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

\[ w(t) = \sum_{j=}^{p} b_j s(t - jT_f) \]  

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:

\[ s(t) = A e^{\frac{-(t - m)^2}{2\sigma^2}} \]  

where:
\[ A \] is the pulse amplitude
\[ T \] is the time
\[ \sigma \] is the standard deviation of the Gaussian pulse, and
\[ m \] is the mean value of the Gaussian pulse.

Authorized licensed use limited to: Illinois Institute of Technology. Downloaded on June 4, 2009 at 16:00 from IEEE Xplore. Restrictions apply.
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.

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

$$P_2 = Q \left( \frac{\sqrt{E_b}}{N_0} \right)$$

(3)

where:

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

$$Q(z) = \frac{1}{\sqrt{2\pi}} \int_{z}^{\infty} e^{-x^2/2} dx,$$

- $E_b$ 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 $E_b/N_0$.

**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.

In Fig. 5, 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

$$y_{nk} = \sum_{i=1}^{N_b} a_i \delta_{nk} \delta_{ni} + n_k$$

(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, $\delta_{nk} \delta_{ni}$, is defined as one if $n$ is the same as $x_i$ and zero otherwise.
The $g_{x,k}$ represents a generic pulse shape, where the peak of the pulse is centered on $t_i$ and the peak is scaled by $\alpha_i$. Finally, an additive noise process, $n_i$, 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: AGTGGCCCGCTGACTGTTG. 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 we 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>
<thead>
<tr>
<th>Base</th>
<th>Bit Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1000000000000100000</td>
</tr>
<tr>
<td>C</td>
<td>0000011111100010000</td>
</tr>
<tr>
<td>G</td>
<td>0101100000001000101</td>
</tr>
<tr>
<td>T</td>
<td>0010000000100010100</td>
</tr>
</tbody>
</table>

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, $n_i$ 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 base 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 into account, b) What is the total error effect on the DNA sequencing when several errors occur (base substitution, base insertion, etc.)

REFERENCES