Genomic Signal Processing

Dimitris Anastassiou

ILLUSTRATION: JIM HANKARD

Genomics is a highly cross-disciplinary field that creates paradigm shifts in such diverse areas as medicine and agriculture. It is believed that many significant scientific and technological endeavors in the 21st century will be related to the processing and interpretation of the vast information that is currently revealed from sequencing the genomes of many living organisms, including humans.

Genomic information is digital in a very real sense; it is represented in the form of sequences of which each element can be one out of a finite number of entities. Such sequences, like DNA and proteins, have been mathematically represented by character strings, in which each character is a letter of an alphabet. In the case of DNA, the alphabet is size 4 and consists of the letters A, T, C and G; in the case of proteins, the size of the corresponding alphabet is 20.

As the list of references shows, biomolecular sequence analysis has already been a major research topic among computer scientists, physicists, and mathematicians. The main reason that the field of signal processing does not yet have significant impact in the field is because it deals with numerical sequences rather than character strings. However, if we properly map a character string into one or more numerical sequences, then digital signal processing (DSP) provides a set of novel and useful tools for solving highly relevant problems.

For example, in the form of local texture, color spectrograms visually provide significant information about biomolecular sequences which facilitates understanding of local nature, structure, and function. Furthermore, both the magnitude and the phase of properly defined Fourier transforms can be used to predict important features like the location and certain properties of protein coding regions in DNA. Even the process of mapping DNA into proteins and the interdependence of the two kinds of sequences can be analyzed using simulations based on digital filtering. These and other DSP-based approaches result in alternative mathematical formulations and may provide improved computational techniques for the solution of useful problems in genomic information science and technology.

The Nature of Biomolecular Sequences

There are many good textbooks dealing with the topic of molecular biology. One excellent reference of particular suitability for introduction to the subject, for people outside the field, consists of chapters 6-10 of a simplified version [1] of a classic biology textbook. Here we present a brief summary of the main concepts.

DNA

A single strand of DNA is a biomolecule consisting of many linked, smaller components called nucleotides. Each nucleotide is one of four possible types designated by the letters A, T, C, and G and has two distinct ends, the 5′end and the 3′ end, so that the 5′ end of a nucleotide is linked to the 3′ end of another nucleotide by a strong chemical bond, thus forming a long, one-dimensional chain (backbone) of a specific directionality. Therefore, each DNA

1053-5888/01/$10.00©2001IEEE

single strand is mathematically represented by a character string, which, by convention specifies the 5′ to 3′ direction when read from left to right.

Single DNA strands tend to form double helices with other single DNA strands. Thus, a DNA double strand contains two single strands called *complementary* to each other because each nucleotide of one strand is linked to a nucleotide of the other strand by a chemical bond, so that A is linked to T and vice versa, and C is linked to G and vice versa. Each such bond is weak (contrary to the bonds forming the backbone), but together all these bonds create a stable, double helical structure. The two strands run in opposite directions, as shown in Fig. 1, in which we see the sugar-phosphate chemical structure of the DNA backbone created by strong (covalent) bonds, and that each nucleotide is characterized by a base that is attached to it. The two strands are linked by a set of weak (hydrogen) bonds. The bottom left diagram is a simplified, straightened out depiction of the two linked strands.

For example, the part of the DNA double strand shown in Fig. 1 is

```
5′ - C-A-T-T-G-C-C-A-G-T - 3′
3′ - G-T-A-A-C-G-G-T-C-A - 5′
```

Because each of the strands of a DNA double strand uniquely determines the other strand, a double-stranded DNA molecule is represented by either of the two character strings read in its 5′ to 3′ direction. Thus, in the example above, the character strings CATTGCCAGT and ACTGGCAATG can be alternatively used to describe the same DNA double strand, but they specify two different single strands which are complementary to each other. DNA strands that are complementary to themselves are called self-complementary, or *palindromes*. For example AATCTAGATT is a palindrome.

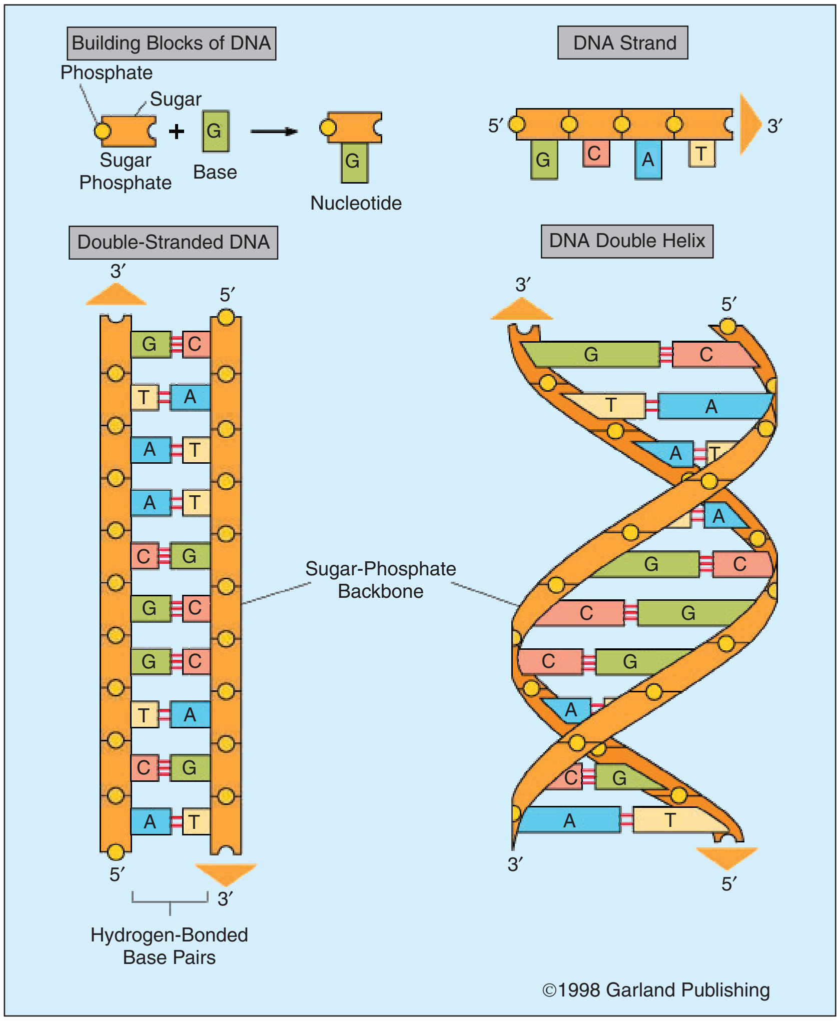
DNA molecules store the digital information that constitutes the genetic blueprint of living organisms. This digital information has been created and reliably stored throughout billions of years of evolution during which some vital regions of DNA sequences have been remarkably preserved, despite striking differences in the body plans of various animals. Fig. 2 compares two related DNA sections between humans and whales; the identical nucleotides are highlighted.

Proteins

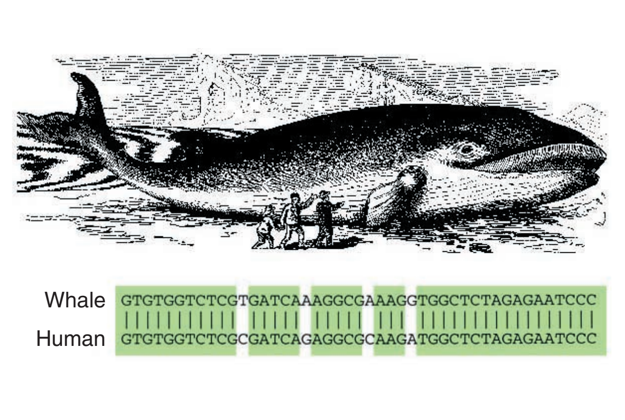
A protein is also a biomolecule consisting of many linked, smaller components called amino acids. There are 20 possible types of amino acids in proteins and, just as is the case in DNA single strands, they are connected with strong bonds, one after the other, forming a long one-dimensional chain (backbone) of a specific directionality. Therefore, as in DNA, a character string mathematically represents each protein. The length of a character string representing a protein molecule is relatively small, typically in the hundreds, while the length of a character string representing a DNA molecule in the living cell is typically in the millions, or even hundreds of millions.

Protein molecules tend to fold into complex three-dimensional (3-D) structures forming weak bonds between their own atoms, and they are responsible for carrying out nearly all of the essential functions in the living cell by properly binding to other molecules with a number of chemical bonds connecting neighboring atoms. Thus, protein functions are largely determined by their 3-D structures because these geometrical shapes often determine whether a protein can bind to another molecule by a process reminiscent of a hand fitting into a glove.

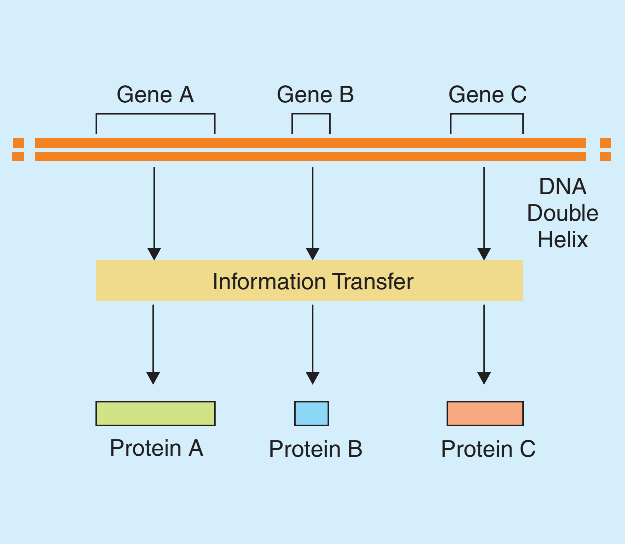
Although we do not yet know how to reliably predict protein 3-D structures from their one-dimensional amino acid sequences (the person who solves this problem will



▲ 1. *DNA and its building blocks (©2001 from* Essential Cell Biology *by Alberts et al. Reproduced by permission of Routledge, Inc., part of The Taylor & Francis Group).*


Whale
GTGTGGTCTCGTGATCAAAGGCGAAAGGTGGCTCTAGAGAATCCC
Human
GTGTGGTCTCGCGATCAGAGGCGCAAGATGGCTCTAGAGAATCCC

▲ *2. Comparison of a section of the sex determination gene from two different animals (©2001 from* Essential Cell Biology *by Alberts et al. Reproduced by permission of Routledge, Inc., part of The Taylor & Francis Group).*



▲ *3. Each gene contains the information to make a protein (©2001 from* Essential Cell Biology *by Alberts et al. Reproduced by permission of Routledge, Inc., part of The Taylor & Francis Group).*

probably receive a Nobel prize), we do know that nearly all proteins in the living cell are uniquely determined by these sequences. Therefore, the amino acid character strings determine the functions of proteins.

In fact, protein functions are ultimately determined by the DNA character string because it is the digital information in the DNA nucleotide sequences that determine the amino acid sequences; each protein character string is generated based on information in genes, which are regions in the DNA character strings. This process is shown schematically in Fig. 3 in which, for simplicity, the intermediate role of another biomolecule (RNA) is omitted, as is the fact that sometimes the same gene may code for multiple proteins through a process called *alternative splicing* which is discussed later.

Protein synthesis is governed by the genetic code which maps each of the 64 possible triplets (codons) of DNA characters into one of the 20 possible amino acids (or into a punctuation mark, like a stop codon, signaling termination of protein synthesis). Fig. 4 shows the genetic code in which the 20 amino acids are designated by both their one-letter and three-letter symbols. A particular triplet, ATG, serves as the START codon and it also codes for the M amino acid (methionine); thus, methionine appears as the first amino acid of proteins, but it may also appear in other locations. We also see that there are three STOP codons indicating termination of amino acid chain synthesis, and the last amino acid is the one generated by the codon preceding the STOP codon.

Coding of nucleotide triplets into amino acids can happen in either the forward or the reverse direction based on the complementary DNA strand. Therefore, there are six possible reading frames for protein coding DNA regions. For example, if the ten nucleotide pairs of the DNA segment shown in Fig. 1 are within protein coding regions, then there are six possibilities for the codons:

```
CAT TGC CAG T..
.CA TTG CCA GT.
..C ATT GCC AGT
ACT GGC AAT G..
.AC TGG CAA TG.
..A CTG GCA ATG
```

AAA: K (Lys)	GAA: E (Glu)	TAA: STOP	CAA: Q (Gln)
AAG: K (Lys)	GAG: E (Glu)	TAG: STOP	CAG: Q (Gln)
AAT: N (Asn)	GAT: D (Asp)	TAT: Y (Tyr)	CAT: H (His)
AAC: N (Asn)	GAC: D (Asp)	TAC: Y (Tyr)	CAC: H (His)
AGA: R (Arg)	GGA: G (Gly)	TGA: STOP	CGA: R (Arg)
AGG: R (Arg)	GGG: G (Gly)	TGG: W (Trp)	CGG: R (Arg)
AGT: S (Ser)	GGT: G (Gly)	TGT: C (Cys)	CGT: R (Arg)
AGC: S (Ser)	GGC: G (Gly)	TGC: C (Cys)	CGC: R (Arg)
ATA: I (Ile)	GTA: V (Val)	TTA: L (Leu)	CTA: L (Leu)
ATG: M (Met)/START	GTG: V (Val)	TTG: L (Leu)	CTG: L (Leu)
ATT: I (Ile)	GTT: V (Val)	TTT: F (Phe)	CTT: L (Leu)
ATC: I (Ile)	GTC: V (Val)	TTC: F (Phe)	CTC: L (Leu)
ACA: T (Thr)	GCA: A (Ala)	TCA: S (Ser)	CCA: P (Pro)
ACG: T (Thr)	GCG: A (Ala)	TCG: S (Ser)	CCG: P (Pro)
ACT: T (Thr)	GCT: A (Ala)	TCT: S (Ser)	CCT: P (Pro)
ACC: T (Thr)	GCC: A (Ala)	TCC: S (Ser)	CCC: P (Pro)

▲ *4. The genetic code.*

In addition to protein coding regions, DNA contains regions serving regulatory functions as well as regions serving yet unknown functions.

One of the most relevant and yet unsolved problems in bioinformatics is to accurately and automatically annotate sequences by identifying such regions using gene prediction [6], [9], [25]. From the above discussion it is clear that the total number of nucleotides in the protein coding area of a gene will be a multiple of three, that the area will be bounded by a START codon and a STOP codon, and that there will be no other STOP codon in the coding reading frame in between. However, given a long nucleotide sequence, it is very difficult to accurately designate where the genes are.

Accurate gene prediction becomes further complicated by the fact that, in advanced organisms, protein coding regions in DNA are typically separated into several isolated subregions called *exons*. The regions between two successive exons are called *introns,* and they are eliminated before protein coding through a process called *splicing*. For example, the three highlighted areas in Fig. 5 indicate the locations of the three exons inside the protein coding region of the human ß-globin gene. Note that the first highlighted region starts from a START codon (ATG), and that the last highlighted region ends at a STOP codon (TAA), and that the number of all nucleotides in the highlighted regions is a multiple of three.

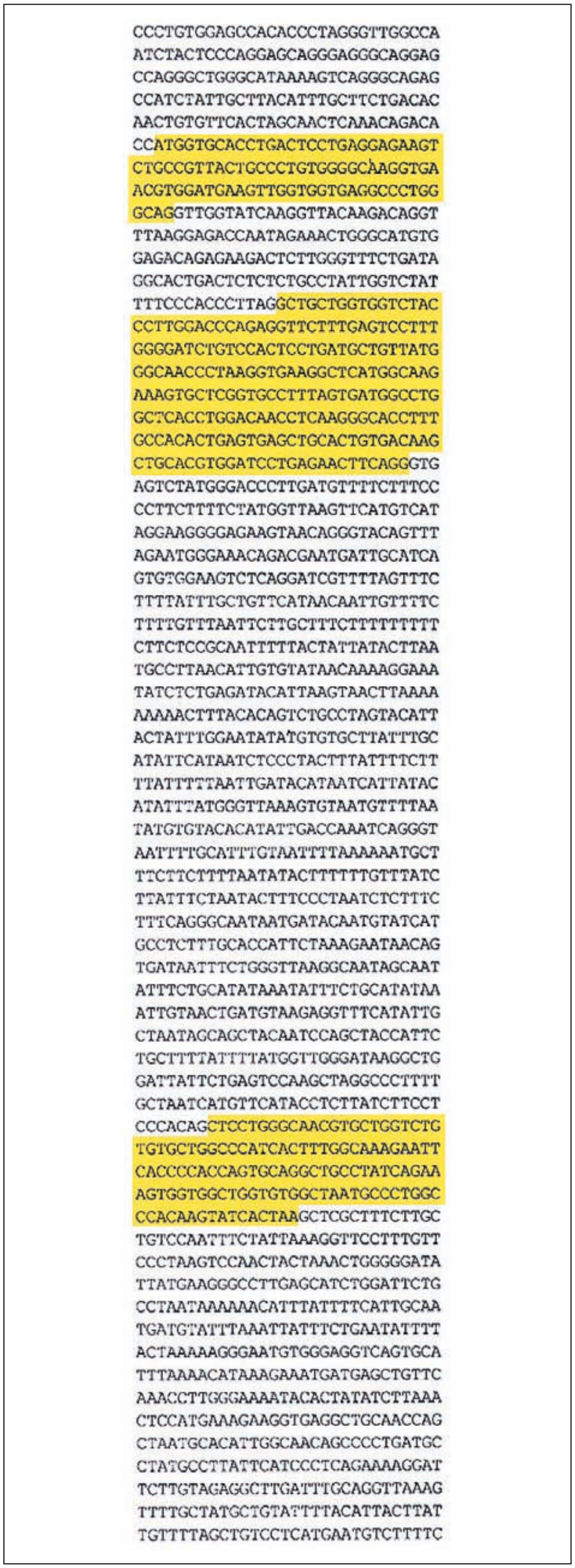
To make things more complicated, sometimes genes of advanced organisms can be spliced in a number of distinct ways to produce different proteins, depending on the cell type in which the gene is being expressed, or the stage of development of the organism. This is called *alternative splicing* and enables several proteins to be produced from the same gene by allowing several ways for some of the exons to be stitched together resulting in different proteins. This property was recently invoked in reference to the fact that the total number of genes in the human genome is lower than originally expected, although the number of different proteins in humans is much higher.

Public Databases

Most of the identified genomic data is publicly available over the Web at various places worldwide, one of which is the Entrez search and retrieval system of the National Center for Biotechnology Information (NCBI) at the National Institutes of Health (NIH). The NIH nucleotide sequence database is called GenBank and contains all publicly available DNA sequences. For example, one can go to http://www.ncbi.nlm.nih.gov/entrez and identify the DNA sequence with Accession Number AF 099922; choose Nucleotide under Search and then fill out the other entry by typing: AF 099922 [Accession] and pressing "Go." Clicking on the resulting accession number will show the annotation for the genes as well as the whole nucleotide sequence in the form of raw data. Similarly, Entrez provides access to databases of protein sequences as well as 3-D macromolecular structures, among other options. As another example, a specialized repository for the processing and distribution of 3-D, macromolecular structures can be found in the Protein Data Bank at www.rcsb.org.

Gene Regulation

A magical interplay between proteins and DNA is responsible for many of the essential processes inside all living cells. Typically, each gene is being activated or *expressed* (starting the process that will eventually lead to

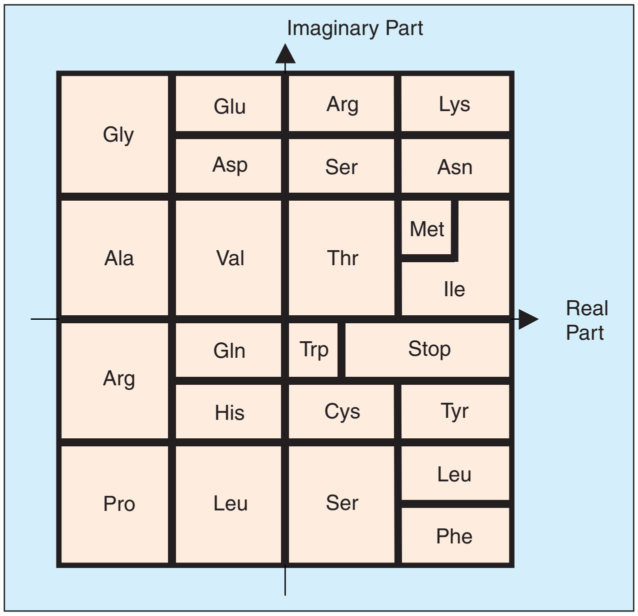


```
CCCTGTGGAGCCACACCCTAGGGTTGGCCA
ATCTACTCCCAGGAGCAGGGAGGGCAGGAG
CCAGGGCTGGGCATAAAAGTCAGGGCAGAG
CCATCTATTGCTTACATTTGCTTCTGACAC
AACTGTGTTCACTAGCAACTCAAACAGACA
CCATGGTGCACCTGACTCCTGAGGAGAAGT
CTGCCGTTACTGCCCTGTGGGGCAAGGTGA
ACGTGGATGAAGTTGGTGGTGAGGCCCTGG
GCAGGTTGGTATCAAGGTTACAAGACAGGT
TTAAGGAGACCAATAGAAACTGGGCATGTG
GAGACAGAGAAGACTCTTGGGTTTCTGATA
GGCACTGACTCTCTCTGCCTATTGGTCTAT
TTTCCCACCCTTAGGCTGCTGGTGGTCTAC
CCTTGGACCCAGAGGTTCTTTGAGTCCTTT
GGGGATCTGTCCACTCCTGATGCTGTTATG
GGCAACCCTAAGGTGAAGGCTCATGGCAAG
AAAGTGCTCGGTGCCTTTAGTGATGGCCTG
GCTCACCTGGACAACCTCAAGGGCACCTTT
GCCACACTGAGTGAGCTGCACTGTGACAAG
CTGCACGTGGATCCTGAGAACTTCAGGGTG
AGTCTATGGGACCCTTGATGTTTTCTTTCC
CCTTCTTTTCTATGGTTAAGTTCATGTCAT
AGGAAGGGGAGAAGTAACAGGGTACAGTTT
AGAATGGGAAACAGACGAATGATTGCATCA
GTGTGGAAGTCTCAGGATCGTTTTAGTTTC
TTTTATTTGCTGTTCATAACAATTGTTTTC
TTTTGTTTAATTCTTGCTTTCTTTTTTTTT
CTTCTCCGCAATTTTTACTATTATACTTAA
TGCCTTAACATTGTGTATAACAAAAGGAAA
TATCTCTGAGATACATTAAGTAACTTAAAA
AAAAACTTTACACAGTCTGCCTAGTACATT
ACTATTTGGAATATATGTGTGCTTATTTGC
ATATTCATAATCTCCCTACTTTATTTTCTT
TTATTTTTAATTGATACATAATCATTATAC
ATATTTATGGGTTAAAGTGTAATGTTTTAA
TATGTGTACACATATTGACCAAATCAGGGT
AATTTTGCATTTGTAATTTTAAAAAATGCT
TTCTTCTTTTAATATACTTTTTTGTTTATC
TTATTTCTAATACTTTCCCTAATCTCTTTC
TTTCAGGGCAATAATGATACAATGTATCAT
GCCTCTTTGCACCATTCTAAAGAATAACAG
TGATAATTTCTGGGTTAAGGCAATAGCAAT
ATTTCTGCATATAAATATTTCTGCATATAA
ATTGTAACTGATGTAAGAGGTTTCATATTG
CTAATAGCAGCTACAATCCAGCTACCATTC
TGCTTTTATTTTATGGTTGGGATAAGGCTG
GATTATTCTGAGTCCAAGCTAGGCCCTTTT
GCTAATCATGTTCATACCTCTTATCTTCCT
CCCACAGCTCCTGGGCAACGTGCTGGTCTG
TGTGCTGGCCCATCACTTTGGCAAAGAATT
CACCCCACCAGTGCAGGCTGCCTATCAGAA
AGTGGTGGCTGGTGTGGCTAATGCCCTGGC
CCACAAGTATCACTAAGCTCGCTTTCTTGC
TGTCCAATTTCTATTAAAGGTTCCTTTGTT
CCCTAAGTCCAACTACTAAACTGGGGGATA
TTATGAAGGGCCTTGAGCATCTGGATTCTG
CCTAATAAAAAACATTTATTTTCATTGCAA
TGATGTATTTAAATTATTTCTGAATATTTT
ACTAAAAAGGGAATGTGGGAGGTCAGTGCA
TTTAAAACATAAAGAAATGATGAGCTGTTC
AAACCTTGGGAAAATACACTATATCTTAAA
CTCCATGAAAGAAGGTGAGGCTGCAACCAG
CTAATGCACATTGGCAACAGCCCCTGATGC
CTATGCCTTATTCATCCCTCAGAAAAGGAT
TCTTGTAGAGGCTTGATTTGCAGGTTAAAG
TTTTGCTATGCTGTATTTTACATTACTTAT
TGTTTTAGCTGTCCTCATGAATGTCTTTTC
```

▲ *5. Sequence of nucleotides in the human β-globin gene (©2001 from* Essential Cell Biology *by Alberts et al. Reproduced by permission of Routledge, Inc., part of The Taylor & Francis Group).*

protein synthesis) as a result of the combined presence, or absence, of certain particular regulatory proteins which bind to specific sites belonging to regulatory regions in DNA (usually in the vicinity of the gene) in a sequence-specific manner. DNA regulatory regions can be as short as ten nucleotide pairs in simple organisms, but can be thousands of nucleotide pairs in more advanced organisms; these nucleotide pairs store some complex digital logic involving chemical binding to complexes of multiple molecules, including several regulatory proteins. Again, chemical binding is dependent on the sequence-specific, 3-D structure of the macromolecules. Deciphering this digital logic in regulatory regions has proved to be a much more challenging task compared to the discovery of the genetic code governing coding DNA regions. We still know very little about these sophisticated regulatory mechanisms that govern the rates of activations of each of the genes.

Things become more complex, and more interesting, by the fact that each of the regulatory proteins are synthesized from other genes, which in turn were activated in relation to another set of regulatory proteins, and so on. A complex system can be defined by a network of many mutually interacting genes; the chemical product of each of these genes influences the activation of other genes in the network. One way of attempting to model this system is by using a set of nonlinear, differential equations involving concentrations of several proteins and other molecules that participate in related pathways. The output of such a system is a script involving the coordinated activation events of many genes; the precise timing of several such events during the lifecycle of the cell plays a crucial role. Even referring to primitive organisms, the term Bacterial Nanobrain [24] has already been used to describe such networks which are indeed described as complex, generalized, artificial neural networks. Such gene regulatory networks are in the heart of genomic information processing, and their analysis is one of the most exciting future topics of research that will require a systems-based approach involving cross-disciplinary collaboration at various levels of abstraction, including a genomic level, a macromolecular binding level, and a higher network level.



▲ *6. The genetic code on the complex plane.*

Character Strings Described by Numerical Sequences

In a DNA sequence of length N, assume that we assign the numbers a, t, c, g to the characters A, T, C, G, respectively. A proper choice of the numbers a, t, c and g can provide potentially useful properties to the numerical sequence $x[n]$.

For example, if we choose complex conjugate pairs $t = a^*$ and $g = c^*$, then the complementary DNA strand is represented by

$$\tilde{x}[n] = x^*[-n+N-1], \quad n = 0,1,\ldots,N-1 \qquad (1)$$

and, in this case, all palindromes will yield conjugate, symmetric numerical sequences which have interesting mathematical properties, including generalized linear phase.

One such assignment (the simplest out of many possible ones) is the following:

$$a = 1 + j, \quad t = 1 - j, \quad c = -1 - j, \quad g = -1 + j. \qquad (2)$$

We may also assign numerical values to amino acids by modeling the protein coding process as an FIR digital filter, in which the input $x[n]$ is the numerical nucleotide sequence, and the output $y[n]$ is the possible resulting numerical amino acid sequence (if $x[n]$ is within a coding region in the proper reading frame):

$$y[n] = h[0]\,x[n] + h[1]\,x[n-1] + h[2]\,x[n-2]. \qquad (3)$$

For example, if we set $h[0] = 1$, $h[1] = 1/2$, and $h[2] = 1/4$, and $x[n]$ is defined by the parameters in (2), then $y[n]$ can only take one out of 64 possible values.

Furthermore, if for example, $x[n]$ corresponds to a forward coding DNA sequence in the first reading frame (i.e., if $x[0]$, $x[1]$, $x[2]$ corresponds to the first codon), then the elements of the output subsequence:

$$y[2],\ y[5],\ y[8], y[11], \ldots, y[N-1]$$

are complex numbers representing each of the amino acids of the resulting protein. In fact, the entire genetic code can be drawn on the complex plane as shown in Fig. 6, in which the center of the square, labeled Met (coded by ATG), is the complex number $(1+j) + 0.5(1-j) + 0.25(-1+j) = 1.17 + 0.88j$.

Each of the entries in Fig. 6 correspond to one of the 20 amino acids or the STOP codon. Therefore, the protein coding process can be simulated by a digital low-pass filter, followed by subsampling via a three-band polyphase decomposition, followed by a switch selecting one of the three bands (reading frames), followed by a vector quantizer as defined in Fig. 6.

In the frequency domain, because of (3), the Fourier transform of the sequence $y[n]$ will be the product of the Fourier transforms of $x[n]$ and of the known finite-duration sequence $h[n]$. Therefore, we can use existing knowledge about the polyphase components to relate the frequency spectra of proteins with those of nucleic acids. Frequency domain, or correlation analysis of nucleotide sequences, has already been recognized as an important tool in bioinformatics by authors outside the DSP community [4], [5], [11], [12], [15], [25], [26].

In other words, certain useful frequency-domain properties of proteins can be evaluated from the corresponding frequency-domain properties of nucleic acids. In the field of multirate signal processing there are several results and equations connecting the frequency spectra of polyphase components, which more accurately relate the frequency spectra of proteins with those of nucleic acids, providing a novel computational framework. Of course, it is possible that other choices of the parameters a, t, c, g, and of the FIR coefficients, may provide a better fit with actual data when solving such bioinformatics problems as alignment of nucleotide or amino acid sequences.

A new field of computer science, bioinformatics, has emerged, focusing on the use of computers for efficiently deriving, storing, and analyzing these character strings to help solve problems in molecular biology.

DNA Spectrograms

It is well known that the appearance of spectrograms provides significant information about signals, to the extent that trained observers can figure out the words uttered in voice signals by simple visual inspection of their spectrograms. Similarly, it appears that spectrograms are powerful visual tools for biomolecular sequence analysis [2], [3].

Here we present a proof-of-concept discussion defining a spectrogram as the display of the magnitude of the short-time Fourier transform (STFT), using the discrete Fourier transform (DFT) as a simple example of a frequency-domain analysis tool. For a numerical sequence $x[n]$ of length N, the DFT $X[k]$ is another sequence of the same length N, providing a measure of the frequency content at frequency k, which corresponds to an underlying period of N/k samples, where the maximum frequency (period 2) corresponds to $k = N/2$, assuming that N is even. The STFT results by applying the DFT over a sliding window of small width to a long sequence, thus providing a localized measure of the frequency content.

In the case of biomolecular sequences, we want these spectrograms to simultaneously provide local frequency information for all four bases; therefore, it is best to avoid using just one choice of assigned numbers a, t, c, g to the characters A, T, C, G, respectively.

We observe that

$$x[n] = au_A[n] + tu_T[n] + cu_C[n] + gu_G[n], \quad n = 0,1,2,\ldots, N-1$$

in which $u_A[n]$, $u_T[n]$, $u_C[n]$, and $u_G[n]$ are called the *binary indicator sequences* which [27] take the value of either one or zero at location n, depending on whether or not the corresponding character exists at location n. For example, in the sequence CATTGCCAGT (one of the strands shown in Fig. 1) all, except for three, of the values of $u_C[n]$ for $0 \le n \le 9$ are 0, and the exceptions are $u_C[0] = u_C[5] = u_C[6] = 1$.

Therefore

$$X[k] = aU_A[k] + tU_T[k] + cU_C[k] + gU_G[k], \quad k = 0,1,\ldots, N-1 \tag{4}$$

For pure DNA character strings (i.e., without assigning numerical values), the sequences $U_A[k]$, $U_T[k]$, $U_C[k]$, and $U_G[k]$ provide a four-dimensional representation of the frequency spectrum of the character string. The quantity

$$S[k] = |U_A[k]|^2 + |U_T[k]|^2 + |U_C[k]|^2 + |U_G[k]|^2 \tag{5}$$

combines the contributions of all four characters, and has been used as a measure of the total spectral content of the DNA character string, at frequency k [25], [22], [15].

Any three of these four binary indicator sequences are sufficient to determine the DNA character string, because they add to 1 for all n, which implies that the corresponding four DFT sequences are also a redundant set:

$$U_A[k] + U_T[k] + U_C[k] + U_G[k] = \begin{cases} 0, & k \ne 0 \\ N, & k = 0. \end{cases} \tag{6}$$

If we wish to reduce the dimensionality from four to three in a manner that is symmetric with respect to all four components, we may adopt the technique [22], in which each of the four letters is assigned to a vertex of a regular tetrahedron in space. Thus, three numerical sequences x_r, x_g, and x_b are defined from corresponding coefficients (a_r, t_r, c_r, g_r), (a_g, t_g, c_g, g_g), (a_b, t_b, c_b, g_b) by considering the four 3-D vectors having magnitude equal to 1 and pointing to the four directions from the center to the vertices of the tetrahedron. For example, we can choose $(a_r, a_g, a_b) = (0,0,1)$, $(t_r, t_g, t_b) = (2\sqrt{2}/3, 0, -1/3)$, $(c_r, c_g, c_b) = (-\sqrt{2}/3, \sqrt{6}/3, -1/3)$, $(g_r, g_g, g_b) = (-\sqrt{2}/3, -\sqrt{6}/3, -1/3)$, hence:

$$x_r[n] = \frac{\sqrt{2}}{3}(2u_T[n] - u_C[n] - u_G[n])$$

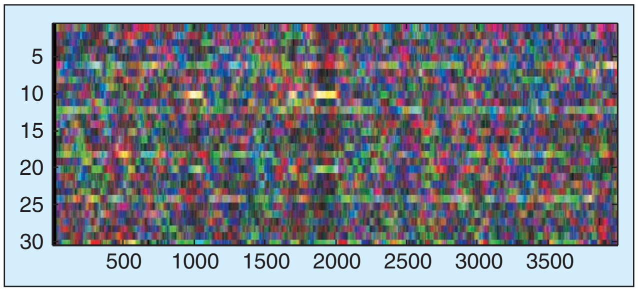
$$x_g[n] = \frac{\sqrt{6}}{3}(u_C[n] - u_G[n])$$

$$x_b[n] = \frac{1}{3}(3u_A[n] - u_T[n] - u_C[n] - u_G[n])$$

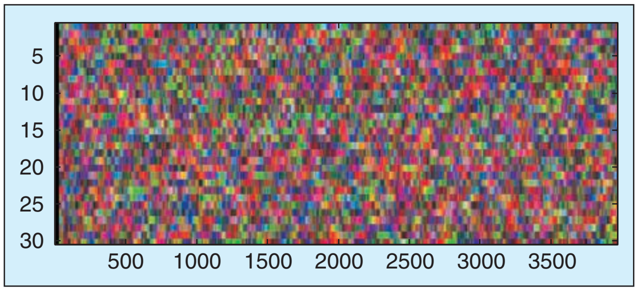
from which we can find the DFTs $X_r[k]$, $X_g[k]$, $X_b[k]$. It has been shown [7] that the power spectra of the two methods (using three or four dimensions) are essentially the same.

We now define the spectrograms of biomolecular sequences that simultaneously provide local frequency information for all four bases by displaying the resulting three magnitudes by superposition of the corresponding three primary colors, red for $|X_r[k]|$, green for $|X_g[k]|$, and blue for $|X_b[k]|$. Thus, color conveys real information, as opposed to pseudocolor spectrograms, in which color is used for contrast enhancement. For example, Fig. 7 shows a spectrogram using DFTs of length 60 of a DNA stretch of 4,000 nucleotides from chromosome III of *C. elegans* (GenBank Accession number NC 000967).

The vertical axis corresponds to the frequencies k from 1 to 30, while the horizontal axis shows the relative nucleotide locations starting from nucleotide 858,001; only frequencies up to $k = 30$ are shown due to conjugate symmetry as x_r, x_g, and x_b are real sequences. The DNA stretch contains three regions (*C. elegans* telomere-like hexamer repeats) at relative locations (953-1066), (1668-1727), and (1807-2028). These three regions are well depicted as bars of high-intensity values corresponding to the particular frequency $k = 10$ (because period 6 corresponds to $N/6 = 10$). Other frequencies also appear to play a prominent role in the whole region of the 4,000 nucleotides.



▲ *7. Color spectrogram of a DNA stretch.*



▲ *8. Color spectrogram of "totally random" DNA.*

For comparison purposes, Fig. 8 shows the texture of a spectrogram coming from a sample of totally random DNA, i.e., in which each type of nucleotide appears with probability 0.25 and independent of the other nucleotides.

In other examples, we have noted the presence of other periodicities (including 10-11 periodicities) and features, indicative of structural patterns. Such periodicities have been observed before [12]-[14], [28], [23], [26].

Of course, there are numerous other ways in which spectrograms can be defined. We may use tapered windows, and adjust their width and shape. Furthermore, more balanced spectrograms can be defined using the wavelet transform rather than the DFT. The wavelet transform has been used to analyze some fractal scaling properties of DNA sequences [4].

Identification of Protein Coding DNA Regions

We now show how frequency-domain analysis of DNA sequences can be a powerful tool for specifically identifying protein coding regions in DNA sequences. The DFT frequency $k = N/3$ corresponds to a period of three samples the length of each codon. It is known [9], [5], [25] that the spectrum of protein coding DNA typically has a peak at that frequency. For example, in Fig. 9 we have plotted the sequence $S[k]$, as defined in (5), for a coding region of length $N = 1320$ inside the genome of the baker's yeast (formally known as *S. cerevisiae*), demonstrating a peak at frequency $k = 440$.

If we define the following normalized DFT coefficients at frequency $k = N/3$:

$$W = \frac{1}{N}X\left[\frac{N}{3}\right]$$
$$A = \frac{1}{N}U_A\left[\frac{N}{3}\right], \quad T = \frac{1}{N}U_T\left[\frac{N}{3}\right],$$
$$C = \frac{1}{N}U_C\left[\frac{N}{3}\right], \quad G = \frac{1}{N}U_G\left[\frac{N}{3}\right] \tag{7}$$

then it follows from (4), with $k = N/3$, that:

$$W = aA + tT + cC + gG. \tag{8}$$

In other words, for each DNA segment of length N (where N is a multiple of three), and for each choice of the parameters a, t, c and g, there corresponds a complex number $W = aA + tT + cC + gG$.

A complex system can be defined by a network of many mutually interacting genes; the chemical product of each of these genes influences the activation of other genes in the network.

We have found [2] that, for properly chosen values of a, t, c, and g, the magnitude of W is a superior predictor, compared to $S[N/3]$ defined in (5), of whether or not the DNA segment is part of a protein coding region; and that, in the former case, the phase $\Theta = \arg\{W\}$ is a powerful predictor of the reading frame that it belongs.

For each DNA segment, there corresponds a set of complex numbers A, T, C, and G, as defined in (7), in which $A + T + C + G = 0$, because of (6). These quantities can be thought of as complex random variables. They have quite different probabilistic characteristics depending on whether or not the DNA sequence is part of a protein coding region, as well as on the corresponding reading frame. Under this interpretation, the quantity W, as defined in (8), is a complex random variable itself, and its properties depend on the particular choice of the parameters a, t, c, and g.

We can quantify the statistical properties of the random variables A, T, C, and G for protein coding regions by collecting statistics from a large sample. For example, we can consider [2] chromosome XVI of *S. cerevisiae* (GenBank accession number NC 001148). We isolated all genes for which there were no introns and for which the evidence was labeled "experimental." If the orientation of a gene was complementary, then we properly transformed its values as if it were a forward coding gene (i.e., starting from the codon ATG). For each of the chosen genes, we evaluated the corresponding numbers A, T, C, and G, thus creating a set of statistical samples. We found that, for that particular chromosome, the average values of A, T, C, and G, scaled by 10^3, were $8.0 - 56.3j$, $-84.1 + 37.4j$, $-46.2 - 23.2j$, and $122.3 + 42.1j$. By comparison, the magnitudes of A, T, C, and G, for nonprotein coding regions are much smaller, typically between one and two.

There have been many proposed protein coding measures used for gene identification [10], [6]. Here, we describe how we can predict whether or not a given DNA segment belongs to a reading frame from the magnitude of a properly defined W, i.e., after optimizing the values of the parameters a, t, c, and g.

We want to maximize the discriminatory capability between protein coding regions (with corresponding random variables A, T, C, and G) and random DNA regions. Using a random number generator, we synthesized a random DNA sequence, with the same number and length as the protein coding statistical sample, thus creating a different set of random variables: A_R, T_R, C_R, and G_R.

Because $A + T + C + G = 0$, if we add any constant value to the coefficients a, t, c, and g, then the value of W in (8) remains the same. To define an optimization problem with a unique solution, we first fix one of the four coefficients (c) to the value of 0, so that $W = aA + tT + gG$, and $c = 0$. (We could have reduced dimensionality in a symmetrical manner, but this would not have enhanced predictive power.)

Therefore, the following problem is naturally formulated once we have available a joint probabilistic model for the complex random variables A, T, and G (in our case coming from our measurements from chromosome XVI of *S. cerevisiae*) and for the complex random variables A_R, T_R, and G_R:

Find the complex numbers a, t, and g maximizing the quantity:

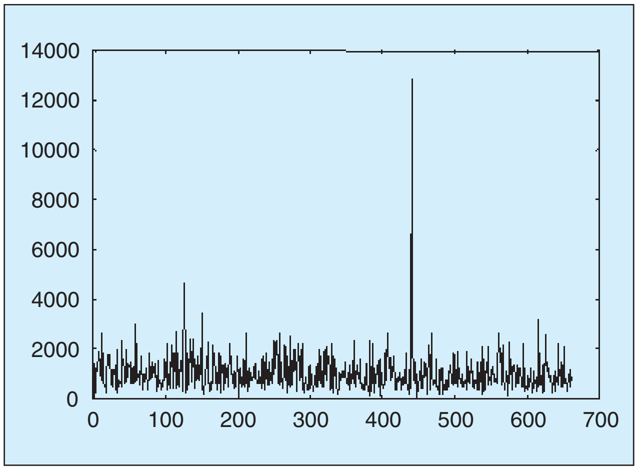
$$p(a, t, g) = \frac{\mathrm{E}\{|aA + tT + gG|\} - \mathrm{E}\{|aA_R + tT_R + gG_R|\}}{\mathrm{std}\,(|aA + tT + gG|) + \mathrm{std}\,(|aA_R + tT_R + gG_R|)}$$

(in which std stands for standard deviation) under the constraining conditions (because W is also invariant to rotation and scaling):

$$E\{\arg\{aA + tT + gG\}\} = 0, \qquad |a| + |t| + |g| = 1.$$

The above mathematical problem (and similar ones defined below) can potentially be solved yielding some closed-form solution as a function of certain statistical coefficients. There is no need for this, however, because conventional optimization techniques, based on iterated random perturbations starting from an initial guess, immediately converge to the optimum values.

For this particular example, using the resulting random variables, the solution is:



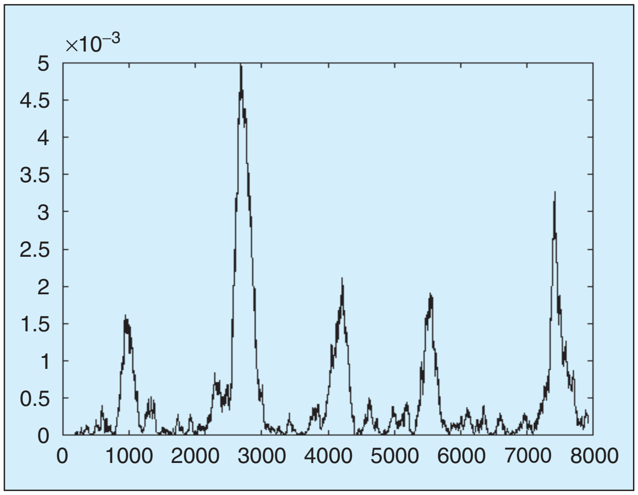
▲ *9. Plot of the spectrum of a coding DNA region, demonstrating peak at frequency k = N / 3.*

$$a = 0.10 + 0.12j \quad t = -0.30 - 0.20j$$
$$c = 0 \quad g = 0.45 - 0.19j \tag{9}$$

corresponding to a value of $p(a, t, g) = 2.18$.

Using the coefficients in (9), we evaluated the magnitude of the 351-point STFT for a DNA stretch of *C. elegans* (GenBank accession number AF 099922), containing 8,000 nucleotides starting from location 7021. The plot of its square is shown in Fig. 10. The DNA stretch contains a gene (F56F11.4) with five exons, all identified by the peaks of the plot, at the positions shown in Table 1, relative to 7021.

In [2], it was shown that the performance of the optimized spectral content measure $|aA + tT + cC + gG|^2$ is significantly superior to that of the traditional one from (5), $|A|^2 + |T|^2 + |C|^2 + |G|^2$, in terms of their capabilities to distinguish between coding and noncoding regions in DNA sequences. The traditional measure, evaluated at $k = N/3$, has been used to create pseudocolor bar renditions accompanying genome maps. Such a bar is included in the journal inset map as a recent result of the Genome Annotation Assessment Project (GASP) for *Drosophila melanogaster* (fruit fly) [19].



▲ *10. Plot of $|aA + tT + cC + gG|^2$ for the five exons shown in Table 1.*

Table 1. Locations and Reading Frames of the Five Exons of the Gene F56F11.4.

Relative Position	Exon Length	Reading Frame
929-1135	207	2
2528-2857	330	2
4114-4377	264	1
5465-5644	180	2
7255-7605	351	1

Reading Frame Identification

We label the three reading frames by the number $\mathrm{mod}(n,3) + 1$ where n is the leftmost location of any codon triplet. According to this definition, the reading frames corresponding to each of the five exons of the gene are given in Table 1.

To distinguish forward-coding reading frames from reverse-coding ones, we augment the notation by including a tilde on the numerical value assigned to a reverse-coding reading frame. For example, the bases at locations (0, 1, 2) form codons at either reading frame 1 or reading frame $\tilde{1}$ depending on whether the codon is forward-coding or reverse-coding. Similarly, the bases at locations (1, 2, 3) form codons at either reading frame 2 or reading frame $\tilde{2}$, and the bases at locations (2, 3, 4) form codons at either reading frame 3 or reading frame $\tilde{3}$.

It has been known [21] that the different reading frames exhibit different statistical characteristics. Here we show that the *phase* Θ of the complex random variable W can be used as the reading frame predictor, making use of the following fact, which can be directly proved from the DFT definition:

Assuming that a DNA segment is part of a forward coding region (reverse coding will be addressed later), we define the angles ϕ_1, ϕ_2, and ϕ_3 to be the expected values of the phase of the random variable $W = aA + tT + cC + gG$ corresponding to the reading frames 1, 2, and 3, respectively. Then, $\mathrm{mod}(\phi_2 - \phi_1) = \mathrm{mod}(\phi_3 - \phi_2) = \mathrm{mod}(\phi_1 - \phi_3) = -2\pi/3$.

If, for example, $\phi_1 = 34°$, then $\phi_3 = 154°$ and $\phi_2 = 274°$, or, equivalently, $\phi_2 = -86°$. Therefore:

If, for a particular choice of the parameters a, t, c, and g, the phase Θ of the complex random variable W has small variance, then the angle Θ will probably take values that will be close to one out of three possible ones, ϕ_1, ϕ_2, and ϕ_3, differing from each other by 120°.

To maximize predictive power, it is desirable to select the parameters a, t, c, and g minimizing some measure of the variability (such as the statistical variance) of Θ.

The definition of a unique meaningful statistical variance of the phase of a complex random variable is complicated by the fact that the phase is not uniquely specified unless restricted to an interval of length 2π, in which case the two ends of the interval correspond to equivalent values. Therefore, we can instead choose the almost equivalent task of maximizing the magnitude of the expected value of the random variable $W/|W| = e^{j\Theta} = \cos\Theta + j\sin\Theta$. We would like that number to be as large, and as close to 1, as possible, because if it is only slightly less than 1, this will imply that $e^{j\Theta}$ is concentrated on a tiny area in the periphery of the unit circle in the complex plane.

As in the previous optimization problem, we reduce the dimensionality of the problem by setting $c = 0$, and we formulate the following optimization problem:

Find the complex numbers a, t, and g maximizing the quantity:

$$q(a,t,g)=\left|E\left\{\frac{aA+tT+gG}{|aA+tT+gG|}\right\}\right| \tag{10}$$

under the constraining conditions (for unique solution):

$$E\{\arg\{aA+tT+gG\}\}=0 \qquad |a|+|t|+|g|=1$$

The solution of the optimization problem, for our data, is given by

$$\begin{aligned} a&=0.26+0.10j & t&=0.03-0.17j \\ c&=0 & g&=0.51-0.21j \end{aligned} \tag{11}$$

corresponding to a value of $q(a,t,g)$, as defined in (10), equal to 0.952, and to a standard deviation of the phase $\Theta = arg\{aA+tT+gG\}$ of 18.2°. This means that the probability that Θ will be within two standard deviations (36.4°) from the mean (0°) is very high.

All statistical data were collected under reading frame 1, and in that case the value of $E\{\Theta\}$ is 0. Therefore, there is a high probability that the value of Θ for reading frame 1 will be within 0°± 36.4°. And so, if the data were corresponding to reading frame 2, there is a high probability that the value of Θ would have been within $-120° \pm 36.4°$. Similarly, if the data were corresponding to reading frame 3, there is a high probability that the value of Θ would have been within $120° \pm 36.4°$. There is still a significant gap between any two of those angular regions.

Because the number of primary colors (red, green, and blue) is the same as the number of possible forward coding reading frames, we can conveniently assign a color-coding scheme in which the value $\Theta = 0°$ is assigned the color red, the value $\Theta = 120°$ is assigned the color blue, and the value $\Theta = -120°$ is assigned the color green. In-between values are color-coded in a linear manner, according to Fig. 11, in which the three axes labeled R, G, and B correspond to the primary colors red, green, and blue.

Color Maps

We use the above color coding for reading frame identification as shown in Table 2.

All STFT windows must be aligned at the same reading frame. Therefore, we have chosen for the sliding window to slide by precisely three locations for each DFT evaluation. Furthermore, we always make sure that the window size is a multiple of three so that the frequency $k = N/3$ is well defined.

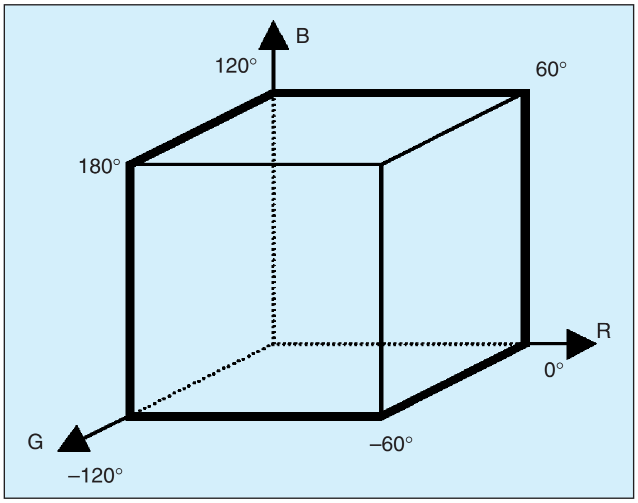
Fig. 10 identifies the five exons based on the magnitude of the STFT using the parameter values of (9). We now use the parameter values of (11) to enrich the information of Fig. 10 in the form of a color map shown in Fig. 12. For each nucleotide location in the color map, the color assigned obeys the rule of Fig. 11, and the intensity is modulated by the square-magnitude multiplied by 700 and clipped to the interval (0, 1).

Note that the color of the third exon is closer to orange than to pure red, but the information is still sufficient to accurately identify its reading frame as 1.

Complementary Sequences

The binary indicator sequences of the complementary DNA strand are

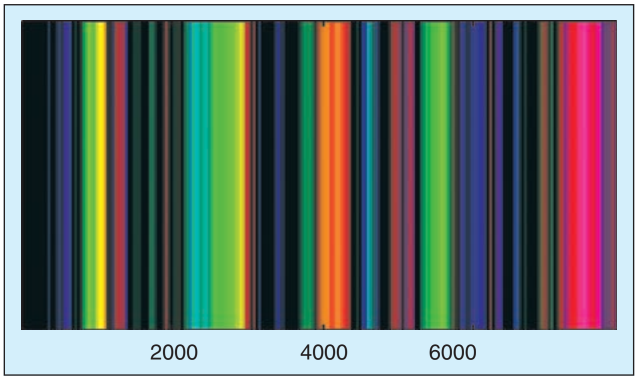
$$\begin{aligned} \tilde{u}_A[n] &= u_T[-n+N-1] \\ \tilde{u}_T[n] &= u_A[-n+N-1] \\ \tilde{u}_C[n] &= u_G[-n+N-1] \\ \tilde{u}_G[n] &= u_C[-n+N-1] \end{aligned}$$



▲ *11. Color coding of the Fourier transform phase*

Table 2. Color-Coded Reading Frame Identification.

Red	Reading frame 1
Green	Reading frame 2
Blue	Reading frame 3



▲ *12. Color map of reading frames for the exons of the gene of Table 1.*

Gene regulation analysis is one of the most exciting research topics that can potentially be addressed using the theory of artificial neural networks.

in which $u_A[n], u_T[n], u_C[n]$, and $u_G[n], n = 0, 1, \ldots, N-1$ are the binary indicator sequences of the corresponding primary DNA strand.

We make use of the following fact which can readily be proved from the DFT definition: If two sequences $x[n]$ and $\tilde{x}[n]$ are related to each other by (1), i.e., if $\tilde{x}[n] = x^*[-n+N-1]$, then

$$\tilde{X}\left[\frac{N}{3}\right] = e^{j\frac{2\pi}{3}} X^*\left[\frac{N}{3}\right]. \qquad (12)$$

We now find the values of A, T, C, and G, as defined in (7), for the numerical sequence of the complementary DNA strand, for which we will use the notation $\tilde{A}, \tilde{T}, \tilde{C}$, and $\tilde{G}$, respectively. If we use the same choice for a, t, c, and g for both strands, it follows from (7) and (12) that

$$\tilde{A} = e^{j\frac{2\pi}{3}} T^*, \ \tilde{T} = e^{j\frac{2\pi}{3}} A^*, \ \tilde{C} = e^{j\frac{2\pi}{3}} G^*, \ \tilde{G} = e^{j\frac{2\pi}{3}} C^* \qquad (13)$$

in which A, T, C, and G are the corresponding values of the primary strand. The value for $\tilde{W} = (1/N)\hat{X}[N/3]$ is:

$$\tilde{W} = a\tilde{A} + t\tilde{T} + c\tilde{C} + g\tilde{G}. \qquad (14)$$

Now, if we define

$$\tilde{a} = e^{-j\frac{2\pi}{3}} t^*, \ \tilde{t} = e^{-j\frac{2\pi}{3}} a^*, \ \tilde{c} = e^{-j\frac{2\pi}{3}} g^*, \ \tilde{g} = e^{-j\frac{2\pi}{3}} c^* \qquad (15)$$

then from (13), (14), and (15) it follows that $\tilde{W} = (\tilde{a}A + \tilde{t}T + \tilde{c}C + \tilde{g}G)^*$.

In conclusion, we can simulate the processing of the complementary strand in the reverse direction with parameter values a, t, c, and g, by processing the primary strand using the values of the parameters $\tilde{a}$, $\tilde{t}$, $\tilde{c}$, $\tilde{g}$ given by (15), and taking the complex conjugate of the resulting W.

It can easily be shown that the identical color code for reading frame identification applies to reading frames $\tilde{1}$, $\tilde{2}$, and $\tilde{3}$ as well.

Example

Consider a DNA stretch from chromosome III of *S. cerevisiae* (GenBank accession number NC 001135). Note that there is no overlap with the collected statistics. The DNA stretch consists of 12,000 nucleotides starting from location 212041. It contains six genes (three forward coding and three reverse coding) at the locations shown in Table 3 relative to 212040.

One problem is that the color map for forward coding will contain some interference from reverse coding regions, and vice versa; you may recall that the parameters a, t, c, and g were optimized to distinguish forward coding regions from noncoding regions and not from reverse coding regions. This problem can be addressed by partitioning the DNA segment into possible forward coding regions and possible reverse coding regions; this approach will fail to detect simultaneously multiple coding areas, but these are rare occasions in most organisms.

Because of (15), the following optimization problem is defined.

Find the parameters a, t, c, and g maximizing the expected value of the following random variable:

$$V = \left|\frac{aA + tT + cC + gG}{t^*A + a^*T + g^*C + c^*G}\right|.$$

To assure unique solution, we may simply pose the constraints $c = 0$ and $g = 1$, in which we find the optimal values $a = 0.049 + 0.149j$ and $t = -0.122 - 0.518j$. The criterion for partitioning is whether or not V is greater or less than one.

The resulting partitioning between forward and reverse coding is another unique feature of our proposed approach compared with existing Fourier analysis tools. Figs. 13 and 14 show the resulting color maps for forward and reverse coding. In Table 3, we see that the six genes were accurately color coded, and we can obtain a sense of the power and the limitations of these tools.

In [2], we provide more details and propose a more sophisticated, soft partitioning scheme by estimating the probabilities $P(F/V)$ and $P(R/V)$ that a particular location belongs to a forward or reverse coding region, given that it belongs to either one, and given the value of V for that location.

Table 3. Locations and Reading Frames of Six Genes.

Relative Location	Gene Length	Reading Frame
761 → 1429	669	2
1687 → 3135	1449	1
3387 → 4931	1545	3
5066 ← 6757	1692	$\tilde{2}$
7147 ← 9918	2772	$\tilde{1}$
10143 ← 10919	777	$\tilde{3}$

Discussion

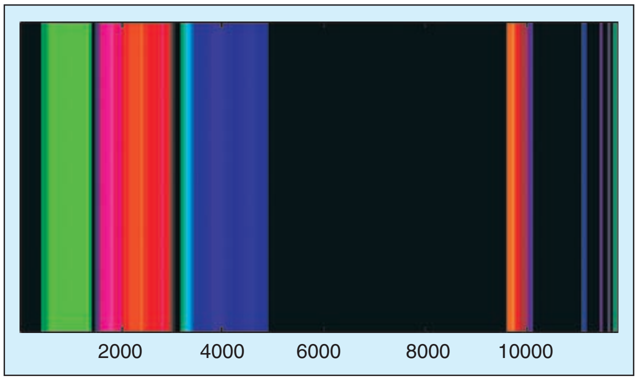
Signal processing-based computational and visual tools are meant to synergistically complement character-string-domain tools that have successfully been used for many years by computer scientists. In this article, we illustrated one of several possible ways that signal processing can be used to directly address biomolecular sequences. The assignment of optimized, complex numerical values to nucleotides and amino acids provides a new computational framework, which may also result in new techniques for the solution of useful problems in bioinformatics, including sequence alignment, macromolecular structure analysis, and phylogeny [8], [20].

An important advantage of DSP-based tools is their flexibility. Spectrograms can be defined in many ways. For example, depending on the particular features that must be emphasized, we may wish to define spectrograms using certain values of parameters. Once a visual pattern appears to exist, we have the opportunity to interactively modify the values of these parameters in ways that will enhance the appearance of these patterns, thus clarifying their significance. It is hoped that visual inspection of spectrograms will establish links between particular visual features (like areas with peculiar texture or color) and certain yet undiscovered motifs of biological sequences.

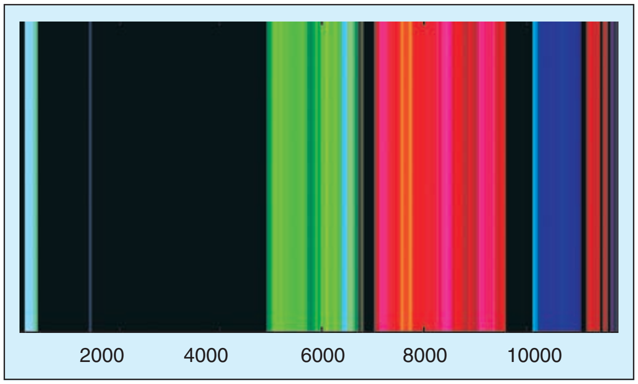
With the explosive growth of the amount of publicly available genomic data, a new field of computer science, bioinformatics, has emerged, focusing on the use of computers for efficiently deriving, storing, and analyzing these character strings to help solve problems in molecular biology. A plethora of computational techniques familiar to the signal processing community has already been used extensively and with significant success in bioinformatics, including such tools as hidden Markov models and neural networks. This is another area in which DSP-based approaches can be of help. For example, in [18] it is shown that a combination of signal processing and statistical pattern recognition methods may significantly improve the base-calling accuracy in DNA sequencing.

Gene regulation analysis is one of the most exciting research topics that can potentially be addressed using the theory of artificial neural networks. The topic of genomics was recently featured in *IEEE Spectrum* magazine [16]. One of the tools providing valuable information about gene expression patterns is the DNA hybridization microarray; this topic was also featured in a recent issue of *IEEE Spectrum* [17]. We believe that there exists a unique opportunity for the DSP community, and the electrical engineering community in general, to play an important role in the emerging field of genomics.

Dimitris Anastassiou received the Ph.D. degree in electrical engineering from the University of California, Berkeley, in 1979. From 1979 to 1983 he was a Research Staff Member at the IBM Thomas J. Watson Research Center, Yorktown Heights, NY. Since 1983, he has been with the Department of Electrical Engineering of Columbia University where he is a Professor and Director of Columbia's Genomic Information Systems Laboratory. He is an IEEE Fellow, the recipient of an IBM Outstanding Innovation Award, a National Science Foundation Presidential Young Investigator Award, and a Columbia University Great Teacher Award. His research is now exclusively focused on applying his expertise in several traditional electrical engineering disciplines to the emerging field of genomics.



▲ *13. Color map for forward coding after partition for the genes shown in Table 3.*



▲ *14. Color map for reverse coding after partition for the genes shown in Table 3.*

References

[1] B. Alberts, D. Bray, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Essential Cell Biology*. New York: Garland Publishing, 1998.

[2] D. Anastassiou, "Frequency-domain analysis of biomolecular sequences," *Bioinformatics*, vol. 16, no. 12, pp. 1073-1082, Dec. 2000.

[3] D. Anastassiou, "Digital signal processing of biomolecular sequences," Tech. Rep. EE000420-1, Apr. 2000. Available: http://www.ee.columbia.edu/cgi-ee-bin/show_archive.pl.

[4] E. Arneodo, E. Bacry, P.V. Graves, and J.F. Muzy, "Characterizing long-range correlations in DNA sequences from wavelet analysis," *Phys. Rev. Lett.*, vol. 74, pp. 3293-3296, 1995.

[5] V.R. Chechetkin and A.Y. Turygin, "Size-dependence of three-periodicity and long-range correlations in DNA sequences," *Phys. Lett. A*, vol. 199, pp. 75-80, 1995.

[6] J.-M. Claverie, "Computational methods for the identification of genes in vertebrate genomic sequences," *Hum. Mol. Genet.*, vol. 6, pp. 1735-1744, 1997.

[7] E. Coward, "Equivalence of two Fourier methods for biological sequences," *J. Math. Biol.*, vol. 36, pp. 64-70, 1997.

[8] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[9] J.W. Fickett, "Recognition of protein coding regions in DNA sequences," *Nucleic Acids Res.*, vol. 10, pp. 5303- 5318, 1982.

[10] J.W. Fickett and C.S. Tung, "Assessment of protein coding measures," *Nucleic Acids Res.*, vol. 20, pp. 6441-6450, 1992.

[11] H. Herzel and I. Große, "Measuring correlations in symbol sequences," *Phys. A*, vol. 216, pp. 518-542, 1995.

[12] H. Herzel, O. Weiss, and E.N. Trifonov, "10-11 bp periodicities in complete genomes reflect protein structure and protein folding," *Bioinformatics*, vol. 15, pp. 187-193, 1999.

[13] I. Ioshikhes, A. Bolshoy, K. Derenshteyn, M. Borodovsky, and E.N. Trifonov, "Nucleosomal DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences," *J. Mol. Biol.*, vol. 262, pp. 129-139, 1996.

[14] I. Ioshikhes, E.N. Trifonov, and M.Q. Zhang, "Periodical distribution of transcription factor sites in promoter regions and connection with chromatin structure," *Proc. Nat. Acad. Sci. USA*, vol. 96, pp. 2891-2895, 1999.

[15] W. Li, T.G. Marr, and K. Kaneko, "Understanding long-range correlations in DNA sequences," *Phys. D.*, vol. 75, pp. 392-416, 1994.

[16] S.K. Moore, "Understanding the human genome," *IEEE Spectr.*, vol. 37, pp. 33-35, Nov. 2000.

[17] S.K. Moore, "Making chips to probe genes," *IEEE Spectr.*, vol. 38, Mar. 2001.

[18] M.S. Pereira, L. Andrade, S. El Difrawy, B.L. Karger, and E.S. Manolakos, "Statistical learning formulation of the DNA base-calling problem and its solution in a Bayesian EM framework," *Discrete Appl. Math. J.*, (special issue on computational molecular biology), vol. 104, pp. 229-258, 2000.

[19] M.G. Reese, N.L. Hartzell, U. Harris, J.F. Ohler, J.F. Abril, and S.E. Lewis, "Genome annotation assessment in Drosophila melanogaster," *Genome Res.* vol. 10, pp. 483-501, 2000.

[20] J. Setubal and J. Meidanis, *Introduction to Computational Molecular Biology*. Boston, MA: PWS, 1997.

[21] J.C.W. Shepherd, "Method to determine the reading frame of a protein from the Purine/Pyrimidine genome sequence and its possible evolutionary justification," *Proc. Nat. Acad. Sci. USA*, vol. 78, pp. 1596-1600, 1981.

[22] B.D. Silverman and R. Linsker, "A measure of DNA periodicity," *J. Theor. Biol.*, vol. 118, pp. 295-300, 1986.

[23] A. Stein and M. Bina, "A signal encoded in vertebrate DNA that influences nucleosome positioning and alignment," *Nucleic Acids Res.*, vol. 27, pp. 848-853, 1999.

[24] J. Stock, "The bacterial nanobrain," in *Proc. NESCI Int. Conf. Complex Systems*, Nashua, NH, 2000, pp. 112-113.

[25] S. Tiwari, S. Ramachandran, A. Bhattacharya., S. Bhattacharya, and R. Ramaswamy, "Prediction of probable genes by Fourier analysis of genomic sequences," *CABIOS*, vol. 113, pp. 263-270, 1997.

[26] E.N. Trifonov, "3-, 10.5-, 200- and 400-base periodicities in genome sequences," *Phys. A*, vol. 249, pp. 511-516, 1998.

[27] R. Voss, "Evolution of long-range fractal correlations and 1/f noise in DNA base sequences," *Phys. Rev. Lett.*, vol. 68, pp. 3805-3808, 1992.

[28] J. Widom, "Short-range order in two eukaryotic genomes: relation to chromosome structure," *J. Mol. Biol.*, vol. 259, pp. 579-588, 1996.