An Analogy between Baseband On-Off Keying and DNA Time-Series Modeling

Efi K. Manopoulou, *Student Member, IEEE,* and Stavros S. Kotsopoulos, *Member, IEEE*

Abstract **— This paper focuses the analogy between baseband On-Off Keying (OOK) modulation, used in telecommunications, and the time–series modeling of the DeoxyriboNucleic Acid (DNA), which is one of the pillars of the current biotechnology revolution. Applying the well-known performance evaluation tools of the Gaussian pulse OOK, important characteristics as the error rate of the DNA sequencing can be studied.**

Keywords **— On-Off Keying (OOK), DeoxyriboNucleic Acid (DNA) time-series, error rate.**

I. Introduction

LAST years many authors pointed out several analogies between biological processes in genetic systems and corresponding ones in communications systems [1]-[5]. Most of them proposed well-known techniques in communications theory as tools to resolve problems in genetic systems. One of them is the modeling of the DNA time series, which can be efficiently used to automatically recover the base sequence in DeoxyriboNucleic Acid (DNA) sequencing. This modeling fundamentally depends on the underlying statistics of the DNA electrophoresis time series. In [5] a formal statistical model of the DNA time series is presented and then it is used to construct the optimal maximum likelihood (ML) processor, for the DNA sequencing recovery. Moreover, in the same work the derived DNA-ML algorithm features Kalman prediction of peak locations, peak parameter estimation, whitened waveform comparison and multiple hypotheses processing using the M-algorithm. Properties of the algorithm are also examined using both simulated and real data.

In this paper, we point out the existed analogy between the baseband On-Off Keying (OOK) modulation, used in telecommunications, and the DNA time series modeling. Using the well-known performance characteristics of the OOK, we can conclude on the errors which occurred in the DNA time series sequencing.

The rest of this paper is organized as follows: In the next section the baseband OOK and the DNA time-series characteristics are presented and discussed. In Section, III we focus on their analogy and Section IV presents some concluding remarks.

II. Baseband OOK and DNA Time-Series

A. Baseband OOK

OOK is a well-known orthogonal modulation technique widely used in optical and ultra wideband communications [6]-[10]. OOK, or otherwise known as unipolar signaling in the analog baseband world, is a simple pulse modulation technique where a pulse is transmitted to represent a binary "1", while no pulse is transmitted for a binary "0". The baseband representation, which is illustrated in Fig. 1, of the transmitted signal is

$$w(t)=\sum_{j\quad\infty}^{\infty} b_j s\left(t-jT_f\right) \qquad (1)$$

where:

$w(t)$ is the transmitted UWB signal

b_j are the data bits 0 or 1

$s(t)$ is the pulse shape

T_f is the time frame period

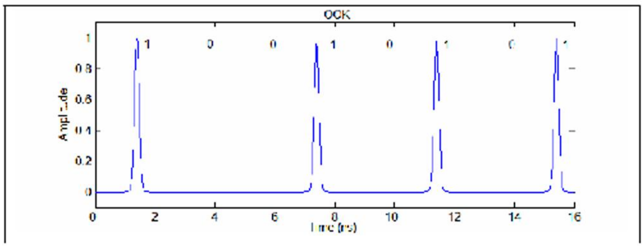

Fig. 1: A Baseband OOK with Gaussian pulses

The most popular pulse shape for OOK communication systems is the Gaussian pulse due to mathematical convenience and ease of generation, which is illustrated in Figure 2 and can be described by:

$$s(t)=Ae^{\frac{(t\quad m)^2}{2\sigma^2}} \qquad (2)$$

where:

A is the pulse amplitude

T is the time

σ is the standard deviation of the Gaussian pulse, and

m is the mean value of the Gaussian pulse.

The authors are with Department of Electrical and Computer Engineering, University of Patras, Patras, 26110, Greece.
E-mail: emanopoulou@yahoo.gr, kotsop@ee.upatras.gr.

1-4244-0049-X/05/$20.00 ©2005 IEEE

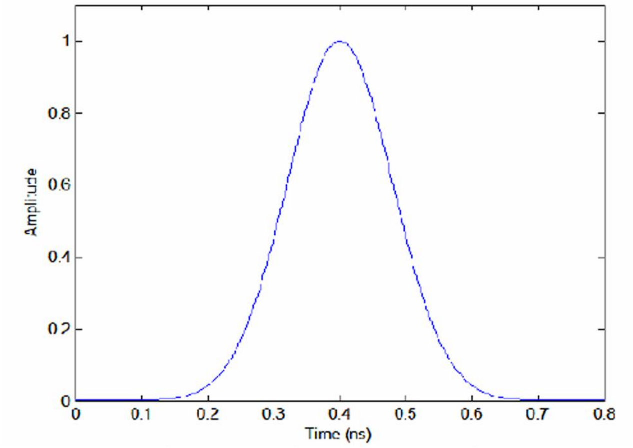

Fig. 2: Gaussian Pulse Example

One obvious advantage to using OOK is the simplicity of the physical implementation, as one pulse generator is necessary, as opposed to two, as is the case with biphase modulation. A single RF switch can control the transmitted pulses by switching on for a "1" data bit and off for a "0" data bit. This effortless transmitter configuration makes OOK popular for less complex communications systems.

The average bit error rate (BER) of the OOK in a Gaussian noise channel when a matched filter receiver is used, is [6]

$$P_e = Q\left(\sqrt{\frac{E_b}{N_0}}\right) \tag{3}$$

where:

Q is the well-known Q-function [6] defined as

$$Q(z) \triangleq \frac{1}{2\pi}\int_z^\infty e^{\frac{x^2}{2}} dx \ ,$$

E_b is the average energy per bit, and

N_0 is the noise power spectral density.

In Fig. 3, the BER performance of the OOK is presented, for several values of the E_b / N_0 .

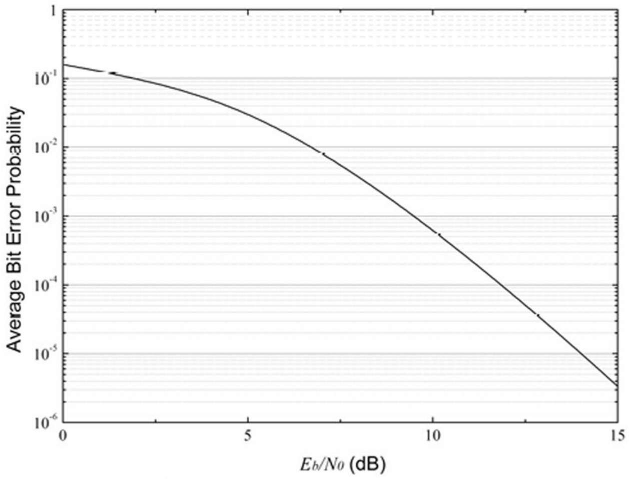

Fig. 3: OOK BER performance

B. DNA Time-Series

Next, we give useful information on the DNA time-series. A representative sample of DNA electrophoresis sensor data from a Pharmacia automated luminescence fluorescent (ALF) DNA sequencer is shown in Fig. 4. This data has had certain large scale trends removed and has been normalized so all channels have approximately the same signal strength. The four channels, corresponding to the four base types: Adenine (A), Cytosine (C), Guanine (G), Thymine (T), have been offset in Fig. 2 to facilitate easier reading.

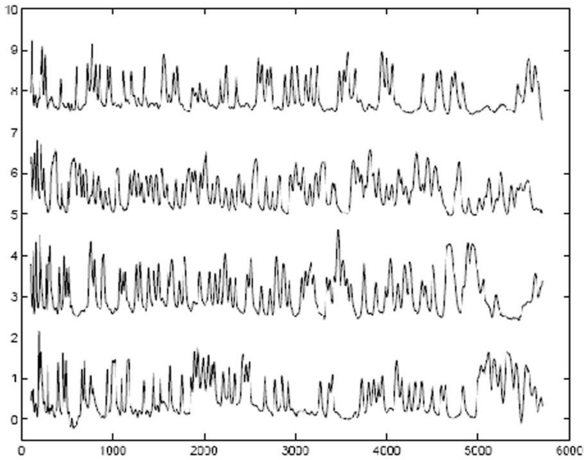

Fig. 4: Selected compensated DNA time series. Top curve is for A channel with C, G and T channels presented in order from top [5]

Fig. 5 is a higher resolution presentation of the compensated time-series for all four bases, is presented. Note, that the individual peaks are of similar shape and that there is evidence of noise. This suggests a time-series model as in

$$\mathrm{y}_{n,k} = \sum_{i\ 1}^{N_b} \alpha_i g_{k,t_i} \delta_{n,x_i} + n_k \tag{4}$$

where n refers to the base type, k is the sample number, the sum is over the base sequence position, i, and there are a total of N_b bases in the sequence. Moreover, the Kronecker delta function, δ_{n,x_i} , is defined as one if n is the same as x_i and zero otherwise.

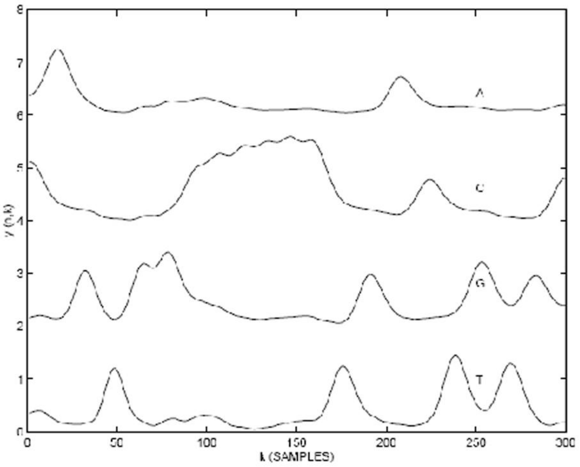

Fig. 5: Sample DNA time-series of Fig. 2 [5].

The g_{k,t_i} represents a generic pulse shape, where the peak of the pulse is centered on t_i and the peak is scaled by α_i. Finally, an additive noise process, n_k, it is evident in Figure 3.

The sequence in Fig. 5 is read from left to right by identifying which channel has a signal peak, as: AGTGGCCCCCCTGACTGTG. Sequence position refers to the number of bases from the start of the sequence up to and including the base whose sequence position is being evaluated. Thus, the first "T" in the sequence is in sequence position 3.

III. The Analogy: DNA Base Bit Sequences

Looking at Fig. 5, the existence of a pulse in the time-series of every base means the existence of the corresponding base in the DNA sequence. Note, that here there are roughly 15 time samples between each signal peak. Also, it is clear from that the pulse associated with each base extends over many samples and interferes with adjacent pulses.

Now, if we translate the existence of a pulse (which it is also means the existence of the base) as one (1) and the non-existence (background noise) as zero (0), then the times series for every base can be represented by a "bit sequence". Per example, using the data of Fig. 2 the bit sequences for the four bases are

Table 1: Base Bit Sequences in Fig. 3

Base	Bit Sequence
A	1000000000000100000
C	0000011111100010000
G	0101100000001000101
T	0010000000010001010

Using this approach the Base Bit sequence can be apparently considered (see Figs. 1 and 5) as a bit sequence of an OOK modulation system.

Concerning, the additive noise, n_k in eq. (4) is assumed to be Gaussian uncorrelated, due to physical phenomena, such as integrated sensor shot-noise and pre-amplifier thermal noise [5]. Under this assumption the average error rate in the DNA sequencing can be approximated using the expression (3) for the corresponding OOK process.

IV. Conclusion

We pointed out the similarities between baseband OOK modulation, used in telecommunications, and the time series modeling of the DNA. The proposed analogy can be used to study the average error rate of the bases time-series and consequently for the DNA sequencing. Additive white Gaussian noise was assumed. There are several open issues for further investigation: a) What happened to the noise component when the chemical noise is also taken in to account, b) What is the total error effect on the DNA sequencing when several errors occur (base substitution, base insertion, etc.)

References

[1] H. Yockey, *Information theory and molecular biology*. Cambridge:Cambridge University Press, 1992.

[2] W.K. T. Schneider, "Theory of molecular machines I. Channel capacity of molecular machines," *Journal Theoretical Biology*, vol. 148, pp. 83 123, 1991.

[3] M. Eigen, "The origin of genetic information: viruses as models," *Gene*, vol. 135, pp. 37 47, 1993.

[4] Z. Dawy, F. Morcos Gonzalez, J. Hagenauer, and J. C. Mueller, "Modeling and Analysis of Gene Expression Mechanisms:A Communication Theory Approach," *in Proc. of IEEE ICC'05*, Korea, May 2005.

[5] S. W. Davies, "Application of Communication Theory to Automatic DNA Sequencing," Ph.D. Thesis, Graduate Department of Electrical and Computer Engineering, University of Toronto, 1999.

[6] J. Proakis, *Digital Communications*, 4th edition, Mc GrawHill publications, 2002.

[7] S. Roy, J. R. Forester, V. S. Somayazulu, and D. G. Leeper, "Ultrawideband radio design: The promise of high-speed, short-range wireless connectivity," *Proc. IEEE*, vol. 92, no. 2, pp. 295 311, Feb. 2004.

[8] M. Z. Win and R. A. Scholtz, "Ultra-wide bandwidth time-hopping spread-spectrum impulse radio for wireless multipleaccess communications," *IEEE Trans. Commun.*, vol. 48, no. 4, pp. 679 691, Apr. 2000.

[9] M. Z. Win and R. A. Scholtz, "Impulse radio: How it works," *Commun. Lett.*, vol. 2, no. 2, pp. 36 38, Feb. 1998.

[10] R. A. Scholtz, "Multiple access with time hopping impulse modulation," in *Proc. IEEE Military Commun. Conf.*, Oct. 1993, Bedford, MA, pp. 447 450.