FINDING NEAR-PERIODIC DNA REGIONS USING A FINITE-FIELD FRAMEWORK

Gail L. Rosen

Georgia Institute of Technology, Center for Signal and Image Processing, Atlanta, GA 30332

gailr@ece.gatech.edu

ABSTRACT

Previously, we investigated the error-corrective properties of DNA. Although there is no evidence of a universal coding structure, we now show finite-field mathematics may yield a promising deterministic framework to analyze DNA. In this paper, we discuss a finite-field representation of nucleotides and explore an application, an algebraic technique for detecting near-periodic structure in DNA. In conclusion, this technique is a robust iterative process for finding periodicities among mutational errors.

1. INTRODUCTION

Since the introduction of Shannon's mathematical theory of communication, scientists have tried to explain DNA within an information theoretic framework (e.g. [1]). MacDonaill believes to have uncovered why DNA chose a nucleotide alphabet of four [2]. Schneider explores how much pattern is stored in the DNA for genetic control systems [3]. May et. al [4] explores how biology might use a block decoding technique to identify translational signals amidst errors and mutation in mRNA sequences. We have examined DNA in this same light but look for an overall encoding scheme throughout entire DNA sequences in our previous study [5] as DNA has been found to incorporate a "proofreading" mechanism in the replication process [6]. So far, there is little evidence of a universal error-correction coding structure in DNA analogous to man-made communications [5], but May's work [4] implies that the type of coding scheme used may vary from region to region and depend on the type of signal encoded.

Thus, we extend our previous study by using finite-field mathematics to search for localized redundancies in DNA and show that this framework is useful for periodic region detection. "Tandem repeats" is a more formal term of these periodic regions, and they can contain simple repeats, repeats of variable length, or multiple period repeats. These regions are associated with human disease, play a role in evolution and are important in DNA fingerprinting. It is difficult to characterize these regions due to imperfect conservation of patterns caused by mutation.

2. DNA REPRESENTATION

Whenever one attempts to tie mathematical theory to the genome, the most important assumption is the representation of the set of nucleotides, $\{A, T, C, G\}$. For four bases, one can choose a simple representation such as $A = 0$, $C = 1$, $G = 2$, $T = 3$ and use modulo operations, but this implies a structure on the nucleotides such that $T > A$ and $C > G$. Also, one can use indicator sequences (binary sequence representing the locations of each element of nucleotides) producing a four-dimensional representation yielding an efficient representation for spectral analysis [7].

A deterministic symbolic framework is needed. Symbolic statistical techniques, using markov models to represent the various nucleotide states, have been developed to predict gene sequences [8]. Therefore, a representation is needed which allows deterministic mathematical operations on a finite set of elements. Finite-field theory offers three fields for DNA analysis: groups, rings, and fields. For a short synopsis, a group is a set of elements on which a binary (usually additive) operation has been defined, a ring can have multiplicative and additive operations but an inverse may not exist (i.e.: subtraction is possible but not division), and a field has both operations and their inverses [9]. (This is not a complete set of conditions for the fields). If one wishes to have a wide range of operations available for linear algebraic analysis of a set of elements, a finite field is the preferred framework.

DNA is a symbolic set and in no way can be characterized as a group, ring, or field. Afterall, the commutivity and associativity of DNA is unknown. As a result, we look solely upon the fact that if linear algebra were available as a tool, ease of analysis awaits. Therefore, using this logic only, we choose to analyze DNA as a finite field and will inspect the results to assess the validity of this framework.

In this paper, we will map nucleotides to a Galois field [9] of four, $GF(4)$. Since $GF(4)$ is an extension field of $GF(2)$ (any $GF(2)$ binary pair corresponds to one of four $GF(4)$ symbols) , we can create labels (Table 1) for the nucleotide elements with $GF(2)$'s primitive polynomial:

$$\alpha^2 + \alpha + 1 = 0 \tag{1}$$

$$
\begin{aligned}
0 &= \quad 0 \quad \Leftrightarrow 0 \Leftrightarrow A \\
\alpha^0 &= \quad 1 \quad \Leftrightarrow 1 \Leftrightarrow C \\
\alpha^1 &= \quad \alpha \quad \Leftrightarrow 2 \Leftrightarrow T \\
\alpha^2 &= \alpha + 1 \Leftrightarrow 3 \Leftrightarrow G
\end{aligned}
$$

Table 1: Exponential root representation, polynomial representation, numerical label, and nucleotide label.

+	0	1	2	3
0	0	1	2	3
1	1	0	3	2
2	2	3	0	1
3	3	2	1	0

*	0	1	2	3
0	0	0	0	0
1	0	1	2	3
2	0	2	3	1
3	0	3	1	2

Table 2: Addition and multiplication tables in $GF(4)$.

This abstraction of elements to integer labels makes finite field theory an attractive framework.

Although each element is meant to be symbolic in nature, the true meaning of these assignments give rise to the questions: What is it meant for A to be the 0 element and to be its own multiplicative inverse? Is this finite field abstraction truly symbolic?

Refer to [9] for a detailed derivation of the $GF(4)$ operations using the polynomial in equation 1. For reference, we show the resulting operation tables in Table 2.

Another question arises when using these operations in a linear space. What does it mean for a vector to be self-orthogonal? In GF(4), the inner-product of $[2\ 1\ 0\ 3]$ with itself is 0. Abstraction of pure mathematics to a physical system introduces anomalies and its implications are not obvious. We will use this framework to analyze redundancy in DNA and draw conclusions about the advantages and disadvantages of a finite-field framework.

3. LINEAR DEPENDENCE TEST

Now with a nucleotide representation and defined arithmetic operations, we can use linear algebraic techniques on DNA sequences. To analyze redundancy, we developed a method, the linear dependence test, to search for localized regions of linear dependence in sequence data. The linear dependence (LD) test indicates the mere existence of a subspace while the subspace partitioning method from our previous paper [5] tells us the subspace's orientation. If we can determine that a subspace exists and is present for a greater portion of the data, we can use this as a starting point for further examination of its orientation (as explored in [5]).

In the LD method, an N^2-length window of the data is reshaped as an $N \times N$ matrix as shown in Figure 1. This matrix occupies a maximum of N-dimensions. In the linear dependence test, the rank of each $N \times N$ window is computed to find its dimensional occupancy; the rank computation is based on a recursive Gaussian-elimination [10] modified for $GF(4)$ arithmetic. Then the data is incrememented

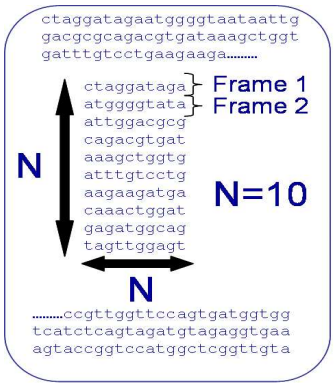

Figure 1: Illustration of how an $N \times N$ window is shaped from the DNA data.

by an N-length frame each time, thereby creating a slowly moving $N \times N$ window which moves by N nucleotides each iteration until the entire sequence has been traversed. A weight, I, is increased linearly, $I = I + 1$, on each iteration if rank-deficiency is found in consecutive windows segments.

The outline of the LD technique:

- For analysis frame length, N, collect N consecutive vectors to form $N \times N$ window.
- Perform a rank computation of the $N \times N$ matrix.
- Increment by one frame for each iteration.
- Note consistent rank-deficiency by linearly increasing I.

By itself, this method is a measure of linear dependence in regions of the data but not necessarily globally as needed for a block coding scheme. For global linear coding to be present, the basis vectors would have to form a consistent subspace over all frames in a sequence. Thus, this algorithm is more localized and detects approximate repeats and even time-varying approximate repeats.

4. RESULTS

4.1. Sequence Data Source

Using the online $GenBank$ database, the entire Yeast sequence (accession number: NC_001133) and a human satellite region (accession number: HSVDJSAT) was selected.

4.2. Sequence Analysis

First, we decided to run the algorithm on whole sequences such as Yeast which can be seen in Figure 2.

In these graphs, the x-axis corresponds to the frame number in which the $N \times N$ window begins, and the y-axis denotes our algorithm running for all $N - 1$ frame offsets needed to test all possible groupings (see Figure 3 for illustration) of the data, accounting for insertion and deletion mutations [6]. This means that if an insertion or deletion occurs and effectively shifts the redundant portion forward or

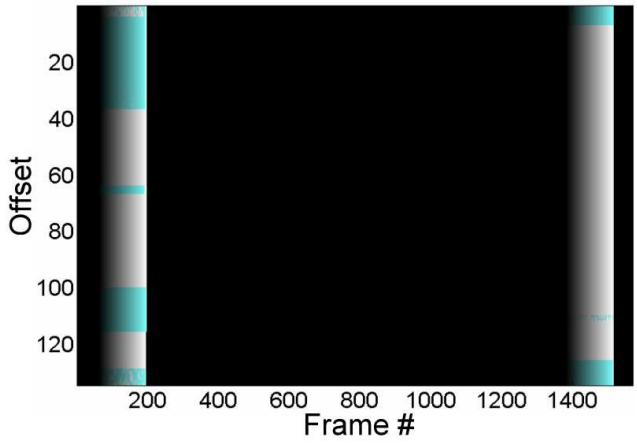

Figure 2: $N = 135$ for NC_001133. The rank computation is computed for each $N \times N$ window starting at the frame number plus offset.

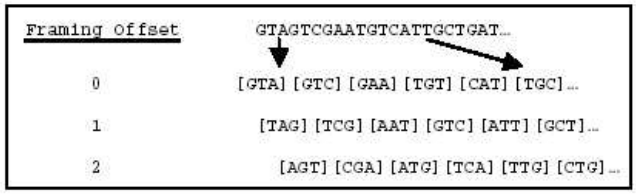

Figure 3: Illustration of vector framing for $N = 3$. Three of these vectors would form the initial $N \times N$ window.

backward, this information is preserved because all shifts of the data are examined. If an $N - 1$ rank subspace is found, it is denoted in white, and subsequently an $N - 2$ rank is denoted in blue, an $N - 3$ rank in magenta, and an $N - 4$ rank in yellow. The intensity of this color is determined by I, the persistence of the linear dependence, as mentioned in the LD method's final step.

In Figure 2, two notable redundant regions of over 17000 bases are found to have rank-deficiency when $N = 135$. Even though data is only deficient by one or two dimensions, visually inspecting a portion of the data in Figure 4 shows the frames are almost identical to each other, indicating that a tandem repeat is present.

In [11], it was found that the HSVDJSAT sequence, a repetitive satellite region of 1985 bases in the human genome, has a tandem repeat of 19 bases. Using the LD test, one can easily see the tandem repeat in Figure 5. While the strong repeat is from $1150 - 1500$, a longer redundant region starting at around base 900 is detected by using an offset of 6.

Exhaustively running the algorithm for many N, it was also found that a strong periodicity of 24 bases exists as seen in Figure 6. At an offset of 12, there is a periodic region of over 1100 bases, which is longer than the periodicity found

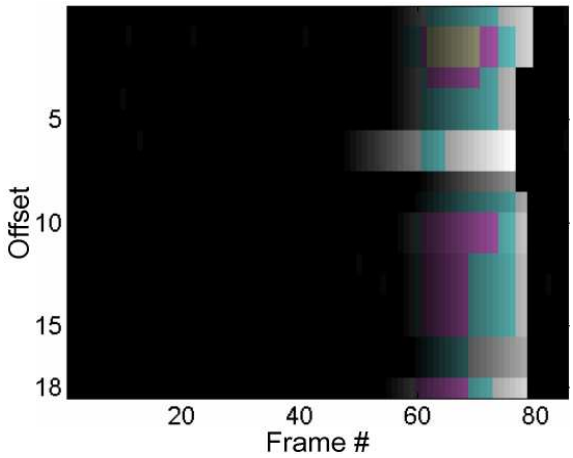

Figure 5: $N = 19$ for HSVDJSAT.

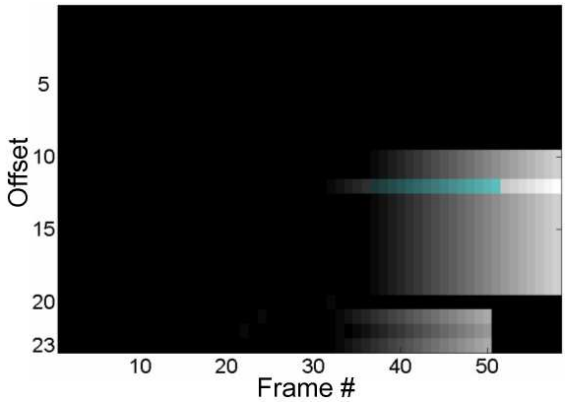

Figure 6: $N = 24$ for HSVDJSAT.

in the $N = 19$ runs. Hauth [11] does not report this periodicity, and with existing tandem repeat algorithms, it may be hard to find. The *mreps 2.5* algorithm [12] did not yield any periodicity of 24 for this sequence. This may be due to the fact that no exact repetition exists. In Figure 7, a portion of the HSVDJSAT region is shown, and no two frames are the same because of mutational errors. For current tandem repeat algorithms, this is a problem because they are based on exact frequencies, but our algorithm detects redundancies and therefore can easily identify near-periodic regions.

The LD test easily found the $N = 24$ redundancy, and a listing of the nucleotides in the strong tandem repeat region can be seen in Figure 7. An interesting note is that the LD algorithm does not search for exact repeats or matching patterns. This can be seen when the algorithm finds similar structures rather than exact repeats. In Figure 7, the red blocks show regions where the nucleotides may have mutated to other nucleotides (known as substitution errors) when the DNA was copied and appended. The blue squares indicate regions where the a deletion error may have occurred. The orange circles in this figure shows where the

Figure 4: A near tandem repeat sequence of 1890 nucleotides in the Yeast NC_001133 sequence (exact nucleotide indices are located at the top).

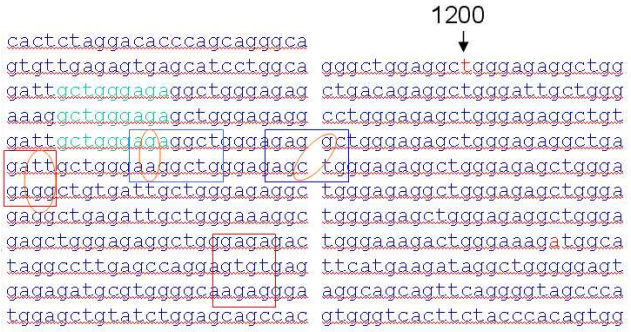

Figure 7: $N = 24$ listing of $1141 - 1672$ for HSVDJSAT.

suspected error occurred; in both deletion cases, a G was deleted. Despite these errors which throws other algorithms astray, the LD algorithm easily detected the periodicity.

5. DISCUSSION: THE GALOIS FIELD AND DNA

There are some clear disadvantages to using $GF(4)$ in conjunction with linear algebra. The clearest is the self-orthogonal vector property. For this to occur, the number of nucleotides in the frame must be an even number; if we look at Table 2, a number added to itself is 0, and this can occur for an inner-product of an even-element symmetric nucleotide vector, $ATAT$ or $[0\ 2\ 0\ 2]$. Before more complex linear algebraic operations can be used, this anomaly must be dealt with.

The proposed analysis in this paper yields a quick way to visually inspect periodicity in DNA regions. We now address the validity of the finite-field framework. From Figure 7, it can be seen that the algorithm is robust amidst various errors. This is due to the fact that the finite field structure preserves the symbolic nature of the nucleotides. Similar structures are linear combinations of each other (e.g. $GAGA = GTGT + ATAT$). If a pure repeat exists, the dimensional occupancy of a analysis window is 1. With added insertions and deletions, the window rank increases but not enough for all basis vectors to be linearly independent of each other, thus still detecting a strong redundancy among errors. Although some analysis cannot be performed because of the self-orthogonal property induced by $GF(4)$, the Galois field allows complex operations on a finite symbolic set and enables powerful tools for DNA analysis.

6. CONCLUSIONS

In this paper, we develop an algorithm which finds near periodic DNA regions, common to many genetic disorders, in a fast iterative process. Also, it is shown that using a finite-field framework enables the use of linear algebra's massive toolbox. Two sequences are analyzed via the LD algorithm, and expected tandem repeats are found in each. A previously undiscovered repeat of 24 bases is found in the HSVDJSAT sequence. This discovery is due to the abundance of mutation errors in this sequence which other algorithms do not tolerate. In conclusion, the linear dependence test is a simple way to find imperfect periodicities and remains robust in substitution, deletion, and insertion errors.

7. REFERENCES

[1] Gatlin L.L., *Information Theory and the Living System*, Columbia University Press, New York, 1972.

[2] Mac Donaill D.A., "A parity code interpretation of nucleotide alphabet composition," *Chemical Communications*, pp. 2062–2063, 2002.

[3] Schneider T.D., "Some lessons for molecular biology from information theory," *Entropy Measures*, vol. 119, pp. 229–237, 2003.

[4] May E.E. et. al., "Coding theory based models for protein translation initiation in prokaryotic organisms," in *Fifth International Workshop on Information Processing in Cells and Tissues (IPCAT)*, September 2003.

[5] Rosen G.L. and Moore J.D., "Investigation of coding structure in dna," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April 2003.

[6] Burdon R.H., *Genes and the Environment*, Taylor and Francis, Inc., Pennsylvania, 1999.

[7] Anastassiou, "Genomic signal processing," *IEEE Signal Processing Magazine*, July 2001.

[8] Borodovsky M. et. al., *GeneMark*, http://opal.biology.gatech.edu/GeneMark/.

[9] Wicker S.B., *Error Control Systems*, Prentice Hall, New Jersey, 1995.

[10] Golub G.H. and Van Loan C.F., *Matrix Computations*, The Johns Hopkins University Press, Maryland, 1996.

[11] Hauth A., *Identification of Tandem Repeats Simple and Complex Pattern Structures in DNA*, Ph.D. thesis, University of Madison, Wisconsin, 2002.

[12] Kolpakov R., Bana G., and Kucherov G., *mreps*, http://www.loria.fr/mreps/.