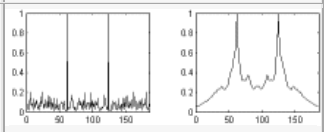
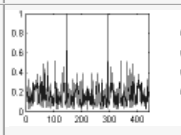
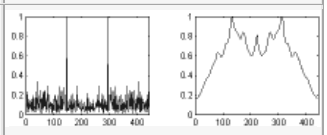
Gene	Indicator sequence (left) and CCP mapping technique (right)	Gene	Indicator sequence (left) and CCP mapping technique (right)
PIGAPAI		GPINCP11A	
AF135499		BOVGAS	
BTTN2PG		CALEGLOBIN	
DMPROTP1		OAMITII	
MACHBGAI		LAYEGLOBIN	

Table 4
Comparison of position values of eukaryotic genes spectral content obtained using CCP mapping and indicator sequence mapping

Gene Id	Length $= N$	CCP Signal Spectrum		Indicator Sequence Spectrum		$N/3$	Difference between positions obtained by both methods methods (CCP-Indicator sequence mapping)
		Maximum Value	Position	Maximum Value	Position		
PIGAPAI	798	1	262	1	266	266	-4
GPINCP1AA	282	1	88	1	93	94	-6
AF135499	1128	1	380	1	376	376	4
BOVGAS	315	1	100	1	99	105	-5
BTTNP2G	396	1	130	1	132	132	-2
CALEGLOBIM	444	1	151	1	148	148	3
DMPROTP1	177	1	56	1	59	59	-3
OAMTII	186	1	62	1	62	62	0
MACHBGA1	444	1	150	1	148	148	2
LAYEGLOBIN	444	1	133	1	148	148	-15

It is observed that in the case of prokaryotic genes the result obtained using indicator sequence mapping and CCP mapping are similar. Both simultaneously fail to highlight coding exons through $N/3$ peaks.

Table 5
Application of indicator sequence mapping and CCP mapping to prokaryotic genes lacking period-3 component

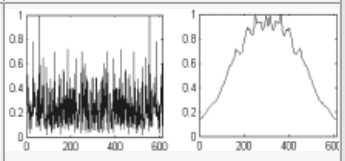
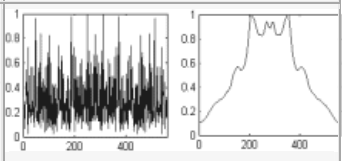
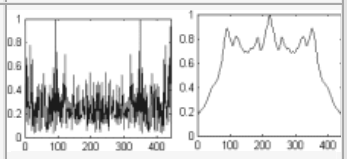
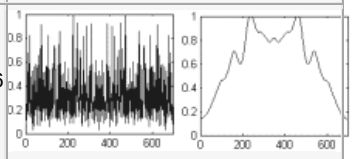
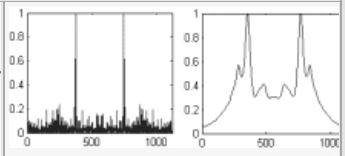
Gene	Indicator sequence (left) and CCP mapping technique (right)	Gene	Indicator sequence (left) and CCP mapping technique (right)
BQ297596		BM085626	
BF599063		DQ338566	
AAPD1000017			

Table 6
Comparison of position values of prokaryotic genes spectral content obtained using CCP mapping and indicator sequence mapping

Gene ID(1)	Length = N(2)	CCP Signal Spectrum Position(3)	Indicator Sequence Spectrum Position(4)	$N/3$(5)	Difference of columns 3 and 4(6)
BQ297576	611	250	60	203	NA
BM085626	560	208	187	186	21
BF599063	442	221	94	147	NA
DQ338566	699	239	239	233	0
AAPD1000017	1118	353	373	372	−20

CONCLUSION

This paper has reported the use of a novel numeric representation of DNA sequences using the concept of categorical periodograms. It has been shown that spectral analysis of the categorical periodogram signals of DNA show prominent peaks which are shown to be about exactly coinciding with the well-known $N/3$ peaks in the spectrum of indicator sequence of DNA. As the categorical periodogram is a single signal representation of DNA sequences (compared to 4 signals in the established indicator sequence representation), due to the effectively identical gene coding signature, then CCP is a better candidate for gene identification. The CCP is to be viewed as a step forward in identifying the ideal mapping of DNA to numerical sequences. The biological significance of the functional equivalence of CCP and indicator sequences of DNA, however, remains to be investigated.

REFERENCES

[1] Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. Biometrika **73**, 13-22.
[2] Abrams, M., Doraswamy, N. and Mathur, A. (1992). Chitra: Visual analysis of parallel and distributed programs in the time, event, and frequency domain. IEEE Trans. Parallel Distrib. Syst. **3**, 672-685.
[3] Anastassiou, D. (2000). Digital Signal Processing of biomolecular sequences. Technical report, Dept. of EE, Columbia University, 2000-20-041.
[4] Anastassiou, D. (2001). Genomic Signal Processing. IEEE Signal Processing Magazine **18**(**4**), 8-20.
[5] Mitra, S. K. (1998). Digital Signal Processing: A Computer Based Approach. McGraw-Hill.
[6] Fickett, J. W. and Tung, C. S. (1992). Assessment of Protein coding measures. Nucleic Acids Res. **20**, 6441-6450.
[7] Anastassiou, D. (2000). Frequency-domain analysis of bimolecular sequences. Bioinformatics **16**, 1073-1081.
[8] Vaidyanathan, P. P. and Yoon, B. J. (2002). Gene and exon prediction using allpass-based filters. *In:* Workshop on Genomic Signal Processing and Statistics (GENSIPS), Raleigh, NC.
[9] Pirogova, E., Fang, Q., Akay, M. and Cosic, I. (2002). Investigation of the structural and functional relationship of oncogene proteins. Proceedings of IEEE **90**, 1859-1866.
[10] Nair, A. S. and Mahalakshmi, T. (2005). Visualization of genomic data using inter-nucleotide distance signals. GSP 2005 -International Conference on Genomic Signal Processing, Romania.
[11] Mahalakshmi, T. and Nair, A. S. (2005). GSP using Binucleotide Distance Signals. ADCOM 2005 at Coimbatore, India.
[12] Cristea P. D. (2002). Conversion of nucleotides sequences into genomic signals. J. Cell. Mol. Med. **6**, 279-303.
[13] Cristea P. D. (2001). Genetic Signals: An emerging concept. Proceedings of IWSSIP, 17-22.
[14] Daffin, S. and Asif, A. (2005). A Fast DFT Based Gene prediction algorithm for identification of protein coding regions. *In*: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '05, Philadelphia, PA, vol. 3, pp. 113-116.
[15] Tiwari, S., Ramachandran, S., Bhattacharya, A., Bhattacharya, S. and Ramaswamy, R. (1997). Prediction of probable genes by Fourier analysis of genomic sequences. Comput. Appl. Biosci. **13**, 263-270.
[16] Fickett, J. W. (1982). Recognition of protein coding regions in DNA sequences. Nucleic Acids Res. **10**, 5303-5318.
[17] Chechetkin, V. R. and Turygin, A. Y. (1995). Size-dependence of three-periodicity and long-range correlations in DNA sequences. Phys. Lett. A **199**, 75-80.
[18] Friendly, M. (2000). Visualizing Categorical Data. SAS Institute, Cary, NC.
[19] Ribler, R., Mathur, A. and Abrams, M. (1995). Visualizing and Modeling Categorical Time Sequence Data. Virginia Polytechnic Institute and State University. *In*: ICASE/LaRC Symposium on Visualizing TimeVarying Data, Williamsburg, VA, pp. 3-19.
[20] Kumar, L., Futschik, M. and Herzel, H. (2006). DNA motifs and sequence periodicities. In Silico Biol. **6**, 0008.