This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the ICC 2007 proceedings.

Applying Techniques from Frame Synchronization for Biological Sequence Analysis

Johanna Weindl and Joachim Hagenauer
Institute for Communications Engineering (LNT)
Technical University of Munich (TUM)
Arcisstr. 21, 80290 Munich, Germany
Email: jweindl@tum.de, hagenauer@tum.de

***Abstract*— During the last years, the amount of genetic data available has increased rapidly which creates the demand for advanced techniques for their interpretation. In the following, we present an approach of applying communication theory for biological sequence analysis. We use an analogy to frame synchronization to gain more insights into transcription, the step of copying a gene into messenger RNA (mRNA). In continuous and packet data transmission, successful decoding of a transmitted data stream at the receiver side strongly depends on the choice of the sync word that indicates the beginning of the message and thus needs to be detected reliably. Analogously, biological sync words indicate the beginning of a gene, i.e. they mark the sequence in the DNA that needs to be copied during transcription. These biological sync words are the -35 promoter region and the -10 promoter region named after their approximate position before the gene. In digital data transmission, the sync word is selected from all possible patterns based on its autocorrelation behavior. Therefore, we use an adapted autocorrelation function to investigate the synchronization properties of the promoter regions revealing that the -35 region is an outstanding synchronization pattern. In contrast to that, the -10 region, though more important for transcription initiation, showed to have worse properties. However, when including sequence constraints imposed through the region's importance for transcription, the -10 region showed to be among the best possible sequences, too. These facts imply that during evolution promoter sequences evolved in a way to optimize their synchronization properties.**

I. Introduction

Surprisingly to biologists, communication and information theory proved to provide powerful tools for the analysis of processes in molecular biology and genetics [1]–[3]. An up-to-date summary of ongoing research can be found in [4]. The genetic information of an organism is stored in the DNA, which can be seen as a digital signal of the quaternary alphabet of nucleotides $\mathcal{A} = \{A, C, G, T\}$. An important field of interest is gene expression, the process during which this information stored in the DNA is transformed into proteins which are responsible for cell functions like oxygen transport etc.. Gene expression in bacteria takes place in two steps: transcription and translation (see Fig. 1).

During transcription, specific parts of the DNA - the genes - serve as a template for construction of the messenger RNA (mRNA). During translation, the mRNA is transformed into a chain of amino acids that forms a protein. Both steps are performed by proteins that interact with the DNA and mRNA, respectively. In the first step of the transcription cycle, the protein RNA polymerase (RNAP) randomly binds to the DNA double helix and slides along it until its sigma subunit detects the two promoter regions that indicate the proximate beginning of the gene [5]. These regions are located around 35 and 10 base pairs before the transcription start site and are separated by a variable spacing of 15 to 21 base pairs (see Fig. 2). The sigma factor is a kidney-shaped protein that confers sequence-specificity on the RNAP in order to enable detection of the promoter regions [6]. Different sigma factors are available in the cell and regulate transcription initiation of specific sets of genes. In our research, we primarily focus on the main sigma factor (σ^{70}), responsible for transcription of housekeeping genes. For more information about transcription initiation see e.g. [7]. In the following, we present a communication theory approach for the analysis of transcription initiation in the bacterium *Escherichia coli* (*E.coli*).

Fig. 1. The process of protein synthesis (gene expression).

II. Frame Synchronization in Digital Communication Systems

Frame synchronization is an essential problem of data transmission in all digital communication systems. It refers to localizing the beginning of a message in received data, i.e. to *"the correct association at the receiver of the received symbols to blocks such as words, bytes or data-frames"* [8]. For this purpose, a fixed binary or quaternary pattern (depending on the modulation scheme), known to both the transmitter and receiver, is inserted into the data stream indicating the

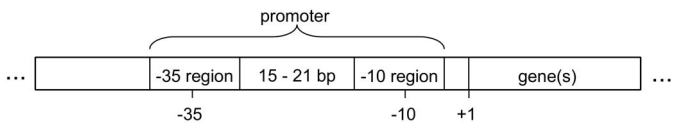

Fig. 2. Structure of bacterial promoters.

1-4244-0353-7/07/$25.00 ©2007 IEEE

beginning of a message. This method of inserting markers (therefore known as "marker concept") was first introduced by R.H. Barker in 1953 [9] and further investigated by J.L. Massey [10]. Since then, researchers have addressed the design and reliable detection of these so-called sync words for both asynchronous packet transmission and continuous transmission (see e.g. [8]).

A. Analogy to Transcription Initiation

All approaches of frame synchronization are based on a correlator that compares the known sync word with the incoming data stream for each position and decides for the position with the highest correlation. This correlation detection was shown in [10] to be suboptimal, however, is widely used. During transcription initiation - the first step of gene expression - the RNA polymerase and its sigma subunit need to recognize two promoter regions that indicate the beginning of the gene (see Section I). Hence, the process of transcription can be considered as a frame synchronization with two sync words surrounded by random data (see Fig. 3). Since the distance between two consecutive promoters varies, it precisely corresponds to the detection of aperiodically inserted sync words. The sliding process of the protein along the DNA represents the correlation taking place in technical communication systems. Detection of promoters during transcription initiation is (among other factors) based on the binding energy between the sigma factor and the DNA, i.e. high binding energies indicate promoter regions [11]. Table I summarizes the comparison of digital communication systems and transcription initiation.

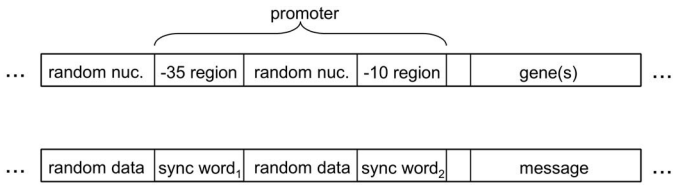

Fig. 3. Analogy between bacterial promoter regions and sync words.

B. Choice of the Sync Words

The sync words in digital data transmission have to be chosen such that they satisfy the following two conditions [9]:

- The probability of a random occurrence of the pattern in the data stream is minimized.
- The structure of the pattern is such that the neighboring symbols cannot yield a shifted sync word, i.e. the pattern should not contain periodicities. Considering e.g. the (binary) pattern $+1+1+1+1+1+1$, the probability is 0.5 that it is followed by a $+1$ (in case of equally probable symbols) which may lead to a shifted synchronization.

While the probability of a random occurrence (first condition) does not depend on the sequence in case of independent symbols, the second condition is to be analyzed

TABLE I

COMPARISON OF FRAME SYNCHRONIZATION AND BACTERIAL TRANSCRIPTION INITIATION.

	digital communication	transcription initiation
data	(generally) binary or quaternary data stream	quaternary DNA sequence
marker	(generally) binary or quaternary sync word	two quaternary promoter regions
device	correlator	sigma subunit of RNAP
parameter	correlation between sync word and data	binding energy between sigma factor and DNA

using the aperiodic autocorrelation function $\varphi_{ss}(\tau)$ of the sync word. $\varphi_{ss}(\tau)$ describes the similarity of a sequence $s = \{s_1, s_2, ..., s_N\}$ surrounded by random data to itself for every shift $\tau \in [-(N-1); +(N-1)]$:

$$\varphi_{ss}(\tau) = \sum_{m=1}^{N-|\tau|} s_m \cdot s^*_{m+|\tau|}, \tag{1}$$

where s^*_m denotes the complex conjugate of s_m and N being the length of the sequence (see e.g. [12] for more information about the autocorrelation function). Since we consider the aperiodic autocorrelation function, the surrounding of the sequence s is assumed to be random and uniformly distributed over the symbol alphabet. Thus, for $|\tau| > N-1$, the autocorrelation function $\varphi_{ss}(\tau)$ represents the expected value (i.e. usually equals to zero). For reasons of clarifying the application to biological sequences in subsequent sections, (1) can be rewritten as

$$\varphi_{ss}(\tau) = \sum_{m=1}^{N-|\tau|} D(s_m, s^*_{m+|\tau|}), \tag{2}$$

where D denotes a matrix defining the multiplication of the elements s_m and $s^*_{m+|\tau|}$ with s_m indexing the rows of D and $s_{m+|\tau|}$ indexing the columns of D. In the binary, antipodal case, i.e. for $s_m \in \{-1; +1\}$, this yields in

$$s_{m+|\tau|} \rightarrow \quad +1 \quad -1 \qquad s_m \downarrow$$

$$D_{bin} = \begin{pmatrix} +1 & -1 \\ -1 & +1 \end{pmatrix}, \quad \begin{matrix} +1 \\ -1 \end{matrix}$$

and in the complex case, i.e. for $s_m \in \{+1, +i, -i, -1\}$, it results in

$$D_{complex} = \begin{array}{c} s_{m+|\tau|} \rightarrow \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{cccc} +1 & +i & -i & -1 \end{array} \\ \left(\begin{array}{cccc} +1 & -i & +i & -1 \\ +i & +1 & -1 & -i \\ -i & -1 & +1 & +i \\ -1 & +i & -i & +1 \end{array} \right) \\ \\ \end{array} \begin{array}{c} s_m \downarrow \\ \\ +1 \\ +i \\ -i \\ -1 \end{array}.$$

In order to maintain the second above condition, the autocorrelation function of the sync word should have a narrow maximum at $\tau = 0$ and

- smallest possible values for $\tau \neq 0$ if unequivocal phase recovery after demodulation is guaranteed.
- smallest possible absolute values for $\tau \neq 0$ if phase ambiguities are expected after demodulation [13], i.e. the autocorrelation function should be as similar as possible to the Dirac delta function $\delta(t)$.

In the 1970s, the peak sidelobe level PSL was introduced as a measure of the synchronization properties of a sequence (also known as minimum peak sidelobe [14] or maximum sidelobe correlation [13]):

$$\text{PSL} = \max_{\tau \setminus \{0\}} |\varphi_{ss}(\tau)|. \tag{3}$$

It rates the impulse-type effect of non-ideal autocorrelation properties [15]. In the following, we focus on the assumption of correct phase recovery. In this case, the absolute values in (3) are omitted since negative values indicate strong dissimilarity and therefore minimize the probability of false synchronizations:

$$\text{PSL'} = \max_{\tau \setminus \{0\}} \varphi_{ss}(\tau). \tag{4}$$

Thus, the task in designing a reliable synchronization system lies in the search for sync words with smallest possible values of the peak sidelobe level, which was first addressed by R.H. Barker in 1953 who found several binary sequences up to length $N = 11$ with PSL ≤ 1 [9].

III. Applying Frame Synchronization to Biological Sequence Analysis

Our aim is to use the methods known from frame synchronization to rate the synchronization properties of DNA sequences. Therefore, we have to adapt the autocorrelation function to the quaternary alphabet of nucleotides $\mathcal{A} = \{A, C, G, T\}$ and redefine the product in (1) with respect to its biological meaning, i.e. in a way that it rates the effect of nucleotide matches and mismatches on the synchronization quality of the sequence. In order not to violate the properties of aperiodic autocorrelation functions,

$$\varphi_{ss}(0) = N, \tag{5}$$

$$\varphi_{ss}(\tau) = 0 \quad \forall \quad |\tau| > (N-1), \tag{6}$$

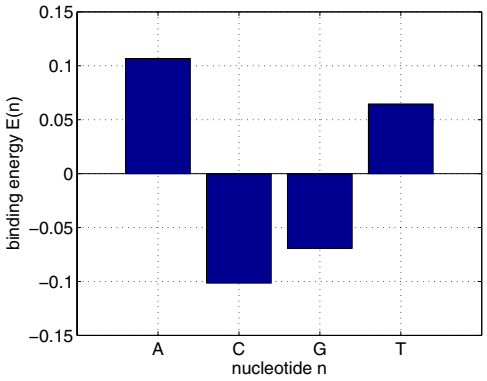

Fig. 4. Average contribution to binding energy between sigma factor and promoters in *E.coli*.

we rate an accordance of nucleotides by 1 and a divergence of nucleotides with a negative value such that mismatches are punished with an overall weight of -1. As mentioned in Section II-A, the binding energy decides about detection of the promoter regions (i.e. correct synchronization). This implies that if during autocorrelation shifted versions of the investigated DNA sequence yield high binding energies, they might cause shifted synchronizations. Thus, we have to relate the adapted autocorrelation to the binding energy of the shifted sequences. Therefore, the individual values rating mismatches are derived from the binding energy between sigma factor and DNA. In [11], Kiryu *et al.* calculated the effect of the nucleotides on the binding energy depending on its position in the promoters of the bacterium *Escherichia coli* (*E.coli*). Figure 4 shows the average effect of the 4 nucleotides on the binding energy.

It can be seen that the nucleotides `A` and `T` have on average a strengthening effect on the contact between sigma factor and DNA sequence ($E = 0.1067$ and $E = 0.0646$, respectively), whereas the nucleotides `C` ($E = -0.1016$) and `G` ($E = -0.0694$) make the contact loose. We rate a mismatch during autocorrelation by the absolute difference $|E(n_i) - E(n_j)|$ between the binding energies of nucleotides n_i and n_j, where $i, j \in \{1, 2, 3, 4\}$ index the nucleotides $n \in \mathcal{A} = \{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}\}$:

$$d(n_i, n_j) = \begin{cases} 1 & \text{for} \quad i = j \\ c \cdot |E(n_i) - E(n_j)| & \text{for} \quad i \neq j \end{cases}. \tag{7}$$

The constant c is still to be determined since in addition to reflecting the differences of binding energies, the values of $d(n_i, n_j)$ have to satisfy (6), i.e. the expected value $E\{d(n_i, n_j)\}$ has to be zero if assuming random and uniformly distributed nucleotides (see Section II-B):

$$E\{d(n_i, n_j)\} = 0, \tag{8}$$

which corresponds to

$$\sum_{\forall i,j} d(n_i, n_j) = 0, \quad (9)$$

$$\Rightarrow \underbrace{\sum_{\substack{i,j \\ i=j}} d(n_i, n_j)}_{\stackrel{(7)}{=}4} + \sum_{\substack{i,j \\ i\neq j}} d(n_i, n_j) = 0,$$

$$\Rightarrow \sum_{\substack{i,j \\ i=j}} d(n_i, n_j) \stackrel{!}{=} -\sum_{\substack{i,j \\ i\neq j}} d(n_i, n_j) = -4.$$

Equation (9) is fulfilled if scaling the individual energy differences according to (7) by the value

$$c = \frac{-4}{\sum\limits_{\substack{k,l \\ k\neq l}} |E(n_k) - E(n_l)|} = \frac{-4}{1.56} = -2.56. \quad (10)$$

In order to adapt the autocorrelation function to the quaternary alphabet of nucleotides and detection by the RNA polymerase and its sigma factor, we use (2) with a matrix D_{nuc} containing the values of $d(n_i, n_j)$, i.e. $D_{nuc}(s_m, s_{m+|\tau|}) = d(n_i = s_m, n_j = s_{m+|\tau|})$, which results for the presented case of *E.coli* promoters in

$$D_{nuc} = \begin{array}{c} \begin{array}{cccc} s_{m+|\tau|} \rightarrow & \texttt{A} & \texttt{C} & \texttt{G} \quad \texttt{T} \end{array} \\ \begin{pmatrix} 1 & -0.55 & -0.46 & -0.11 \\ -0.55 & 1 & -0.08 & -0.44 \\ -0.46 & -0.08 & 1 & -0.36 \\ -0.11 & -0.44 & -0.36 & 1 \end{pmatrix} \end{array} \begin{array}{c} s_m \downarrow \\ \texttt{A} \\ \texttt{C} \\ \texttt{G} \\ \texttt{T} \end{array}.$$

Therefore, the autocorrelation function of *E.coli* promoter sequences is expressed by

$$\widetilde{\varphi}_{ss}(\tau) = \sum_{m=1}^{N-|\tau|} D_{nuc}(s_m, s_{m+|\tau|}). \quad (11)$$

This adapted autocorrelation function allows us to evaluate the synchronization properties of promoter sequences. It has to be mentioned that the matrix values of D_{nuc} are calculated based on the data from [11], i.e. the adapted autocorrelation is based on the interaction between sigma factor and promoter regions in *E.coli* and is, therefore, specific for this biological synchronization process.

A. Autocorrelation Properties of E.coli Promoter Sequences

As mentioned before, transcription initiation corresponds to the process of synchronization used in digital data transmission, since two sync words - the promoter regions - need to be detected by the sigma factor. In order to gain more insights into promoter detection, we determine the synchronization properties of the -35 and the -10 promoter region in the bacterium *E.coli* by applying the adapted autocorrelation function. The consensus (i.e. most frequently detected) sequences are `TTGACA` for the -35 region and `TATAAT` for the -10 region, respectively (see e.g. [7]). Fig. 5 and Fig. 6 show the autocorrelation function calculated using (11). As mentioned in Section II-B, the autocorrelation function of sync words should have small and possibly negative sidelobes to minimize the probability of false synchronizations. This criteria seems to be well satisfied for the -35 region (Fig. 5), whereas the autocorrelation function of the -10 region (Fig. 6) has relatively high sidelobes for $|\tau| = 2$ and $|\tau| = 3$, which indicates periodicities in the sync word that may lead to shifted synchronizations. Calculation of the peak sidelobe level for both promoter regions according to (4) confirms this observation:

$$\text{PSL'}_{-35} = |\widetilde{\varphi}_{-35}(|\tau| = 2)| = 0.45,$$

$$\text{PSL'}_{-10} = |\widetilde{\varphi}_{-10}(|\tau| = 3)| = 1.89.$$

To rate the autocorrelation properties of the promoter sequences, we additionally calculate the values of PSL' for all $4^6 = 4096$ possible nucleotide sequences of length $N = 6$. The mean value and the standard deviation of the resulting values are listed in Table II. Fig. 7 shows the histogram of PSL' with the values of the -35 and -10 region highlighted by vertical lines. It can be seen that the value of the -35 promoter sequence is below average, whereas those of the -10 promoter sequence lies above the mean value. In fact, only 11.06 % of all possible sequences of length $N = 6$ have better autocorrelation properties with respect to the peak sidelobe level than the -35 region. Opposed to that, 75.39 % of all sequences have higher values of PSL' compared to the -10 region.

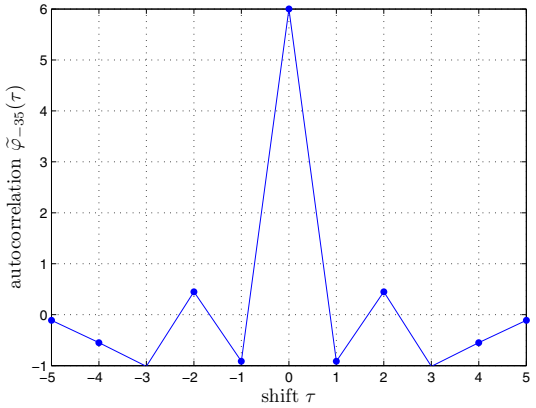

Fig. 5. Autocorrelation function of the -35 promoter region (sequence TTGACA, $N = 6$).

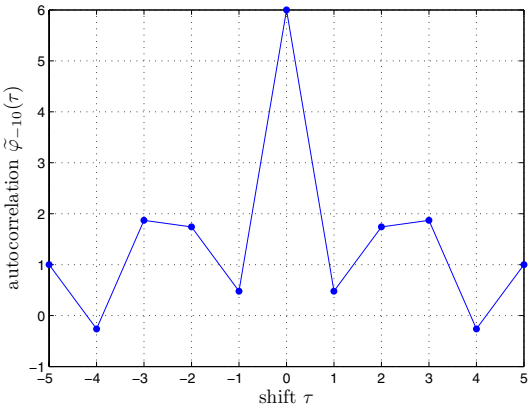

Fig. 6. Autocorrelation function of the -10 promoter region (sequence TATAAT, $N = 6$).

TABLE II

MEAN AND STANDARD DEVIATION OF PSL' FOR ALL POSSIBLE SEQUENCES OF $\mathcal{A} = \{A, C, G, T\}$ WITH LENGTH $N = 6$.

	mean	st. deviation
PSL'	1.32	0.77

B. Interpretation

The excellent PSL' value of the -35 region compared to those of the -10 region suggests that the synchronization takes place in two steps: firstly, the -35 region has to be detected out of all possible sequences with high accuracy to enable localization of the transcription start site (see Fig. 8, (A)). In the second step, the -10 region is detected, however, due to the synchronization conducted before, the sigma factor only needs to detect the -10 region out of around 7 sequences based on the shape and limited deformability of the sigma factor that allow a variable spacing of 15 to 21 base pairs between the two promoter regions (see Fig. 8, (B)). Therefore, the sequence of the -10 promoter region is less important for synchronization. This brings up the conclusion that the two promoters evolved to serve two tasks with different priorities and during different steps of transcription initiation: While the -35 region is indispensable for indicating the close-by transcription start site and, thus, needs to have excellent synchronization properties, the sequence and structure of the -10 region seems to play a more important role during later steps of transcription initiation. These steps may e.g. impose stronger constraints on the AT-richness (i.e. a high content of the nucleotides A and T) than on the sequence's detectability: The DNA double helix is easily opened and unwind in AT-rich regions which is necessary during transcription initiation [16]. Therefore, we assume that the AT-richness of the -10 region played a more important role during evolution than its synchronization properties and, thus, the latter evolved with lower priority.

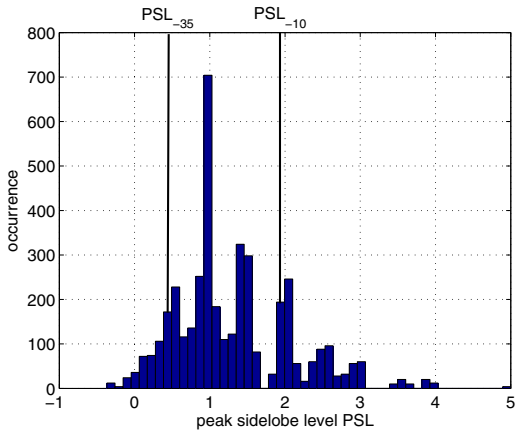

Fig. 7. Histogram of the peak sidelobe level for all possible sequences of $\mathcal{A} = \{A, C, G, T\}$ with length $N = 6$.

C. Synchronization properties of the -10 promoter region

In Section III-B, we hypothesized that the AT-richness might have been a stronger constraint on the evolution of the -10 promoter region than its synchronization properties, i.e. its structure with regard to periodicities and its ability to be distinguished from the random surrounding (see Section II-B). To corroborate this statement, we calculate the PSL' values for all $2^6 = 64$ possible nucleotide sequences of length $N = 6$ made up of only A and T. The mean value and the standard deviation of the resulting values are listed in Table III. Fig. 9 shows the histogram of PSL' for the considered sequences. Recalling the calculated value of the -10 region (PSL' $= 1.89$) shows clearly that it belongs to the sequences with highly below-average values if restricting the alphabet to $\mathcal{A}' = \{\mathtt{A}, \mathtt{T}\}$. In fact, no (!) other sequence of the 64 ones considered has a better PSL' value. This astonishing result strongly supports the conclusion that the bacterial promoter sequences evolved with respect to their synchronization properties: while the -35 region is an excellent synchronization pattern, the -10 region seems to constitute a good trade-off between the AT-richness required for DNA opening / unwinding and the sequence's detectability.

IV. CONCLUSIONS

We presented an application of communication theory for biological sequence analysis. Making use of the analogy between frame synchronization in digital data transmission and

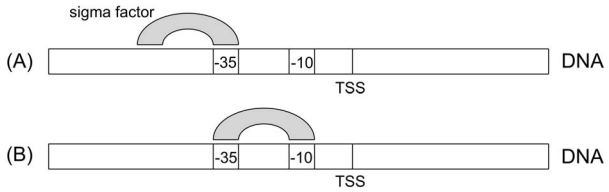

Fig. 8. Detection of promoters by the sigma subunit.

TABLE III

MEAN AND STANDARD DEVIATION OF PSL' FOR ALL POSSIBLE SEQUENCES OF $\mathcal{A}' = \{A, T\}$ WITH LENGTH $N = 6$.

	mean	st. deviation
PSL'	2.89	0.76

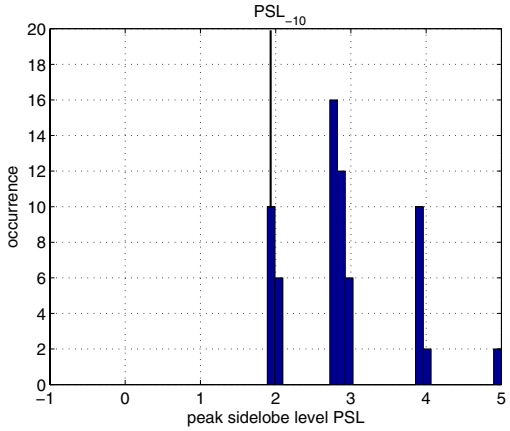

Fig. 9. Histogram of the peak sidelobe level for all possible sequences of $\mathcal{A}' = \{A, T\}$ with length $N = 6$.

transcription initiation provided a powerful tool for promoter analysis, an important aspect of biological research on gene expression. Promoters can be seen as biological sync words that need to be detected reliably by the protein sigma factor. In technical systems, the sync words are chosen based on their autocorrelation properties which are evaluated using the peak sidelobe level. Therefore, we adapted the autocorrelation function to the quaternary alphabet of nucleotides to rate the synchronization properties of promoter sequences in the bacterium *E.coli*. Subsequent calculation of peak sidelobe levels brought up that the -35 promoter region is an outstanding synchronization pattern that enables reliable detection of the close-by start of the gene. The -10 promoter region at first seemed to have worse properties, however, we were able to show that this is due to its importance for other steps of transcription initiation. If taking this constraint into account (high content of nucleotides `A` and `T`), it showed to be among the group of sequences with best possible synchronization properties. Both facts imply that during evolution promoter regions evolved such that they are easily and accurately detectable to ensure expression of the respective gene. Research in molecular biology has focussed on bacterial promoter regions for decades, however, without addressing the presented aspects of a sequence's detectability. Our approach helps to bridge this gap which demonstrates once more the importance of communication theory for the interpretation of processes in molecular biology.

ACKNOWLEDGMENTS

We thank Jakob Mueller from the Max Planck Institute for Ornithology and Juergen Zech from the Institute for Medical Statistics and Epidemiology of the Technical University of Munich for their support regarding the biological details of our research.

REFERENCES

[1] E. E. May, M. A. Vouk, D. L. Blitzer, and D. I. Rosnick, "Coding theory based models for protein translation initiation in prokaryotic organisms," *BioSystems*, vol. 76, pp. 249–260, August-October 2004.

[2] Z. Dawy, F. Gonzalez, J. Hagenauer, and J. C. Mueller, "Modeling and analysis of gene expression mechanisms: a communication theory approach," *proceedings of the IEEE International Conference on Communications (ICC)*, May 2005.

[3] Z. Dawy, B. Goebel, J. Hagenauer, *et al.*, "Gene mapping and marker clustering using Shannon's mutual information," *IEEE Transactions on Computational Biology and Bioinformatics*, vol. 3, no. 1, pp. 47–56, January-March 2006.

[4] "DNA as Digital Data - Communication Theory and Molecular Biology," *IEEE Engineering in Medicine and Biology*, vol. 25, no. 1, January/February 2006.

[5] M. Guthold, X. Zhu, C. Rivetti, *et al.*, "Direct observation of one-dimensional diffusion and transcription by *Escherichia coli* RNA polymerase," *Biophysical Journal*, vol. 77, pp. 2284–2294, October 1999.

[6] E. A. Campbell, O. Muzzin, M. Chlenov, *et al.*, "Structure of the bacterial RNA polymerase promoter specificity σ subunit," *Molecular Cell*, vol. 9, pp. 527–539, March 2002.

[7] B. Lewin, *GENES VIII*. Pearson Education International, 2004.

[8] P. Robertson, *Optimal frame synchronization for continuous and packet data transmission, Fortschritt-Berichte VDI*, ser. 10, no. 376. Duesseldorf: VDI-Verlag, 1995.

[9] R. H. Barker, "Group synchronization of binary digital systems," *Communication Theory*, pp. 273–287, 1953.

[10] J. L. Massey, "Optimum frame synchronization," *IEEE Transactions on Communications*, vol. 20, no. 2, pp. 115–119, April 1972.

[11] H. Kiryu, T. Oshima, and K. Asai, "Extracting relations between promoter sequences and their strengths from microarray data," *Bioinformatics*, vol. 21, no. 7, pp. 1062–1068, October 2005.

[12] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed. McGraw-Hill, 1991.

[13] B. K. Levitt, "Long frame sync words for binary PSK telemetry," *IEEE Transactions on Communications*, vol. 23, pp. 1365–1367, November 1975.

[14] M. J. E. Golay, "Sieves for low autocorrelation binary sequences," *IEEE Transactions on Information Theory*, vol. 23, no. 1, pp. 43–51, January 1977.

[15] H. D. Lueke, *Korrelationssignale*. Berlin: Springer-Verlag, 1992.

[16] B. Shomer and G. Yagil, "Long W tracts are over-represented in the *Escherichia coli* and *Haemophilus influenza* genomes," *Nucleic Acids Research*, vol. 27, no. 22, pp. 4491–4500, November 1999.