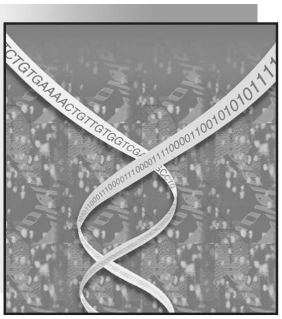


© EYEWIRE

Examining Coding Structure and Redundancy in DNA

How Does DNA Protect Itself from Life´s Uncertainty?

BY GAIL L. ROSEN

The genetic code instructs proteins on the translation of nucleotides to amino acids, but this example is only one of many signals encoded in DNA. It is well known that these protein-coding regions have the lowest mutation rates in the DNA strand. So, the question arises: how does DNA protect itself from error? A review of DNA signal content, redundancy, and mutational mechanisms is presented. Then, mutation-robust methods are developed to detect a linear coding structure and approximate tandem repeats.

Ever since the introduction of the Watson-Crick model of DNA, scientists have been trying to decipher the long sequence of millions (or, for complex organisms, billions) of bases. The genetic code, the mapping of nucleotide triplets (codons) to amino acids, or "protein-coding," was one of the first discoveries. Signals in DNA could then be paralleled to digital signals. After 30 years, many functions and signals in DNA still remain unknown, and scientists have conjectured that nonprotein-coding regions, which compose 97% of human DNA, are unused junk [1]. On the contrary, recent studies reveal that binding sites and initiation signals exist in these nonprotein-coding areas, and mutation errors in these regions cause diseases [2]. Nonprotein-coding regions contain a finite amount of "algorithmic" content [1], [3]. Discovering the signals and function in these areas is just the beginning of genome discovery. In this article, we develop a method to uncover an error-correction coding structure in the nucleotide sequence, and show that our framework is efficient for detecting approximate tandem repeats, such as microsatellite regions.

DNA Composition and Repeats

DNA is composed of four bases or nucleotides. A (adenine) and G (guanine) are considered purines (R), and T (thymine) and C (cytosine) are considered pyramidines (Y), with purines being the larger of the two. This size imbalance between nucleotides creates an affinity between purines and pyramidines, and stability is only reached with complementary pairing: A bonds to T (two weak hydrogen bonds) and C bonds to G (three weak hydrogen bonds). The weak complementary bonds make DNA easy to unzip in replication, but they can also make it susceptible to interfering molecules; thus for error protection in its stable state, the double strand curls into a helix. It has been found that certain nucleotide repeats help DNA to wrap into the curved state. The dinucleotides, AA and TT, are placed at certain phases from each other and cause an average periodicity of 10.55 ± 0.01 base pairs in the DNA sequence; AG and CT also aide to the helical twist [4].

A fascinating nucleotide series is the telomere, the end of the chromosome used to buffer genes from the environment. Due to the way the replication mechanism truncates, the DNA strand shortens each iteration; human DNA shortens by 50 base pairs (bp) on every cell division [5]. To prevent nucleotide loss from eventually interrupting a gene, telomerase elongates a chromosome's ends with repetitive sequences such as TTAGGG, sometimes for thousands of bases [5]. As we age, telomerase expression weakens, genes no longer have protection from being cropped, and cells die. On the contrary, when telomerase is overexpressed, cells tend to live much longer and divide more frequently, resulting in cancer [6]. Ninety percent of tumoric growths exhibit excessive amounts of telomerase! Sequence periodicity and repeats play a vital role in the stability of the overall structure.

Some DNA regions are correlated to specific functions or signals, and a famous function is that of protein coding, also known as the coined "genetic code." These identified patterns and sites already give seemingly random DNA a clear deterministic structure. Schneider presents a comprehensive list of DNA signals recognizable by pattern and information content [7]. In our methods, we begin to examine the underlying redundancy and tandem repeats present in the nucleotide sequence.

Mutations and the Replication Process

Scientists give a rough error rate of 10^{-10} mutations/nucleotides when DNA is copied. So, what are these mutations and how can we quantify them? Substitution mutations mostly occur due to 1) accidental bonding of Brownian-motioned biological elements to DNA or 2) electromagnetic radiation providing enough energy to break bonds in the structure. As an example of 1), one of the most common mutations is the hydrolysis of C to T, known as cytosine deamination. Water molecules do not have as easy an access to nucleotides in DNA's stable helical structure as they do when DNA is unzipped for replication. In fact, cytosine deamination is 100 times more likely in replication [8]. Temperature, geometry, and environment are key factors in studying DNA mutation rates.

In addition to errors/mutations caused by clumsy molecules bumping into DNA, replication itself (or the copying mechanism) can introduce errors that appear structured. For example, microsatellite regions, an excess of repetitive sequences, result from replication slippage [9]. Microsatellites in human DNA are associated with 14 neurodegenerative genetic disorders found in [2]. Repeats from telomerase slippage causes increased cell division and highly correlates with malignant cancer growth.

The replication procedure alone has an error rate of 10^{-3} to 10^{-5} [8], but DNA has an internal "proofreading" mechanism. When copied, the helical structure unzips and forks into two separate strands; complementary bases then attach themselves to complete the new ladders. When a substitution error occurs, usually a purine replaces a purine ($C \rightarrow T$) or a pyramidine replaces a pyramidine ($A \rightarrow G$) in the complementary attachment. This causes a kink to develop due to the mismatch, and no more bases are added until the correct nucleotide is restored. This simple proofreading reduces the error rate to approximately 10^{-10} [8]. Can understanding these repair pathways lead to better error-correcting technologies? Overall, it is important for the computational biologist to be wary of the various mutational errors when examining DNA sequences.

Nucleotide Representation

When analyzing DNA, the mathematical representation of the nucleotides, $\{A, T, C, G\}$, is the fundamental first step. It has even been contemplated why nature chose such an alphabet in [10]. Many representations have been proposed and adapted to the type of analysis. For example, purines (A and G) and the pyrimidines (C and T) can be represented with a binary representation. In addition, a simple representation can be chosen for the four bases such as $A = 0, G = 1, C = 2, T = 3$ (modulo operations), but this implies a structure on the nucleotides such that $T > A$ and $C > G$. For a model of the translation process, Anastassiou defines a complex representation to the nucleotides: $A = 1 + j, T = 1 - j, G = -1 + j$, and $C = -1 - j$ [11]. The geometric interpretation of this representation still imposes a structure such that the Euclidean distance between A and C is greater than the distance between A and T, yet for the application, nucleotide quantization to amino acids, it is useful [11]. Various representations, including the one proposed in this article, can be seen in Table 1.

Symbolic statistical techniques, using Markov models to represent the various nucleotide states, have been developed to predict gene sequences [14]. But a representation is needed that allows deterministic mathematical operations on a finite set of elements. A field has addition, multiplication, and their inverse operations (subtraction and division) unlike groups or rings [15]. If one wishes to have these four operations available to analyze a sequence of symbols, a finite field framework is preferred.

In [13], we propose a mapping of nucleotides to a Galois field of four, noted as *GF(4)* [15]. 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 2) for the nucleotide elements with *GF(2)*'s primitive polynomial:

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

The abstraction of elements to integer labels is an attractive property of the finite field representation.

The polynomial in (1) can be manipulated in addition, multiplication, subtraction, and division in *GF(4)*. Refer to [15] for a detailed derivation. For reference, we show the resulting addition and multiplication operation tables in Table 3.

Information Theoretic Studies

Inspired by information theory, Gatlin developed entropy and divergence measures to quantify complexity in DNA [16]. The entropy, or information capacity of a sequence, is maximized when all four nucleotides are equiprobable:

Table 1. Table of DNA mathematical representations found in the literature. An example sequence, GCATT, with its complement and characteristic property given for each representation.

Example Sequence: G C A T T | | | | | A A T G C	Representation	Sequence GCATT	Complement AATGC	Property
Simple integer assignment	$A = 0, G = 1, C = 2, T = 3$	1 2 0 3 3	0 0 3 1 2	Uses modulo operations
Complex assignment (QPSK) (13)	$A = 1 + j, G = -1 + j,$ $C = -1 - j, T = 1 - j$	$-1 + j, -1 - j,$ $1 + j, 1 - j, 1 - j$	$1 + j, 1 + j, 1 - j,$ $-1 + j, -1 - j$	Reverse and conjugate to get complement
PAM representation (13)	$A = -1.5, G = -0.5,$ $C = 0.5, T = 0.5$	$-0.5, 0.5, -1.5,$ $1.5, 1.5$	$-1.5, -1.5, 1.5,$ $-0.5, 0.5$	Reverse and negate to get complement
Binary indicator sequence (12)	$S_i(n) = 1$ where $S(n) = i$ $S_i(n) = 0$ where $S(n) \neq i$	A: 0 0 1 0 0 G: 1 0 0 0 0 C: 0 1 0 0 0 T: 0 0 0 1 1	A: 1 1 0 0 0 G: 0 0 0 1 0 C: 0 0 0 0 1 T: 0 0 1 0 0	Four-dimensional representation
Galois field assignment (15)	$A = 0, C = 1,$ $T = 2, G = 3$	1 2 0 3 3	0 0 3 1 2	Uses symbolic Galois field operations

$$H = -\sum_i p_i \log_2(p_i)$$
$$= -\sum_{i=1}^{4} \frac{1}{4} \log_2\left(\frac{1}{4}\right) = 2\,\text{bits}. \qquad (2)$$

In many species, the bases are not equiprobable, but temperature dependent. Three bonds exist in C and G bases, while only two exist in A and T. Thus, it takes more energy to break the bond between C and G, and it has been found that GC content is higher in a warmer-environment than colder-environment organisms. For example, Micrococcus Lysodeikticus, which inhabits warm spots, has the following base frequencies: $\Pr(C) = \Pr(G) = 0.355$ and $\Pr(A) = \Pr(T) = 0.145$ [16]. By way of (2), the entropy for this organism is 1.87 b and this nucleotide imbalance implies redundancy.

A simple entropy measure like (2) indicates nucleotide bias in a sequence. In recent years, new measures have been developed such as entropic profiles of various-length genomes [17]. Schneider illustrates DNA nucleotide bias for each nucleotide position through an easy-to-read sequence logo graph [18]. Techniques for studying information content and bias have begun to quantify DNA's implicit structure. In our work, we show how coding theory and signal processing methods can be used to investigate this structure.

Coding Models of DNA

Since DNA is a finite, symbolic sequence, it is a natural to extend the use of coding theory to sequence analysis. Battail has stated that DNA evolves from a series of repeats heavily altered by mutation, such as ill-conserved introns, and he presents a replication decoding framework [19]. His "multiple unfaithful repetition" model only uses partial knowledge of the coding constraints in order to decode a message; this property makes the model attractive since little is known about the DNA encoding structure. Inspired by this model, we use partial knowledge methods in our work.

Also, much research has been done by May et al. to study *E. coli* translation initiation sequences using block and convolutional coding models [20]–[23]. mRNA is viewed as a noisy encoded signal, and the ribosome, which translates the sequence, is seen as the decoder. Several biological and chemical factors are used to parameterize the ribosomal decoding model. The block code model is effective in recognizing the ribosomal binding site, and the convolutional model easily distinguishes between translated and untranslated sequences. May et al. show that coding models are effective in signal recognition and inspire us to ask whether there is an inherent coding structure in DNA.

Table 2. Exponential root representation, polynomial representation, numerical label, and nucleotide label for the *GF(4)* representation.

$$\alpha^0 = 1 \Leftrightarrow 1 \Leftrightarrow \mathrm{C}$$
$$\alpha^1 = \alpha \Leftrightarrow 2 \Leftrightarrow \mathrm{T}$$
$$\alpha^2 = \alpha + 1 \Leftrightarrow 3 \Leftrightarrow \mathrm{G}$$
$$0 = 0 \Leftrightarrow 0 \Leftrightarrow \mathrm{A}$$

Table 3. Addition and multiplication tables in *GF(4)*.

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

x	0	1	2	3
0	0	0	0	0
1	0	1	2	3
2	0	2	3	1
3	0	3	1	2

Determination of an Underlying Linear Code

As discussed in the first section, DNA's repair mechanism detects and fixes irregularities in the sequence and significantly reduces the error rate of the replication process. Also, the bonds between the complementary pairs and the shape of the strands introduce constraints on the sequence, making the sequence less random than previously thought. The genetic code introduces the strictest rules in regions where every three nucleotides produces an amino acid; the 64 nucleotide combinations correspond to 20 amino acids and imply inherent error-protection. This has led Battail and others [13], [24] to consider the existence of error protection in the assembly of nucleotides and that there might be more to DNA repair than just polymerase detection of irregularities in the sequence. Is there a universal block or convolutional code in the sequence where the proofreading mechanism is the decoder? Already, Mac Dónaill has hypothesized that a parity check code is present in the chemical bonds of the four bases [10], providing a foundation for further investigations into sequencing coding structure.

Liebovitch presents the first search for an error-correction code in DNA using a single parity-bit search method [24]. While his methodical coding-theory-based investigation does not reveal the presence of a consistent single parity-bit code, the experiment provides inspiration for future investigations and context for the complexity of the problem. Thus, there is a need for a general approach to find *k*-parity bits placed in any order in any *n*-size code to discern an *(n, k)* block coding structure from a DNA sequence. We introduce subspace partitioning (SP), developed from classical coding theory, as a way to search/test for such codes without prior knowledge of the *n* or *k* values, which are usually known in communication channel error-correction codes. In biology, we lack these values, thus we develop a novel, generalized method to look for any *(n, k)* block-coding structure. Second, we account for DNA frame shift mutations, which are also usually not an issue in telecommunications applications. Third, the symbolic framework of the Galois field allows the four different bases to be solely symbolic, as they are in nature. So, while our method is founded in error-correction code theory, we tailor it to our biological application.

Modeling the Replication Channel

Communication channel models can be paralleled to DNA processes. In one doctrine, the channel is assumed to be the amino acid translation from nucleotide triplets [16]. In May et al., the channel is the actual replication process, and the DNA is the medium in which genetic information is transmitted from generation to generation [25]. The latter is good for mutation modeling since transcription and copying of DNA is a noisy process. From the first section, when the activity of the proofreading mechanism is blocked, replication error rates increase. This leads us to the hypothesis that there is a sequence coding structure to protect against replication noise.

In Figure 1, we assume that the DNA is the sequenced genomic data available in GenBank [26] and that our goal is to examine the dashed-line-encompassed area and uncover the encoder scheme; in other words, we wish to infer structure from

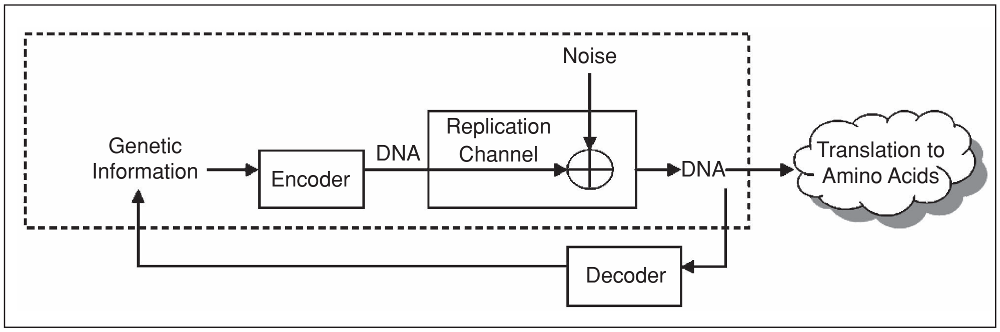


Fig. 1. Our noisy channel model of genome replication with underlying coding assumption.

the noisy output to retrieve the original genetic information. Also, if our assumption is correct and DNA is encoded in a linearly redundant fashion, our analysis will uncover it. In this system, we know nothing about the encoder or the original information; therefore, system identification and deconvolution methods cannot be used. We will assume that the encoder is linear and try to characterize it given such output.

Subspace Partitioning for (N, K) Codes

In our investigation [13], our primary goal is to identify and characterize any linear constraints that might appear in regions of a sequence. Lacking the benefit or prior knowledge regarding the location, duration, or dimensionality of subspace partitioning in the sequence, we propose a method that generates a complete orthogonal basis set oriented to a local region of data. The basis set is used to decompose the sequence (equivalent to a coordinate transformation). The consistent presence of nulls in the transformed sequence indicates both the presence and the dimension of linear subspace partitioning in the data.

The first assumption is a fixed codeword length *n*. The *N*n*-length DNA sequence is grouped into a matrix, $\mathbf{V} = [\boldsymbol{v}_1\ \boldsymbol{v}_2 \ldots \boldsymbol{v}_N]$ where $\boldsymbol{v}_i$ is the *i*th column vector of length *n*. The alignment of the frames relative to the starting point will be referred to as the *framing offset*. A choice of a particular framing offset will be referred to as the *frameset*. Given the frame length *n*, there are *n* unique framesets. See Figure 2 for an illustration of all frameset groupings.

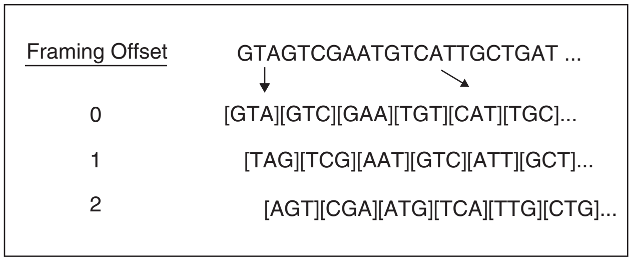
We apply the Gram-Schmidt algorithm using finite field operations to the sequence of vectors to yield a complete set of orthogonal basis vectors, $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots \boldsymbol{e}_n\}$. Once an orthogonal basis is formed from the first *j* frames of data, the $\boldsymbol{v}_i$**'s** for $i > j$ are decomposed into components of each of the basis vectors. This is simply a coordinate transformation and can be described by:

$$t_{\mathrm{i}} = \mathbf{G}\, v_i \qquad \text{where} \qquad \mathbf{G} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{\mathrm{n}} \end{bmatrix}.$$

Provided that the data has been framed correctly when applying the Gram-Schmidt algorithm, a linear coding redundancy can be detected by noting consistent null coordinates over a region in the transformed sequence of length-*n* vectors, $\{\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_{N-j}\}$. This null detection would indicate whether a subspace of the actual *n*-dimensional space exists.

Subspace Partitioning Algorithm Outline

1) Obtain the orthonormal basis $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots \boldsymbol{e}_n\}$ by Gram-Schmidt orthogonalization of *j* number of v_i frames where $j \geq n$. Form the transform matrix **G** from this set.
2) Decompose the sequence into its basis components, $\{\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_{N-j}\}$ across all possible framing offsets.



Framing Offset	GTAGTCGAATGTCATTGCTGAT ...
0	[GTA][GTC][GAA][TGT][CAT][TGC]...
1	[TAG][TCG][AAT][GTC][ATT][GCT]...
2	[AGT][CGA][ATG][TCA][TTG][CTG]...

Fig. 2. Illustration of vector framing for $n = 3$.

3) Note the persistence of nulls in $\boldsymbol{t}_i$**'s**. Calculate confidence by comparing against the probability of sequential sets of randomly chosen vectors having the same subspace partitioning.

Given the copious volume of data produced by iterating the algorithm over numerous frame shifts and codeword lengths, a visualization method is devised to aid in the search for consistent subspace partitioning. A probabilistically based value increases to indicate confidence in the presence of subspace partitioning. We can then plot the confidence as a function of sequence index *i* across all possible framing offsets.

Results of the Subspace Partitioning Method

The algorithm is capable of detecting and characterizing linear subspace partitioning in any sequence provided that such a structure is manifest in the data. For a given sequence, all such structures can be found provided that the algorithm is run for every possible framing offset and for every possible codeword length.

By way of illustration, a test sequence is generated to occupy a five-dimensional subspace of an eight-dimensional vector space. This constitutes an (8, 5) linear block code in *GF(4)*. Running the algorithm on this sequence for $n = 8$ yields the confidence image shown in Figure 3(a). Interstitial symbols are introduced throughout the sequence to illustrate the robustness of the algorithm to framing offsets. When frame shift mutations occur, the region of subspace consistency simply migrates to the corresponding row in the diagram.

The linear SP algorithm is then tested using an *E. coli* K-12 MG1655 sequence (GenBank [26] accession code NC_000913). The result is shown in Figure 3(b). A consistent linear block code is not observed to be present throughout the whole sequence, but some regions are oriented in the same subspace for several consecutive frames, denoted by the aggregated intensity of the light bars.

The SP algorithm requires two conditions from the sequence. Firstly, the algorithm uses nulls in ***t*** to indicate subspace partitioning. This requires that the coordinate system **G** to be properly oriented. So, the sequence has to have the same coding present throughout; otherwise, the vector basis will be misaligned and may make even a regional coding structure impossible to detect.

The algorithm can be generally applied to any sequence for which it is suspected that coding properties are present. It identifies a block code in a symbol sequence independent of framing, provided that the structure is present from the outset. The algorithm could readily be adapted in a classification scheme for data of unknown origin or for cryptographic/cryptanalysis tasks in which the code or encryption scheme is unknown.

Redundancy and Tandem Repeat Detection

From structural studies, we know DNA (especially eukaryotic) has repetitive regions. There are various techniques to classify these [27]–[29]. Most tandem repeat algorithms use complex heuristic, combinatorial, or dynamic programming approaches. In [28], a periodicity transform is used to plot several periodic/near-periodic regions versus position in a simple graph. It is one of the most flexible algorithms (by using different detection thresholds) and efficient representations (periodicities versus nucleotide position), but only base substitution mutations, not frame shift mutations, are taken into account.

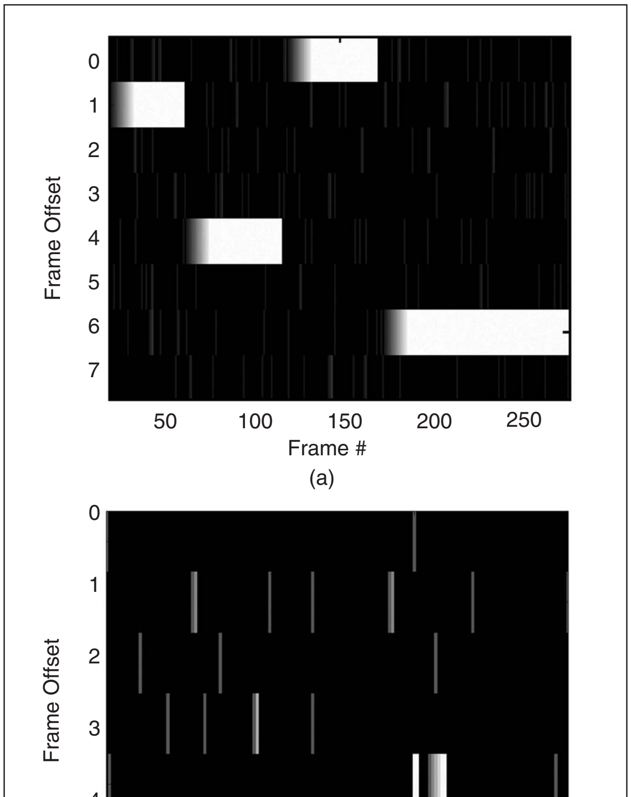


Fig. 3. (a) Linear subspace partitioning results for a (8, 5) block-coding test data. It detects a block code despite interstitial frame shift mutations. (b) Linear subspace partitioning results for a subsection of an $n = 6$ *E. coli* K-12 MG1655 sequence. Regional block codes are detected, but there is no evidence of a universal code.

Now with a nucleotide representation and field-defined arithmetic operations, we can extend the linear algebraic techniques used in the SP method to detect periodicities. To analyze redundancy, we develop a method, the linear dependence (LD) test, to search for localized regions of linear dependence in sequence data. The LD test indicates the mere existence of a subspace, while the subspace partitioning method from the previous section tells us the subspace's orientation. Biologically speaking, the SP method tests for strict block-coding structure, while the LD test detects a "rough" redundancy, such as an approximate repeat. If we can determine that a subspace exists and is present for a portion of the data, we can use this as a starting point for further examination of its orientation (as explored in [13]). The LD test determines local redundant regions and is a good starting point for further analysis such as the detection of tandem repeats.

Linear Dependence Test

In the LD method [30], an N^2-length window of the data is reshaped as an $N \times N$ matrix. 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 algorithm [31] modified for *GF(4)* arithmetic. Then the data is incremented by an N-length frame each iteration, thereby creating a slowly moving $N \times N$ window which moves by N nucleotides each time until the entire sequence has been traversed. A weight, I, is incremented, $I = I + 1$, on each iteration if rank deficiency is found in consecutive window segments.

The outline of the LD technique:

1) For analysis frame length N, collect N consecutive vectors to form $N \times N$ window.
2) Perform a rank computation of the $N \times N$ matrix.
3) Increment by one frame for each iteration.

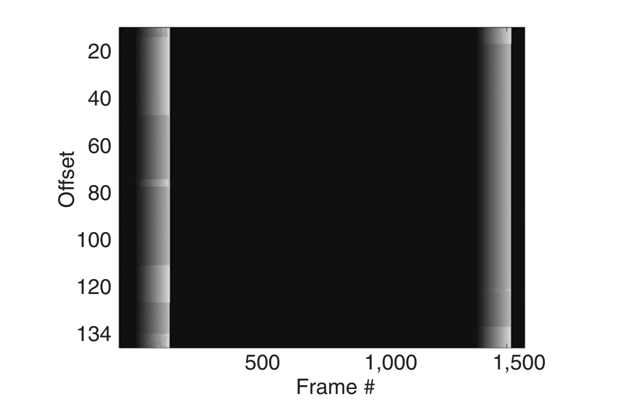


Fig. 4. $N = 135$ LD test for the Yeast Chromosome I sequence, NC_001133. Intensity increases proportionally to the length and level of the rank deficiency of consecutive $N \times N$ windows, each starting at a particular frame number. Two regions associated with the FLO9 gene are shown to be highly repetitive with the LD Test.

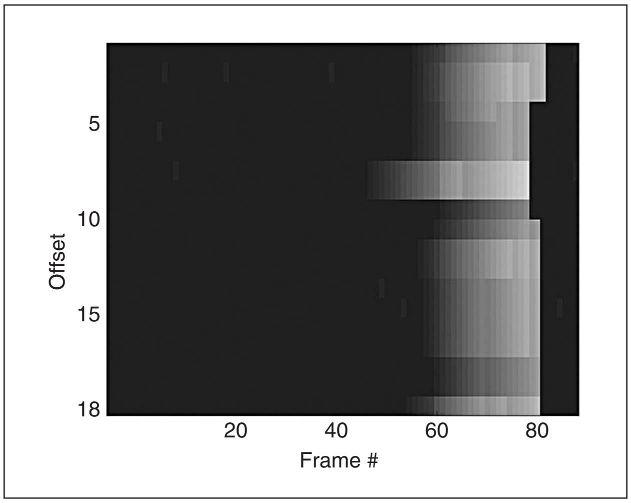


Fig. 5. $N = 19$ LD test for a human satellite sequence, HSVDJSAT. Intensity increases with length and level of the rank deficiency. At offset 6, an 893-base region exhibits a 19-base repeat.

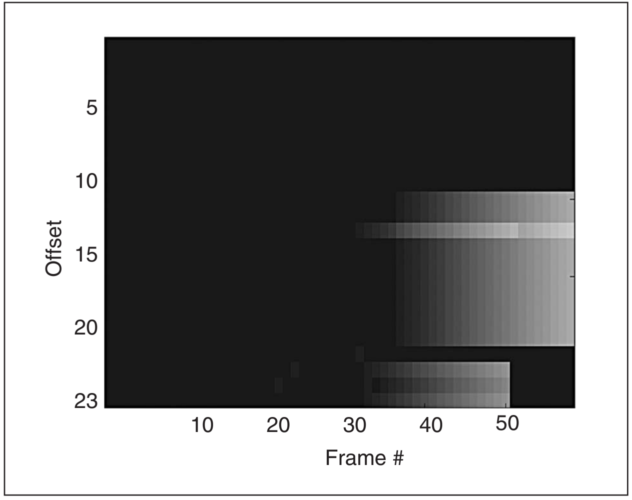


Fig. 6. $N = 24$ LD test for HSVDJSAT. At offset 12, a 1,200 base region exhibits a 24-base repetition. This is longer than the $N = 19$ case. Even though this region has high mutation, the algorithm still finds the redundancy.

4) Note consistent rank-deficiency by incrementing I.

By itself, this method is a measure of regional linear dependence and finds variation in dimensional occupancy between overlapping windows.

Linear Dependence Test Results

In the first section, we discuss the vital role that nucleotide repeats play in chromosome buffering and neurodegenerative disorders. In this section, we show how the LD test is effective in finding tandem repeats, especially those that are highly altered by mutational errors.

Using the online Genbank database [26], we select the Yeast Chromosome I sequence (accession code: NC_001133) and a human satellite region (accession code: HSVDJSAT) for our experimental data.

We introduce a way to highlight periodic regions across all frame offsets to ease visual inspection of periodicities. In Figures 4–6, the x-axis values correspond to the `frame_numbers` (*0* to `sequence_length`/$N - N$), and the y-axis denotes our algorithm running for all N frame offsets needed to test all possible groupings of the data (see Figure 2 for an illustration). If an insertion or deletion occurs and effectively shifts a repetitive portion forward or backward by a few bases, the highlighted segment will still be shown but in another frame offset since all frame shifts are examined. If an $N - 1$ rank subspace is found, it is denoted in dark gray, and the lower the rank of the subspace (up to $N - 4$ for the examples), the brighter the intensity; also, the higher the linear dependence persistence indicator I, the brighter the shading intensity. Therefore, the brightness of the graph is a function of two factors: the strength and length of the redundant region.

First, the algorithm was run on the Yeast Chromosome I sequence which can be seen in Figure 4. In Figure 4, two notable redundant regions of over 17,000 bases are found to have a periodicity of $N = 135$. The number of bases producing a highlighted region can be calculated from the graph by `number_of_frames` $\times N + (N \times N)$. Even though the regions are only rank deficient by one or two dimensions (not a strong linear dependence), the frames of nucleotides are almost identical to each other, indicating a 135-base tandem repeat.

In [27], it is found that the HSVDJSAT sequence, a repetitive satellite region of 1,985 bases in the human genome, has a tandem repeat of 19 bases from 1,195–1,553. The LD test highlights the tandem repeat across all framesets in Figure 5. While the strong repeat is from 1,150–1,728, a longer redundant region starting around base 900 is detected with an offset of 6. Therefore, the LD test can find the longest periodic region by testing all frame shifts.

Exhaustively running the algorithm over various N, a strong periodicity of 24 bases is discovered (as seen in Figure 6). At an offset of 12, there is a periodic region of over 1,100 bases, which is longer than the periodicity found in the $N = 19$ runs. Hauth has recently reported a periodicity of 48 from 1,190–1,553 [32],

```
CACTCTAGGACACCCAGCAGGGCA
gtgTtgAGagtgagCAtCctGGCA GGGCTGGAGGCTGGGAGAGGCTGG
GatTgctGgGaGAGgcTgggaGag ctGacaGAGGCTGGGAttGctgGG
aAagGCTGGGAGAGctgGGagagG CctgggagaGCTGGGAgaGgctGt
gAttGCTGGGAGAGgctGGgagaG gCTGGGAGAGCTGGGAGAGGCTGa
GATTGCTGGGAagGctgGGagagc tggGaGAGgctgGGagagctggGA
GAggctgtGattGctgGgagagGC TGGGAGAGGCTGGGAGAGCTGGGA
GAGGCTGaGATTGCTGGGAaAGGC TGGGAGAGctgGGagagGCTGGGA
GAGctgGgagagGCTGGGAgAGaC TGGGAaAGactGGgaAaGaTGGcA
tAGgccttgagccagGaGtGtGAg TtcatgAagaTaGGctgGggGagt
gAGagaTgcgtgggGcaagagGga aggcaGcAGtTcaGggGtaGccca
tgGaGcTGtaTctGGagcagccac gtGggtCAcTTCtacccacagtgg
aGGtGgacTcTtgtagcCAGagct GTGGacaACcTCTcagaACcagaa
gacccttgctgccctGtatGccaa GgtctCctCCggcCtGggtCtcAg
GgatgccagctgcaaactgGgagg GccaTtgTaCaGaCactaggTggc
tGAgGtaccagttAcAgcctGgtc ttggTgGccacatagaggtccaGC
ctcacTcagctTgAtgGCCaaGct ggtGgGttaggATttgGagtCtGC
agCctTgAGgccttcccaaggtaa aaccaaaTtGtccTgGcttagaat
```

Fig. 7. Annotation of an $N = 48$ HSVDJSAT region (bases 1,141 → 1,976). The annotation scheme used in Figure 7 is used here. An approximate repeat can be seen among insertion and deletion errors. Uppercase letters denote conserved portions, underlined letters denote an insertion from the previous frame, and bold letters denotes a region retained after/around a deletion occurring from the previous frame. Italics denote a region before a deletion, lowercase letters denote substitution errors/sequence differences, and light-gray letters denote portions where multiple base substitutions occur for a particular base.

but with most tandem repeat algorithms, the $N = 24$ periodicity or multiples is difficult to find. For example, the maximal tandem repetition (mreps) 2.5 algorithm [29] did not yield a 24-base periodicity (or multiples of 24) for this sequence. This may be from the lack of exact repeats present. To illustrate, a portion of the HSVDJSAT region is shown in Figure 7, and no two frames are equivalent because of mutational errors. For current tandem repeat algorithms, this is a problem because they are based on exact frequencies, but our algorithm detects approximate repeats and, therefore, can easily identify near-periodic regions.

In Figure 7, the lowercase and light gray nucleotides show regions where the nucleotides may have mutated to other nucleotides (known as substitution errors) in replication. The light gray bases are interesting because they represent substitution of one or more nucleotides, usually dinucleotides, and also occur quite often in this example.

The LD algorithm does not search for exact repeats or matching patterns. Instead, the rank-deficiency of the nucleotide window indicates similar structures, or redundancy, between the segments. Despite these errors, which throw other algorithms astray, the LD algorithm easily detected the periodicity and multiples of 24 as seen in Figure 7.

Conclusions

The subspace partitioning method is based on the hypothesis that there is an underlying coding structure in DNA used for error recovery in replication, but our preliminary results do not indicate a universal block code. In our method, we assume consistent error correction would occur in both protein-coding and nonprotein-coding regions. On the contrary, mutation rates vary from region to region in the genome, and these areas may need separate treatment. For example, nonprotein-coding regions are more susceptible to mutation than protein-coding regions.

In the linear dependence test, we develop an algorithm which finds near-periodic DNA regions, common to genetic disorders, in a fast iterative process. In addition, we show 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. An unexpected approximate repeat of 24 bases is found in the HSVDJSAT sequence. The discovery is due to the algorithm's ability to detect redundancy amidst an abundance of mutation that other algorithms do not tolerate. The linear dependence test is a simple way to find imperfect periodicities and remains robust in substitution, deletion, and insertion errors.

Acknowledgments

This article is based upon work supported under a National Science Foundation Graduate Research Fellowship and an AT&T Research Laboratories Grant. The author would like to thank Jeffrey D. Moore, now at Khimetrics, Inc., for invaluable discussions and his contribution to subspace partitioning method.

Gail Rosen received both a B.S. (highest honors) and M.S. in electrical engineering from the Georgia Institute of Technology (Georgia Tech), in 1999 and 2002, respectively. Since 2002, she has been pursuing a Ph.D. in electrical engineering at Georgia Tech's Center for Signal and Imaging Processing (CSIP).

She is a recipient of numerous awards, including a National Science Foundation (NSF) Graduate Research Fellowship, an NSF STEP Fellowship, an AT&T Research Laboratories Grant, and a Georgia Tech ECE Outstanding Teaching Award. Her main research interests are reverse-engineering biological systems and analyzing DNA structure, mutations, and repair.

Address for Correspondence: Gail Rosen, Center for Signal and Image Processing, Georgia Institute of Technology, Atlanta, GA 30332-0250 USA. E-mail: gailr@ece.gatech.edu.

References

[1] Y.K. Huen, "Brief comments on junk DNA: Is it really junk?" *Complexity Int.*, vol. 9, pp. 1–12, 2002.

[2] R.R. Sinden, V.N. Potaman, E.A. Oussatcheva, C.E. pearson, Y.L. Lyubchenko, and L.S. Shlyakhtenko, "Triplet repeat DNA structures and human genetic disease: Dynamic mutations from dynamic DNA," *J. Biosci.*, vol. 27, no. 1, Feb. 2002.

[3] N. Stambuk, "On circular coding properties of gene and protein sequences," *Croatia Chemica Acta*, vol. 72, no. 4, pp. 999–1008, 1999.

[4] E.N. Trifonov, "3-, 10.5-, 200-, and 400-base periodcities in genome sequences," *Physica A*, vol. 249, no. 1–4, pp. 511–516, 1998.

[5] B. Windle, "Telomerase: Target of immortality" [Online]. Available: http://www.people.vcu.edu/~bwindle/Telomerase/telomerase.html

[6] W. Hahn. "Telomerase and cancer: Where and when?" *Clinical Cancer Res.*, vol. 7, no. 10, pp. 2953–2954, Oct. 2001.

[7] T.D. Schneider, G.D. Stormo, L. Gold, and A. Ehrenfeucht, "Information content of binding sites on nucleotide sequences," *J. Mol. Biol.*, vol. 188, no. 3, pp. 415–431, 1986.

[8] R.H. Burdon, *Genes and the Environment*. Philadelphia, PA: Taylor and Francis, 1999.

[9] D. Metzgar, J. Bytof, and C. Wills, "Selection against frameshift mutations limits. Microsatellite expansion in coding DNA," *Genome Res.,* vol. 10, no. 1, pp. 72–80, 2000.

[10] D.A. Mac Donaill, "The role of error-coding in shaping the nucleotide alphabet: Nature's choice of A, U, C, and G," in *Proc. IEEE EMBS Intl. Conf. Special Session Commun. Theory*, *Coding Theory Molecular Biol.*, pp. 3850–3853, 2003.

[11] D. Anastassiou, "Genomic signal processing," *IEEE Signal Processing Mag.*, pp 8–20, July 2001.

[12] N. Chakravarthy, A. Spanias, L.D. Iasemidis, and K. Tsakalis, "Autoregressive modeling and feature analysis of DNA sequences," *Eurasip Journal on Applied Signal Processing*, vol. 1, pp. 13–28, 2004.<

[13] G.L. Rosen and J.D. Moore, "Investigation of coding structure in DNA," *IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Apr. 2003.

[14] "GeneMark: A family of gene prediction programs" [Online]. Available: http://opal.biology.gatech.edu/GeneMark/

[15] S.B. Wicker, *Error Control Systems.* Upper Saddle River, NJ: Prentice Hall, 1995.

[16] L.L. Gatlin, *Information Theory and the Living System*. New York: Columbia Univ. Press, 1972.

[17] R. Ramon-Roldan, P. Bernaola-Galvan, and J.L. Oliver, "Application of information theory to DNA sequence analysis: A review," *Pattern Recognition*, vol. 29, no. 7, pp. 1187–1194, 1996.

[18] T.D. Schneider, "Some lessons for molecular biology from information theory," *Entropy Measures, Maximum Entropy Principle and Emerging Applications*. New York: Springer-Verlag, 2003, pp. 229–237.

[19] G. Battail, "Replication decoding revisitied," in *Proc. IEEE Information Theory Workshop*, Apr. 2003.

[20] E.E. May, M.A. Vouk, D.L. Bitzer, and D.I. Rosnick, "An error-correcting code framework for genetic sequence analysis," *J. Franklin Instit.*, vol. 341, no 1–2, pp. 89–109, Jan.–Mar. 2004.

[21] E.E. May, M.A. Vouk, D.L. Bitzer, and D.I. Rosnick, "Coding theory based models for protein translation initiation in prokaryotic organisms," in *Proc. Fifth Int. Workshop Information Processing Cells Tissues (IPCAT)*, 2003.

[22] E.E. May, "Towards a biological coding theory discipline," *New Thesis*, vol. 1, no. 1, pp. 19–37, 2004.

[23] D.C. Schmidt and E.E. May, "Visualizing ECC properties of E. coli K-12 translation initiation sites," in *Proc. 2nd IEEE Workshop Genomic Signal Processing Statistics*, 2004.

[24] L.S. Liebovitch, Y. Tao, A.T. Todorov, and L. Levine, "Is there an error correcting code in the DNA?" *Biophysical J.*, vol. 71, no. 3, pp. 1539–1544, 1996.

[25] E.E. May, "Analysis of coding theory based models for initiating protein translation in prokaryotic organisms," Ph.D. dissertation, NC State Univ., 2002.

[26] GenBank, "National Center for Biotechnology Database" [Online]. Available: http://www. ncbi.nlm.nih.gov

[27] A. Hauth, "Identification of tandem repeats simple and complex pattern structures in DNA," Ph.D. dissertation, Univ. of Madison,WI, 2002.

[28] M. Buchner and S. Janjarasjitt. "Detection and visualization of tandem repeats in DNA sequences," *IEEE Trans. Signal Processing*, vol. 51, no. 9, Sep. 2003.

[29] R. Kolpakov, G. Bana, and G. Kucherov, "mreps Tandem Repeat Finder" [Online]. Available: http://www.loria.fr/mreps/.

[30] G.L. Rosen, "Finding near-periodic DNA regions using a finite-field framework," in *Proc. 2nd IEEE Workshop Genomic Signal Processing Stat.*, May 2004.

[31] G.H. Golub and C.F. Van Loan, *Matrix Computations*, Baltimore, MD: Johns Hopkins Univ. Press, 1996.

[32] A. Hauth [Online]. Available: http://www.cs.wisc.edu/gensoft/beyondTR/static/HSVDJSAT.html