Wavelet Analysis in Current Cancer Genome Research: A Survey

Tao Meng, Ahmed T. Soliman, Mei-Ling Shyu, Yimin Yang, Shu-Ching Chen, S.S. Iyengar, John S. Yordy, and Puneeth Iyengar

Abstract—With the rapid development of next generation sequencing technology, the amount of biological sequence data of the cancer genome increases exponentially, which calls for efficient and effective algorithms that may identify patterns hidden underneath the raw data that may distinguish cancer Achilles' heels. From a signal processing point of view, biological units of information, including DNA and protein sequences, have been viewed as one-dimensional signals. Therefore, researchers have been applying signal processing techniques to mine the potentially significant patterns within these sequences. More specifically, in recent years, wavelet transforms have become an important mathematical analysis tool, with a wide and ever increasing range of applications. The versatility of wavelet analytic techniques has forged new interdisciplinary bounds by offering common solutions to apparently diverse problems and providing a new unifying perspective on problems of cancer genome research. In this paper, we provide a survey of how wavelet analysis has been applied to cancer bioinformatics questions. Specifically, we discuss several approaches of representing the biological sequence data numerically and methods of using wavelet analysis on the numerical sequences.

Index Terms—Cancer genome, wavelet analysis, driver mutation, passenger mutation

1 Introduction

Cancer represents one of the greatest medical causes of mortality. It is responsible for one in eight deaths worldwide. Critical strides in developing systemic and local therapies for cancer have been made utilizing the increasing knowledge of the human genome and the relevant genetic changes found in tumors. A thorough understanding of cancer genome is a requirement for better treatment but poses a fundamental challenge due to the depth and sheer volume of data collected that must be evaluated and interpreted.

The early work which identified the role of the genome in the development of cancer dates back to the late 19th and early 20th century. David von Hansemann and Theodor Boveri examined dividing cancer cells under a microscope and observed the presence of strange chromosomal aberrations [1]. These findings suggested that cancers could be related to abnormalities in chromosomes, only found to be the relevant hereditary material half a century later. Following the discovery of DNA as the molecular substrate of inheritance, significant research has ensued to understand the mechanisms of cancer on a molecular level and to show that specific and recurrent genomic abnormalities are associated with cancers. For example, as early as 1981, Reddy et al. [2] found that the single base $G > T$ substitution of the $HRAS$ gene leads to the activation of that specific oncogene function in T24 human bladder carcinoma cells.

Currently, a generalizable concept of cancer states that malignancies result from accumulated mutations in genes that increase the "fitness" of a transformed cell over the cells surrounding it. The transformed cells sometimes acquire a set of sufficiently advantageous mutations that allow for unlimited proliferation and these cells, thus, become transformed, leading to malignancy. In addition, some cancer cells acquire the capability to spread to distant sites, presumably through the development of mutations, leading to metastases and increased patient mortality.

Mutations often occur in genes encoding proteins, the natural building blocks of all the components of the human body. Genes are determined by four subunits of DNA that are oriented in unique sequences, as are the resulting proteins. Current efforts to understand how mutations in DNA lead to the development of cancers have been partly limited by the general inability to sift through the vast quantities of data generated by cancer genome sequencing projects and the studies of individual investigators. As a consequence, there is a need for tools to parse through this large sum of data to present relevant gene changes that may be critical for either understanding how cancers develop or/and determining how they could ultimately be treated. From signal processing point of view, biological sequences, consisting of DNA and protein encoded data, could be viewed as one-dimensional signals. As a result, signal processing approaches have been applied to perform analysis on these types of data. The characteristics of most real-world signals are that they vary in both time and frequency domains. The Fourier transform (FT) is one way

- *T. Meng, A.T. Soliman, and M.-L. Shyu are with the Department of Electrical and Computer Engineering, University of Miami, MEB 406, 1251 Memorial Drive, Coral Gables, FL 33146. E-mail: {t.meng, a.soliman}@umiami.edu, shyu@miami.edu.*
- *Y. Yang, S.-C. Chen, and S.S. Iyengar are with the School of Computing and Information Sciences, Florida International University, 11200 SW 8th Street, ECS 354, Miami, FL 33199. E-mail: {yyang010, chens, iyengar}@cs.fiu.edu.*
- *J.S. Yordy and P. Iyengar are with the Department of Radiation Oncology, Harold C. Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical Center, Dallas, TX 75235. E-mail: {John.Yordy, Puneeth.Iyengar}@utsouthwestern.edu.*

Manuscript received 21 May 2013; revised 9 Sept. 2013; accepted 18 Oct. 2013; published online 28 Oct. 2013.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2013-05-0149.
Digital Object Identifier no. 10.1109/TCBB.2013.134.

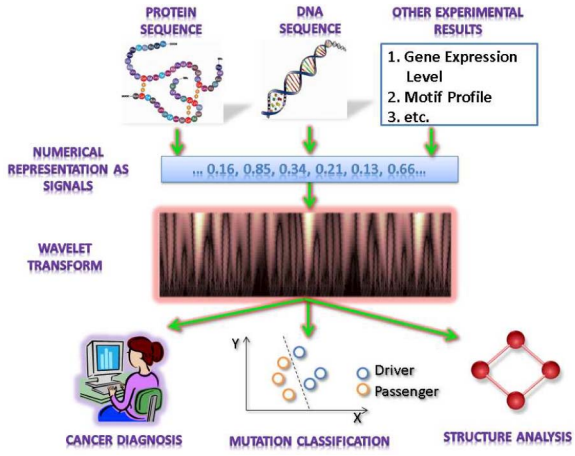


Fig. 1. A sample procedure of applying wavelet transform in cancer genome analysis.

to find the frequency content and measure signal composition in frequency. However, the classic Fourier transform does not give access to the signals' spectral variations during this time interval. In other words, the time and frequency information cannot be seen at the same time, and thus, a time-frequency representation of the signal is needed. With the help of better signal processing techniques and methods to represent genomic data, there may be ways to identify the critical changes within cancer genomes that contribute to progression, therapeutic resistance, desire for metastases, and so on.

Wavelet analysis, unlike traditional FT, is able to decompose time series into time-frequency space and has, thus, been getting more attention as a potential tool to study cancer genomic data. In this paper, we thoroughly survey the existing work and efforts that apply wavelet analysis in cancer genome bioinformatics. Rather than delving into details immediately, we first present an overview of a paradigm of applying wavelet transform techniques in biological sequences analysis relevant to cancer in Fig. 1 to provide the readers with a broad overview. Generally speaking, there are three main steps in this framework. First, to apply wavelet analysis, the original data need to be converted to a one dimensional (1D) signal. This step is critical to ensure the success of the analysis for the protein and DNA sequence data because a proper representation of the biological sequences retains their important characteristics. In contrast, some biological assays generate numerical values naturally, such as the DNA expression values in microarray analysis. In this case, the numerical representation step becomes trivial. Second, different wavelet transform techniques are applied to converted signals so a set of wavelet coefficients are obtained. An important research question in this step is: Which wavelet transform is the best to use? This is an open question and requires a case-by-case analysis. Third, the coefficients gained from the wavelet analysis step serve as features that are utilized further in many applications, among which are cancer driver mutation classification, cancer related protein structure analysis, and cancer diagnosis.

Following this guideline, we give a detailed review of the state-of-the-art techniques in the following sections. Since

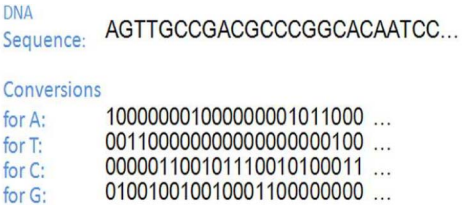


Fig. 2. Voss mapping.

the numerical representation is an important and nontrivial task for DNA and protein sequence analysis, different approaches are reviewed in Section 2. In Section 3, current state-of-the-art wavelet transform techniques are introduced to provide background to readers. In addition, applications utilizing wavelet techniques in general biological sequence analysis are reviewed to give readers further background and context and a more intuitive understanding. Based on Section 3, in Section 4 we summarize current progress in applying wavelet analysis to an important research topic in the bioinformatics domain, namely cancer genome analysis. In Section 5, we demonstrate an initial study applying wavelet analysis in distinguishing mutations that are at the heart of driving cancer development and progression, also known as driver mutations, and segregate these driver mutations from mutations that may be secondary to the actual tumor transformation process, i.e., passenger mutations. Experimental results are given and some observed insights are discussed. Section 6 concludes this review and proposes some future directions that deserve further exploration.

2 Numerical Representation of Biological Sequences

To process biological sequences as signals, the biological sequences need to be encoded in a suitable format that can be used by data analysis and data mining tools. This is usually achieved by assigning a numeral to each symbol that forms the biological sequence. There are two fundamental kinds of biological sequences relevant to cancer genomic/proteomic evaluation, namely DNA nucleotide sequences and protein amino acid sequences.

2.1 Numerical Representation of the DNA Sequences

The DNA sequences consist of four nucleotides—A, T, C, and G. There are various approaches to represent the DNA sequences as numerical sequences, which are introduced as follows:

2.1.1 Voss Mapping and Z-Curve

Voss mapping [3] is the most widely used approach for converting the DNA sequence to a sequence of numerals. It represents one DNA sequence using four binary indicator sequences for each nucleotide (A, T, C, and G). For example, Fig. 2 shows the DNA sequence segment of human Homo Sapiens Hexosaminidase A (HEXA) gene. One original sequence is converted to four binary sequences corresponding to four nucleotides, where "1" indicates the nucleotide appears in the sequence and "0" otherwise. In other words, each nucleotide is represented by a four-dimensional vector

of three "0"s and one "1." The advantages of this approach are its simplicity and its efficiency in spectral analysis of DNA sequences. The problem with this representation is that it does not capture the relationship between the four sequences. This representation has been used in predicting the coding regions (exons) in genes. Genes in eukaryotic cells have two subregions, exons and introns. The exons contain DNA sequences that will be transcribed and translated to protein sequences. Exons exhibit a period-3 property because of the codon structure involved in the translation of base sequences into amino acids. Based on this observation, the Fourier transform on the Voss mapping sequences was used to efficiently classify exons identified by a peak at frequency $1/3$ and introns identified by no peaks [4].

Abo-Zahhad et al. [5] illustrated the efficiency of applying short time discrete fourier transform (STDFT) and Voss Mapping to identify coding regions of the gene F56F11.5 of C. elegans. An extended version based on Voss Mapping is the Z-curve method [6]. The Z-curve is a three-dimensional curve that uniquely represents any given DNA sequence. The Z-curve is constructed from a set of $3d$ nodes, P_i. The number of nodes equals the size of a DNA sequence. For a DNA sequence of length T, $i = 1, \ldots, T$ and $P_i = (x_i, y_i, z_i)$, where

$$x_i = (A_i + G_i) - (C_i + T_i); \tag{1}$$

$$y_i = (A_i + C_i) - (G_i + T_i); \tag{2}$$

$$z_i = (A_i + T_i) - (C_i + G_i). \tag{3}$$

Here, A_i, G_i, C_i, and T_i are the cumulative occurrences of A, G, C, and T from the start of the sequence to the ith base. It is worth noting that the DNA sequence can be reconstructed from the representative Z-curve. It is one of the tools used to visualize genomes [7] and it is also used for gene identification [8].

2.1.2 Tetrahedron

The tetrahedron representation [9] reduces the number of indicator sequences from four to three in a manner symmetric to all four components. In this method, each of the four DNA nucleotides is assigned to a vertex of a regular tetrahedron in space. Each DNA nucleotide can be represented by a three-dimensional (R, G, B) vector as follows:

$$\begin{aligned} &A = (0,0,1); C = \left(-\frac{\sqrt{2}}{3}, \frac{\sqrt{6}}{3}, -\frac{1}{3}\right); \\ &G = \left(-\frac{\sqrt{2}}{3}, -\frac{\sqrt{6}}{3}, -\frac{1}{3}\right); T = \left(\frac{2\sqrt{2}}{3}, 0, -\frac{1}{3}\right). \end{aligned} \tag{4}$$

It is noticed that the representation strategy is similar to the quaternion representation used in [10] but is from a different perspective.

2.1.3 Complex Number Representation

One disadvantage of using the Voss mapping is that the relative weight of absence is not detected. Therefore, Cattani [11] represented the DNA sequence as the complex number. Assume the numbers a, t, c, and g are assigned to the nucleotides A, T, C, and G, respectively. The complex conjugate pairs $t = a^*$ and $g = c^*$ are chosen to represent the pairing structures of A and T, C, and G, respectively. One of the examples is shown in

$$A = 1 + j;\ T = 1 - j;\ C = -1 - j;\ G = -1 + j. \tag{5}$$

In this case, all palindromes yield conjugate and symmetric numerical sequences that have interesting mathematical properties, including the generalized linear phase. Abo-Zahhad et al. [5] demonstrated that the pairs of bases A-T and G-C are expressed by the fact that their representations are complex conjugates, while purines and pyrimidines have equal imaginary parts and real parts of opposite signs by using a slightly different complex representation as shown in (6) resulting in the expression of the two complementary strands of a DNA molecule by digital signals with the sum of zero:

$$A = -1 + j;\ T = 1 - j;\ C = -1 - j;\ G = 1 + j. \tag{6}$$

Bergen and Antoniou [12] proposed a method based on a complex representation, parametric windows function, and STDFT to maximize SNR to identify the coding regions for the gene F56F11.4 (as given in (7)). Anastassiou [13] used the scheme in (8), and the sequence was then represented as the random walk on the DNA sequence:

$$\begin{aligned} &A = 0.10 + 0.12j;\ T = -0.30 - 0.20j; \\ &C = 0;\ G = 0.45 - 0.19j, \end{aligned} \tag{7}$$

$$A = 1;\ T = j;\ C = -j;\ G = -1. \tag{8}$$

2.1.4 Integer Representation

In [14], the DNA nucleotides were mapped to numerals $\{0, 1, 2, 3\}$ using the scheme that $T = 0$, $C = 1$, $A = 2$, and $G = 3$. However, this method implies that a structure on the nucleotides, such as purine (A, G) $>$ pyrimidine (C, T), will introduce bias in the DNA sequence analysis. Also, Zhou and Yan [15] used an integer representation in their proposed approach to analyze short tandem repeats in the DNA sequences.

2.1.5 Physicochemical Property-Based Representation

This kind of representation takes into account of the biochemical properties of the DNA biomolecules. Such a representation carries the characters of the chemicals themselves and is relatively robust and biologically meaningful. In [16], the electron-ion interaction potential (EIIP) indicator was utilized to map the four DNA nucleotides. EIIP was defined as the average energy of delocalized electrons of the nucleotide. By assigning the EIIP values to the nucleotides, a numerical sequence was obtained to represent the distribution of the free electrons' energies along the DNA sequence. This approach has been successfully used to identify coding regions.

In summary, there are many numerical representation techniques proposed to map DNA sequences. Each technique has advantages and disadvantages. Voss Mapping is the most widely used approach. It indicates the frequencies of the bases but it does not capture any mathematical relation between them. It is efficient for spectral analysis of DNA sequences and identification of coding and noncoding regions in DNA sequences. Z-Curve offers numerical and

TABLE 1
The Complex Representation of 20 Amino Acids

Amino Acid Name	Symbol	Complex Number Repr.
Alanine	A	0.61 + 88.3i
Arginine	R	0.60 + 181.2i
Asparagine	N	0.06 + 125.1i
Aspartic	D	0.46 + 110.8i
Cysteine	C	1.07 + 112.4i
Glutamic	E	0.47 + 140.5i
Glutamine	Q	148.7i
Glycine	G	0.07 + 60.0i
Histidine	H	0.61 + 152.6i
Isoleucine	I	2.22 + 168.5i
Leucine	L	1.53 + 168.5i
Lysine	K	1.15 + 175.6i
Methionine	M	1.18 + 162.2i
Phenylalanine	F	2.02 + 189.0i
Proline	P	1.95 + 122.2i
Serine	S	0.05 + 88.7i
Theronine	T	0.05 + 118.2i
Triptophan	W	2.65 + 227.0i
Tyrosine	Y	1.88 + 193.0i
Valine	V	1.32 + 141.4i

graphical representations but it is not suitable for long sequences. Tetrahedron representation reduces the number of indicator sequences from four to three and results in DNA color spectogram that can be used to locate repeating DNA sections visually and to identify CG rich regions (CpG islands). Complex representation projects the tetrahedron components on two planes to reduce the dimensionality to two and it can capture the relations between the four bases and the mathematical properties. Therefore, it is suitable for the detection of exon regions and gene prediction. Integer representation is simple and computationally efficient. It may be capable of mapping some of the nucleotide bases' relations into mathematical properties, on the expense of introducing additional mathematical properties that are not present in the DNA sequence. It is not as efficient as other techniques when applied to gene prediction. Physicochemical property-based representation is a robust and biologically meaningful method to represent the DNA sequence and has been successfully used in detecting coding regions. Physicochemical property-based representation is a robust and biologically meaningful method to represent the DNA sequence and has been successfully used in detecting coding regions.

2.2 Numerical Representation of the Protein Sequences

The protein sequence is more complicated than the DNA sequence since there are 20 amino acids. The left two columns of Table 1 show the amino acids with their symbols.

2.2.1 Orthonormal Representation

The most frequently used encoding strategy is the orthonormal strategy [17]. In this strategy, the 20 amino acids are represented by the 20 orthogonal unit vectors in a 20-dimensional space. Specifically, assume each letter $l_i (1 \leq i \leq 20)$ of the amino acid alphabet $A = l_1, l_2, \ldots, l_{20} = A, R, \ldots, V$ is replaced by an orthonormal vector as shown in

$$l_i = (\delta_{i,1}, \delta_{i,2}, \ldots, \delta_{i,j}, \ldots \delta_{i,20}), \tag{9}$$

where $i, j \in 1, \ldots, 20$ and $\delta_{i,j}$ is the Kronecker delta symbol. This representation is relatively simple but it has several drawbacks. First, the dimension of the feature space is 20 times the sequence length. Second, the distance between two amino acids is always the same and information regarding the similarity of two amino acids is lost. To overcome this disadvantage, researchers group similar amino acids together. In [18], a solution of grouping the amino acids into six groups was proposed (as shown in (10)). Correspondingly, the encoding strategy is changed as in (11), where N is the number of groups. This model is still relatively coarse because it treats all the amino acids in the same group in the same way:

$$G = (\{H, R, K\}, \{D, E, N, Q\}, \{C\}, \{S, T, P, A, G\}, \{M, I, L, V\}, \{F, Y, W\}) \tag{10}$$

$$l \in G_i = (\delta_{i,1}, \delta_{i,2}, \ldots, \delta_{i,N}). \tag{11}$$

2.2.2 Physicochemical Property Representation

This type of approaches accounts for most of the representation methods. Many previous studies utilized the physicochemical property of the amino acid to represent each amino acid numerically. The amino acid indices database [19] contains 544 different kinds of properties of each amino acid. These different indices cover a broad range of amino acid properties including hydrophobicity, residue volume, steric parameter, and so on. In addition, to apply machine learning techniques, some researchers utilized the pseudo amino acid composition (PAAC) [20], [21] as the features to represent each protein sequence. To compute the PAAC features, each amino acid was represented using hydrophobicity, hydrophilicity, mass, PK1, PK2, and PI. Based on these representations, the high-level sequence information was coded as the feature. The advantage of using the physico-chemical property to represent the amino acid is that the number has some physical meanings. The authors of [22] used the hydropathy, flexibility, electronic charge concentration, isotropic surface area, and solvent accessibility area to represent each amino acid and used the wavelet analysis to perform the decoy discrimination. In [23], an EIIP scheme was proposed to represent each amino acid as electron-ion interaction potential. This representation has the similar idea to the one introduced in Section 2.1.5 but for amino acids. However, the property used often depends on the specific application and requires a case-by-case analysis.

In addition to the single-value representation, there are some studies that represent each amino acid as vectors or complex numbers. Swanson [24] represented each amino acid as a two-dimensional vector. The two dimensions of the plane are the size and hydrophobicity of amino acids. This showed that the vector representation could bring four advantages in protein sequence analysis. First, the "similarity" can be explicitly modeled. Second, some new mathematical properties can be attributed to the sequence. For example, the numerical value for the conservativeness of a site could be defined. Third, the protein sequence can be pictured, which helps visualize the sequence information. Fourth, it helps the detection of homology of different proteins, and demonstrated additional applications of the vector representation of amino acid.

Based on similar ideas, some researchers adopted similar strategies to represent each amino acid using the complex number representation with the real part and imaginary part representing different properties of the amino acid. For example, a complex number representation approach was proposed in [25], where the hydrophobicity is the real part and the residue volume is the imaginary part. Such a representation is shown in Table 1.

2.2.3 Two-Dimensional and Three-Dimensional Representations of Amino Acids

In recent years, the geometric representation of the protein sequence has become increasingly popular in sequence comparison. The authors of [26] represented the proteins based on the concepts of virtual genetic code and a four-color map, which help the researchers visualize the similarity/dissimilarity between proteins. In their study, they also developed a novel protein descriptor, which is a 10-dimensional vector derived by the structure matrix associated with the map. There are other approaches that represent each amino acid using an N-dimensional vector and represent the overall sequence as the curve connecting the vertices, which are the summation of the amino acid vectors, and so on. In [27], each amino acid was represented using a three-dimensional vector. The three dimensions correspond to three physicochemical properties, namely Hydrophilicity, pK1, and pI. Specifically, given a protein sequence $S = s_1 s_2 \dots s_n$, the 3D space point $P_i(x_i, y_i, z_i)$, $1 \leq i \leq n$ is computed using

$$x_i = \sum_{k=1}^{i} S_k^1;\ y_i = \sum_{k=1}^{i} S_k^2;\ z_i = \sum_{k=1}^{i} S_k^3. \tag{12}$$

Here, $S_k^j (j = 1, 2, 3)$ represents the jth component of vector corresponding to s_k. p_0 is set to $(0, 0, 0)$. When i runs from 1 to n, P_i becomes the vertices and the curve connecting the vertices forms the protein curve in 3D space.

The authors of [28] proposed a scheme to represent all the amino acids as the vertices of a dodecahedron. In this way, a protein sequence is represented as a 3D curve. The authors showed that such a representation could help compare different sequences. The problem of this representation is that the protein sequence information could not be reflected at the node.

2.2.4 Subalphabets-Based Representation

The property of the amino acid is just one factor that determines the structure of the proteins [29], [30]. The context information of the amino acid is also important. Accordingly, instead of representing each amino acid individually, there is research trying to represent a pair of amino acids or a triplet of amino acids together. The authors of [31] evaluated the subalphabets by searching directly for the sequence codings that improve protein secondary-structure prediction. They discovered that protein alphabets composed of 13 to 19 groups could increase the predictability of the secondary structure from sequences.

In summary, a proper numerical representation of protein sequences serves as the foundation for the following wavelet analysis on sequences. Each representation strategy has its pros and cons, so the choice of the proper strategy is at the discretion of the researcher and needs a case-by-case analysis. Generally speaking, the orthonormal representation is more suitable for local sequence alignment or comparison as it carries no extra amino acid properties. The physiochemical property-based representation is applicable under the scenario that the global comparison of two sequences, or the functional analysis, needs to be carried out, since the properties of the amino acid with the sequence together determine the properties of the overall protein. The geometric representation, which becomes increasingly popular, incorporates the advantages from the aforementioned two approaches. However, it requires some sophisticated geometric expertise to carry out the analysis and also increases computational complexity. Finally, subalphabets that include the context information of one amino acid play an important role in protein structure analysis.

3 Review of Wavelet Analysis and Its Applications in Biological Sequence Analysis

This section provides a brief overview of wavelet analysis including background, evolution, and possible applications. Finally, it concludes with a summary of the wavelet applications in the analysis of gene and protein sequences.

3.1 Review of Wavelet Analysis

In this section, we present an overview of the evolution of wavelet analysis techniques, introduce the transition from Fourier transformation to wavelet analysis, give an example of applying wavelet analysis on biological sequences, and finally provide a list of commonly used wavelets families in biological sequence analysis.

3.1.1 Background

The origin of the wavelets theory can be traced back to the harmonic analysis developed by a French mathematician, Jean Baptiste Joseph Fourier (1768-1830) [32]. He was the first to develop a method of expressing any periodic function in terms of a weighted sum of cosine and sine functions, i.e., Fourier Trigonometric series. In 1909, Alfred Haar developed Haar Wavelets family [33]. It is the simplest wavelets set and can be used to analyze a given signal in terms of functions that are more finite in time than the harmonic functions used in the Fourier analysis. The Haar Wavelets family was later proven to be more accurate in modeling functions because of its scaling property. In 1980s, Jean Morlet replaced the Gabor window used in STFT [34] by the stretched and compressed versions of unique oscillating windows, which get more reliable and accurate analyses and are known as the Morlet Wavelets. In 1989, the idea of multiresolution, which is the base theory of versatile wavelets families, was proposed [35]. Based on the multiresolution concept, Daubechies [36] created the well-known and frequently used Daubechies wavelets family.

As can be inferred from the evaluation of wavelet analysis, it originated from FT. The characteristics of most real-world signals vary in both time and frequency domains. They are nonstationary signals.[1] FT is one way to find such

1. A stationary signal is a signal where there is no change in the properties of signal, while a nonstationary signal is a signal where there is change in the properties of signal.

TABLE 2
Comparison of Wavelet Analysis and FT

Properties	Fourier transform (FT)	Wavelet analysis
Stationary signal	Yes	Yes
Non-stationary signal	No	Yes
Time domain	No	Yes
Frequency domain	Yes	Yes
Scale	Yes	Yes
Shift	No	Yes

frequency content and measure the signal composition in frequency. The FT is calculated using (13), where F is the frequency in Hertz and Ωt is the phase in radians:

$$FT\{x(t)\} = X(\Omega) = \int_{-\infty}^{\infty} x(t)e^{-j\Omega t}dt,\ \Omega = 2\pi F. \tag{13}$$

The FT defines the global representation of the frequency content of a signal over a total period of time. However, it does not give access to the signal's spectral variations during this interval of time. In other words, the time and frequency information cannot be seen at the same time, and thus, a time-frequency representation of the signal is needed.

To circumvent this localization problem, Gabor [37] proposed the STFT to analyze only a small section of the signal at a time by using a technique called windowing the signal. This obtains the specific contents of each of the analyzed sections separately. The segment of signals in each section is assumed stationary. Let $g(t)$ be the sliding window of a fixed size. STFT is defined in (14), where $g(t-b)e^{-j\Omega t} = \psi^*_{\Omega,b}(t)$ is the complex conjugate of $\psi_{\Omega,b}(t)$:

$$\begin{aligned} STFT_{g(\Omega,b)}\{x(t)\} &= \int_{-\infty}^{\infty} x(t)g(t-b)e^{-j\Omega t}dt \\ &= X_g(\Omega, b). \end{aligned} \tag{14}$$

However, STFT has its own limitations due to the fixed window. That is, a narrow window results in a poor frequency resolution, whereas a wide window leads to poor time resolution. In addition, one cannot determine the time intervals where a certain frequency exists. Therefore, the wavelet transform was proposed as an alternative approach to STFT to overcome the resolution problem. The definition of continuous wavelet transform is as follows:

$$\begin{aligned} CWT^{\psi}_{x}(a,b) &= X_{\psi}(a,b) \\ &= \frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} x(t)\left[\psi^*\left(\frac{t-b}{a}\right)\right]dt \\ &= \langle x(t), \psi^*_{a,b}(t)\rangle, \end{aligned} \tag{15}$$

where a and b are the scaling and translation parameters, respectively, and $\psi^*_{a,b}(t) = \frac{1}{\sqrt{a}}\psi^*(\frac{t-b}{a})$ is the mother wavelet (base function), a prototype for generating the other window functions. All the windows are the dilated, compressed, and shifted versions of the mother wavelet. There are various wavelet basis functions, which will be introduced later.

In summary, wavelet analysis techniques outperform the traditional FT in the following perspectives [38]:

1. wavelets are suitable for analysis on both stationary and nonstationary signals, while FT is less useful in analyzing nonstationary signals;
2. wavelets are well localized in both time and frequency domains, whereas the standard FT is only localized in frequency domain;
3. the base functions of wavelets can both be scaled and shifted, while the FT can only be scaled; and
4. wavelets have solid mathematics foundation and a wider range of applications than FT, for example, nonlinear regression and compression.

A brief summary of the comparison is shown in Table 2.

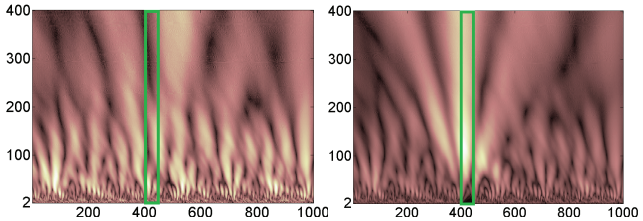


Fig. 3. Visualization of wavelet transform coefficients on a segment of homo sapiens gene p53 (TP53). The x-axis and y-axis represent the nucleotides indices and the scale numbers, respectively. The green rectangles in the figures indicate the mutation spots. The green bar highlights a visualized regional difference in wavelet coefficients based on a mutation in the DNA sequence (see text for details).

3.1.2 An Example of Applying Wavelet Analysis on Biological Sequence

Over the past few decades, tremendous research effort has been dedicated to decipher the entire human genome sequence, which has become one of the most exciting challenges facing scientists today [39]. Taking DNA sequences as an example, an organism's entire genome is usually represented by a large size of DNA sequences. It is desirable to transform this long sequence into a more manageable data set. To be more specific, a DNA sequence can be regarded as a discrete signal composed of a finite number of nucleotides. This observation suggests the potential to use standard digital signal processing techniques to analyze the DNA sequence as a discrete-time sequence. However, a prerequisite step is needed for interpreting the original symbolic signals to proper numerical representations (as discussed in Section 2).

Compared with traditional Fourier-based techniques for signal processing, wavelet-based techniques are more appealing due to their attractive properties, such as time-frequency domain representation, local feature identification, and multi-resolution scalability as introduced in Section 3.1.1.

As an initial example of the effect of wavelet transforms on biological sequences, we apply the Daubechies wavelets function to visualize mutations of a segment of p53 (TP53) gene. Figs. 3a and 3b visualize the wavelets coefficients before and after the mutations of a segment in the gene. Specifically, the following steps were carried out:

1. Extract the first 1,000 nucleotides from the original DNA sequence (TP53).
2. Map the sequence segment to complex numbers using (5).
3. Apply Daubechies transform to the complex numbers in scales 2 to 400 at a step length of 2 to obtain a coefficients matrix (visualized in Fig. 3a).

TABLE 3
Summary of Wavelets Used in Biological Sequence Analysis

Wavelet Name	Wavelet Function	Description and Properties	Applications in Biological Sequence Analysis
Haar wavelet	Mother wavelet: $\psi(t)=\begin{cases}1 & 0\le t<\frac{1}{2}\\ -1 & \frac{1}{2}\le t<1\\ 0 & \text{otherwise}\end{cases}$ Scaling function: $\phi(t)=\begin{cases}1 & 0\le t<1\\ 0 & \text{otherwise}\end{cases}$	The oldest and simplest wavelets Also called $D2$ It can approximate any continuous real function by linear combination of shifting and scaling	Feature extraction for splice sites identification [40] Speed DNA sequence search and gene sequence analysis [41] [42] Structure analysis in conserved protein motif detection [43]
Daubechies wavelets	They can be defined by the wavelets coefficients	A family of orthogonal wavelets defining discrete wavelet transform It cannot be written in closed-form. It is characterized by the number of vanishing moments (order) The number of coefficients= $2*order$	Noise reduction for gene identification [44] [45] CpG islands identification [46] Feature extraction and noise reduction for exons and introns prediction [45] Structure analysis in conserved protein motif detection [43]
Meyer wavelet	$\psi(w)=\begin{cases}\frac{1}{\sqrt{2\pi}}\sin\left(\frac{\pi}{2}v\left(\frac{3\lvert w\rvert}{2\pi}-1\right)\right)e^{\frac{iw}{2}} \text{ if } \frac{2\pi}{3}<\lvert w\rvert<\frac{4\pi}{3},\\ \frac{1}{\sqrt{2\pi}}\cos\left(\frac{\pi}{2}v\left(\frac{3\lvert w\rvert}{4\pi}-1\right)\right)e^{\frac{iw}{2}} \text{ if } \frac{4\pi}{3}<\lvert w\rvert<\frac{8\pi}{3},\\ 0 \text{ otherwise}\end{cases}$ where $v(x)=\begin{cases}0 \text{ if } x<0\\ x \text{ if } 0<x<1\\ 1 \text{ if } x>1\end{cases}$	A family of orthogonal continuous wavelets defined in frequency domain It is indefinitely differentiable with infinite support Dmey wavelet is the discrete variant of Meyer wavelet	Noise reduction for exonic regions identification [47] [48]
Mexican_hat wavelet	$\psi(t)=\frac{2}{\sqrt{3\sigma}\pi^{\frac{1}{4}}}\left(1-\frac{t^2}{\sigma^2}\right)e^{\frac{-t^2}{2\sigma^2}}$	It is the negative normalized second derivative of a Gaussian filter function A special case of the family of continuous wavelets	Noise reduction for protein coding regions prediction [49] Noise reduction for repeating motifs detection [50]
Morlet wavelet	$\psi_\sigma(t)=c_\sigma\pi^{-\frac{1}{4}}e^{\frac{-t^2}{2}}\left(e^{i\sigma t}-K_\sigma\right)$ where $K_\sigma=e^{\frac{-\sigma^2}{2}}$ $c_\sigma=\left(1+e^{-\sigma^2}+2e^{-3\sigma^2}\right)^{-\frac{1}{2}}$	Also known as Gabor wavelet A family of continuous wavelets It is composed of complex exponential multiplied by a Gaussian envelope It allows trade-off between time and frequency resolution The frequency of Morlet wavelet is conventionally taken as $w_\psi\approx\sigma$	Identification of protein coding regions [51] [52] Analysis of Human DNA [53] Protein secondary structure prediction [54]
Shannon wavelet	Real Shannon wavelet: $\psi(t)=\frac{\sin(\frac{3\pi}{2}t)\cos(\frac{3\pi t}{2})}{\frac{\pi t}{2}}$ Complex Shannon wavelet: $\psi(t)=\frac{\sin(\pi t)}{\pi t}e^{-j\pi t}$	A family of continuous wavelets It is indefinitely differentiable with infinite support This family is obtained from the frequency B-Spline wavelets by setting to 1	Analysis of Human DNA [53]

4. Manually mutate 50 base pairs (bp) from the segmented sequence with nucleotides indices from 401 to 450.
5. Perform Daubechies transform again on the artificially mutated sequence segment to obtain a new coefficients matrix (visualized in Fig. 3b).

As illustrated in Fig. 3, the ability of wavelet transform to capture the variations in a DNA sequence due to mutation at different scales is visually obvious. This initial analysis sheds lights on the possibility of utilizing wavelet techniques for DNA and protein sequence analysis.

3.1.3 Wavelet Families for Biological Sequences

This section illustrates some of the wavelet families commonly used in biological sequence analysis. Wavelet families generally belong to one of the following types. Table 3 summarizes some of the wavelet families commonly used in biological sequence analysis applications:

- *Orthogonal wavelets with scaling finite impulse responses (FIR) filters*. These wavelets are defined through a low-pass scaling filter. Predefined families of such wavelets include: Haar, Daubechies, Coiflets, and Symlets.
- *Biorthogonal wavelets with scaling finite impulse responses filters*. These wavelets are defined through two scaling filters, for reconstruction and decomposition, respectively. The BiorSplines wavelet family is an example of a predefined family of this type.
- *Wavelets with scaling function*. These wavelets are defined using a wavelet function, the mother wavelet, and a scaling function, the father wavelet, in the time domain. The Meyer wavelet family is a predefined family of this type.
- *Wavelets without scaling filters and without scaling function*. These wavelets are defined through the definition of the wavelet function. The wavelet has a time-domain representation only. Predefined families of such wavelets include Morlet and Mexican_hat.

3.2 Wavelet Analysis Application in Gene Sequence Analysis

This section summarizes previous applications of wavelets in DNA analysis. Generally, wavelet approaches have been applied in gene finding and gene sequence analysis.

3.2.1 Gene Finding

A major goal of genomic research is to understand the functions of each individual gene and their interactions. The eukaryotic DNA strand is divided into genes and intergenic spaces. Genes are further divided into exons and introns. The exons carry the code for the synthesis of proteins, and are called the protein-coding regions. Nowadays, a very important task in genomic research is to find the locations of the genes in the genome and, in a deeper sense, the protein-coding regions in the DNA strand. Mena-Chalco et al. [51] utilized a modified Gabor wavelet transform of the DNA sequence to identify the protein coding regions on the DNA strand. Given the three base periodicity (TBP) patterns of the protein coding regions and the different scales of protein coding regions, a modified Gabor wavelet transform was utilized to decompose the DNA sequence and the threshold values were computed for the projection coefficients to make the decision. The Gabor wavelet was modified using the scale parameter to keep the complex exponential frequency constant while varying the Gaussian standard deviation. The resulting modified Gabor wavelet enables analyzing a signal in a specific frequency and multiple scales. The proposed approach consists of the following four steps:

1. mapping the DNA sequence to four binary sequences,
2. applying Modified Gabor Wavelet Transform (MGWT) to each binary sequence,
3. projecting the sequence spectra onto the position axis, and

4. thresholding the projection coefficients to identify the edges among coding regions.

A parallel implementation of the MGWT-based approach [51] on multicore systems was proposed in [52], which used the single program multiple data (SPMD) parallel paradigm.

In [44], Daubechies discrete wavelet transform was used for DNA signal denoising by setting the appropriate frequency component thresholds for approximate and details coefficients corresponding to the low- and high-scale frequency components. First, the DNA sequence was mapped to a DNA signal using $X(A) = 0.260, X(T) = 0.375, X(G) = 0.125$, and $X(C) = 0.370$. Then, a three level noise reduction technique based on Daubechies discrete wavelet transform of order three was applied to the DNA signal. In the first level, the DNA signal was decomposed using a high- and low-pass filters to approximate ($A1$) and detail ($D1$) coefficients and downsampled by 2. The first level approximate signal was decomposed and downsampled again using high- and low pass filters to approximate ($A2$) and detail ($D2$) coefficients. The same process was applied to the second level approximate signal ($D2$) resulting in $A3$ and $D3$. Finally, the DNA signal was reconstructed again by the mirror reconstruction filters using $D1$, $D2$, $D3$, and $A3$ after setting the appropriate frequency components thresholds. A similar technique for noise suppression was proposed by using the Dmey mother wavelet, also known as Discrete meyer wavelet, to decompose the DNA signal into detail and approximation signals, and then to filter out the detail signals [47], [48]. By removing the detail signals and considering only the approximation signal, the output power signal was smoothed and extra frequencies were removed. Decreasing the noise effect enhanced the accuracy of exonic region identification. In [40], an approach based on discrete wavelet transform and support vector machines (SVMs) was proposed to identify splice sites in the human genome. It employed one-dimensional and two-dimensional Haar wavelet transforms to transform the binary coded DNA sequence into wavelet coefficients. Wavelet coefficients form the SVM's input feature vector. The binary coded DNA sequence was generated by mapping A, G, C, and T to column vectors $[0001]$, $[0010]$, $[0100]$, and $[1000]$. DasGupta et al. [46] considered an approach based on wavelet analysis and Hidden Markov Tree (HMT) to identify CpG island, regions characterized by a higher concentration of C-G nucleotides than elsewhere, locations in the human genome. The numerical representation of the DNA sequence was subject to a multilevel wavelet decomposition, to generate a sequence of wavelet trees. A single HMT was used to model the wavelet trees sequence. Also, genetic algorithms and lifting algorithm were used to design adaptive wavelets matching the CpG islands. The results indicated that the performance based on the Daubechies wavelet is comparable to performance based on adaptive wavelets. This suggests that the CpG islands might be too heterogeneous to be characterized by a single wavelet and HMT was able to adjust its parameters to suite the wavelet under consideration.

Gupta et al. [45] proposed a SVM classifier to identify the exons and introns. The classifier was based on a novel wavelet variance coefficient feature vector, where features were extracted using the maximal overlap discrete wavelet transform (MODWT). The DNA sequence was represented using the Z-curve and then MODWT was applied to the information extracted from the Z-curve to generate the exons and introns features vector. Daubechies wavelet function and a maximum level of decomposition of 6 were used in MODWT. Despite the lack of comparison with other feature extraction techniques, the demonstrated results seemed promising. Deng et al. [49] introduced a method to predict protein coding regions in DNA sequences using Fourier and Wavelet transforms. A continuous wavelet transform using a Mexican hat wavelet function was used to eliminate the high-frequency noise in the Fourier spectrum representing the DNA sequence. The proposed method is not valid if the DNA sequences lack the three-base periodicity characteristics.

3.2.2 Gene Sequence Analysis

This area of research is the main impetus for using wavelets in DNA sequence analysis. For example, Jiang and Yan [55] used the Hilbert-Huang transform to study the properties of short genes which are genes whose exons sequence length is below 70 base pairs. In their paper, a wavelet subspace algorithm combined with the empirical mode decomposition (EMD) was proposed to create subdivided intrinsic mode functions (IMF) and a cross-correlation analysis was applied to remove pseudo spectral components. In [41] and [42], the symmetrical shapes in the wavelet coefficients of DNA sequences appear when the short wavelet transform mapped the numerical representation of the DNA sequence into the space of wavelet coefficients. The short wavelet transform was achieved by the subdivision of the DNA sequence into four-length segments and then the Haar wavelet transform was applied to each segment.

In [56], the nucleotide sequences were mapped onto a DNA walk to produce the fractal landscapes. This method determines the singularity spectrum of the considered DNA sequence, and provides a complete multifractal analysis. The wavelet transform is also used in characterization and detection of repeating motifs,[2] small conserved regions that usually carry specific structural or functional significance, in DNA sequences. Repeating motifs are thought to be the modular building blocks, allowing for an economic way of constructing complex proteins. Murray et al. [50] employed techniques based on the Mexican hat wavelet transform, a method of continuous wavelet transform, to classify the TIM barrels motif, propeller blades motif, coiled coils motif, and leucine-rich motifs. El-Zanaty et al. [57] applied the Haar Wavelet transform to the binary representations of the DNA sequence and claimed that their proposed method improves the search performance and reduces the storage requirements. In [53], six wavelet approaches including Haar, Ricker (also called the Mexican hat), Shannon, Hermitian hat, Shannon complex, and Morlet wavelets were compared for their performances on analyzing the human genome. It showed that applying the six different wavelets would yield different results, and indicated that the Shannon real wavelet is more promising in terms of discovering hidden patterns in the human DNA sequences.

2. In genetics, a sequence motif is a widespread nucleotide or amino-acid sequence pattern with a biological significance.

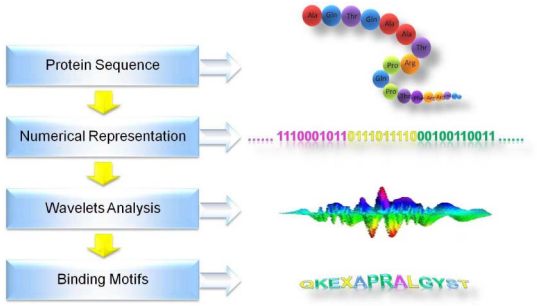


Fig. 4. The procedure of finding protein motifs using wavelet analysis.

Lin and Linton [58] used wavelet packet decomposition, which was represented by a subtree of the complete decomposition tree given by all possible dyadic decompositions of a signal with two filters fulfilling the power complementarity condition to analyze the sequence Colacium elongatum MI-11. Chandra and Rizvi [59] used Morlet and Haar wavelets to analyze the HIV-1 genome. Their results demonstrated the significance of using Morlet wavelets for lower scales analysis and Haar wavelets for higher scale analysis. Similar work in this direction includes [60], [61], [62], [63], [64], [65].

3.3 Wavelet Analysis Application in Protein Sequence Analysis

Wavelet analysis has also been applied in protein sequence analysis including motif searching, sequence comparison, and so on. In this section, we summarize the applications of wavelet in protein sequence analysis and give some examples in each application.

3.3.1 Protein Motif Searching

Protein motifs are small conserved regions within the protein sequences carrying specific structural or functional significance. The detection of a common protein motif serves as an indicator for the function of the protein and is a good feature for protein classification. Since the wavelet analysis can not only capture the global information as the spectral analysis in discrete Fourier transform but is also able to capture the location features, it is a useful approach to detect local repeating motifs on the sequences. An exemplar study was carried out in [66]. In this study, the researchers utilized the hydrophobicity and relative accessible surface area values, which belong to the physicochemical property representation introduced in Section 2.2.2, to represent each amino acid and converted the protein sequences into 1D signals. The reason for choosing these two properties of amino acids was that they give a strong indication of the protein's overall geometry as shown in previous studies [67]. Next, the Mexican hat wavelet analysis was applied to the series because it is good at highlighting periodic structures. The authors then plotted the coefficients using scalogram. They successfully identified six different protein motifs: seven bladed β propellor, domain repeat and eightfold $\beta\alpha$ motif repeat, $\beta\alpha$ Leucine-rich repeat, $\alpha\alpha$ Heat repeats, four repeating β sandwich domains, and coiled coil heptad repeat. In a recent study [43], the authors utilized another physicochemical property-based representation, where each amino acid was mapped using 10 bits with each bit representing the presence or the absence of a certain property like positive charge, polarity,

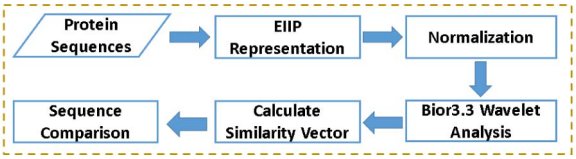


Fig. 5. The procedure of using wavelet analysis to perform sequence comparison.

and so on. The authors showed that Haar wavelet and Daubechies wavelet are suitable for protein sequence signal analysis based on physico-chemical properties. Therefore, they were used to analyze the converted amino acid string. The converted region was found by picking up the peak of the transformed curve. The authors successfully identified several conservative protein motifs on the sequences. The general framework of using wavelet analysis in protein motif search is shown in Fig. 4.

3.3.2 Sequence Comparison

Protein sequence comparison is one of the most important areas in bioinformatics research. The conventional BLAST-based approach focuses on the local pairwise amino acid match. However, two protein sequences with low sequential identity may show similarities in physiochemical properties and tertiary structure, which indicates a functional correlation between the two proteins. The conventional sequence-based comparison is challenged in identifying this kind of similarity. In this case, wavelet analysis becomes an useful alternative as it is able to capture the multiscale information that enables the comparison of protein sequences at different resolutions. An example framework is presented in [68]. Their proposed framework is shown in Fig. 5. First, the protein sequences are converted to numerical sequences using the EIIP representation introduced in Section 2.2.2. Then, the numerical series are normalized to zero mean and unit standard deviation and zero-padded to have an identical sequence length. Next, M level Bior3.3 biorthogonal wavelet analysis, which allows more flexibility compared with the orthogonal wavelet, is applied to each sequence to generate a set of coefficients at each level. Finally, a similarity vector is computed according to the coefficients and used to compare different sequences. Using this approach, the authors successfully identified the functionally correlated proteins even though they show little similarity at the sequence level. For example, the sperm wale myoglobin and lupine leghemoglobin only have 15 percent identical residues. However, they both contain a heme group and have similar secondary and tertiary structures. Their work indicated that the wavelet-based sequence comparison could discover information missed by the sequence alignment-based approach. A similar framework was proposed in [69].

3.3.3 Prediction of the Secondary Structure

The secondary structure of the proteins includes α-helix, β-sheets, and the short peptides connecting them. The prediction of this secondary structure is a classic research problem. Since the wavelet analysis is able to capture both the spectral and temporal information of a sequence, it can provide significant features for a sequence. These features

could be utilized in state-of-the-art machine learning frameworks to annotate secondary structures automatically. Based on this thinking, Qiu et al. [70] proposed a framework for classifying the protein sequence into four classes according to four secondary structures: α/β, $\alpha+\beta$, all-α, and all-β. Because hydrophobic property plays a crucial role in the process of forming secondary structures and folding into tertiary structures [54], the amino acid was represented using hydrophobicity, which is introduced in Section 2.2.2. Next, Morlet wavelet transform was applied to the numerical series because it is suitable for protein high-order structure analysis. Four kinds of features, which are the maximum, mean, minimum, and standard deviation of the wavelet coefficients, were extracted in each sub-band to form a feature vector, and SVM classifier was adopted to annotate four different classes. Other similar studies include [54] and [71].

In summary, wavelets techniques are able to extract both spectral and local information and perform multiscale analysis on DNA/protein sequences. Using the proper numerical representation strategy introduced in Section 2 according to different applications (such as hydrophobicity in the protein structural analysis), the wavelet-based framework could provide extra information, which is difficult to mine from the traditional alignment-based approaches.

4 Wavelet and Cancer Genome Research

The elucidation of the mechanisms of cancer development on the molecular level is one of the preeminent questions in molecular and cell biology and is an essential part of any plan to fight cancer. As the development and commercialization of massively parallel DNA sequencing lowers the cost per sequenced nucleotide by several orders of magnitude, a multitude of DNA sequences or protein sequences have been collected from tumor cells. This rapid growth of sequence data calls for more advanced engineering approaches to gain insight into the development of malignancies. Among a plurality of analysis approaches, wavelet analysis holds a place on this stage for its advantages in simultaneous localization in time and frequency domains and multiscale analysis. In this section, based on the description of the general discussion of wavelet analysis in biological sequence in the previous section, we focus on reviewing existing work applying wavelet analysis in cancer genome research. In Section 4.1, a brief summary of common understandings regarding the cancer genome is given. We then introduce the state-of-the-art work that capitalizes wavelet analysis in cancer genome research and provide suggestions for researchers in this domain. It is important to point out that there is a large body of work utilizing wavelet techniques in analyzing biomedical images, such as mammography images and ultrasound images, for cancer diagnosis. This type of work is beyond the scope of this survey. Researchers interested in this topic are referred to [72], [73].

4.1 Cancer Genome Mutations

Nowadays, the general consensus about cancer is that it arises due to accumulation of mutations in critical genes that alter normal programs of cell proliferation, metabolization, differentiation, apoptosis, and so on. Ever since Boveri [74] proposed that cancer is caused by chormosomal derangements, research into the cancer genome has centered around identifying the gene mutations that are causally implicated in oncogenesis and elucidating the mechanism of their effects on the physiology of the cell. These mutations could occur in either oncogenes or tumor suppressor genes and render the cells capable of unlimited proliferation.

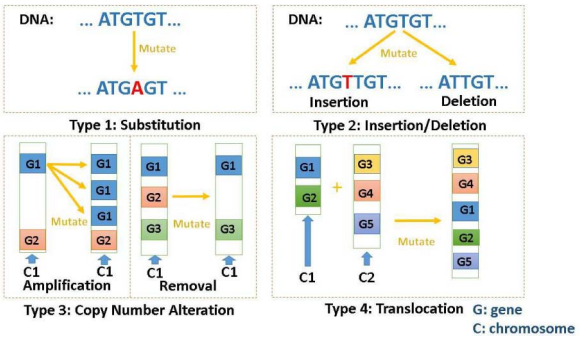


Fig. 6. The four types of mutations.

Generally speaking, there are four types of common mutations in the cancer genome, which are substitution, insertion or deletion (indel), copy number alterations, and translocations. They are illustrated in Fig. 6. In substitution, one nucleotide is substituted by another ("T" is replaced by "A" in the example). The change of the nucleotide could lead to the change of amino acid in the protein sequence, which consequentially may lead to a functional change in the proteins [2]. The second type of mutation is insertion or deletion of nucleotides of DNA in the DNA sequence, which is also called the Indel mutation. For example, in Type 2 mutation illustrated in Fig. 6, one extra "T" is inserted (insertion) or the original "G" is deleted (deletion). Such mutations could shift the codon reading frames in tumor-suppressor genes and cause a loss of function [75]. The third type of mutation is the copy number alteration, in which the number of genes is increased (amplification) or decreased (deleted), which is illustrated in the type 3 mutation in Fig. 6. Such mutations could lead to over expression of a certain gene, which causes a change in the physiology of normal cells and leads to pathogenesis of cancer [76]. The fourth type of mutation is translocation, which is illustrated in Type 4 mutation in Fig. 6. In this example, certain sections of chromosome 1 (C1) with all the genes in that section are relocated to chromosome 2 (C2). This relocation could accidentally activate other genes in chromosome 2, which may contribute to the progression of certain types of cancer, such as myelogenous leukemia [77].

In summary, the four aforementioned fundamental categories of mutations are identified as key factors in the cancer genome. These mutations lead to complex modifications in processes, such as signal transduction pathways, metabolism, histone modification, RNA splicing and protein homeostatsis, and so on. Therefore, cancer is now understood as an intricate network, integrating variations at the genomic, epigenetic, and transcriptomic levels. A more detailed description about the molecular mechanisms of cancer can be found in [1].

4.2 Wavelet Analysis in Cancer Genome Study

Dulbecco [78] argued that the complete sequence of the human genome would be an essential tool for systematically

discovering the genes that drive cancer. A system-level analysis of the cancer cell genome provides significant insights in genome mutations. Compared with the previously prohibitive cost of sequencing, the second-generation sequencing technology makes the whole cell genome analysis feasible for individual cancer. As data about the epigenome and transcriptome on a genome-wide scale of cancer grow exponentially [79], more advanced data analysis techniques are adopted. In this section, the main avenues of existing work applying wavelet analysis in cancer genome research are reviewed to illustrate how wavelet analysis can benefit cancer genome research.

4.2.1 Wavelet Analysis in Insertion/Deletion Mutations in Cancer Genome

As described in previous sections, cancer is caused by different mutations in the cancer genome. Recent studies find that the form and rate of mutations depend on the context and location of the mutation [80]. Wavelet analysis finds its application in this scenario because it provides a multiscale analysis on the sequence without predefined knowledge or parameters. Therefore, it is suitable for detecting spatial patterns of the sequences around the mutation point without any prior knowledge. In a previous study [81], the authors identified the spatial distributions of seven types of mutation related motifs, such as deletion hotspots, DNA pol pause/frameshift hotspots, and so on, with respect to insertion/deletion break points. The authors first computed the motif frequency to generate the motif frequency profile. Because of the computational simplicity, Haar wavelet analysis was applied to decompose the frequency profile. The coefficients' second raw moments on a multiscale basis were computed and they were used to measure the size of the difference between motifs occurrence patterns in insertion/deletion flanks versus control regions. Their study identified the significant spatial distribution patterns of mutation motifs. The identified motifs could be utilized as targets for some cancer medicine. In an another study presented in [82], the authors collected 1,625 spontaneous base-pair substitutions in the MutL2 strain of $Escherichiacoli$ and analyzed the spatial distribution of these mutations across the $E.coli$ genome. To accommodate the total number of mutations and describe the data clearly, the researchers generated 46 bins, each of which contains 100-kb nucleotides, starting at the origin of replication. A histogram was generated to show the distribution of missense mutations. Next, the fourth-order Daubechies wavelet transforms were applied because it is able to remove jumpy appearance of the Haar averaged signals. The analysis found that these mutations are not distributed at random but, instead, fall into a wave-like spatial pattern that is repeated almost exactly in a mirror image in the two separately replicated halves of the bacterial chromosome. These findings give some insight on different mutations occurring in the cancer genome.

4.2.2 Wavelet Analysis in Copy Number Alterations

As described in Section 4.1, copy number alterations represent a common type of structure variation in cancer genome. In general, there are two main approaches for detecting copy number alterations, which are the classic array-based gene expression approach and the more recent next generation sequencing-based approach. Wavelet analysis finds its applications in both approaches. For the first approach, the study introduced in [83] serves as a good example. In that paper, the expression level values of each gene is viewed as a time series along the chromosome coordinates. The goal of detecting the copy number alterations can be interpreted as extracting distinctive information through the curve, for example, the sharp peaks and drops of the signal in the high-noisy background. Even though the Fourier transform is useful and important in signal processing to transform the time series to frequency domain, it loses the information regarding the position of signal changes. In contrast, wavelet transform can represent the signal simultaneously in both the frequency and time domains and is well suited for detecting these sharp discontinuities. In this work, they first decomposed the signal profile into a family of multiresolution sub-bands using Haar wavelet. For each sub band, they assigned p-values to the Haar coefficients based on a null-distribution estimated from normal reference samples. Further, they selected significant coefficients by setting the threshold for false discovery rate and used the selected coefficients to identify the copy number alterations. The Haar wavelet was chosen here because it is good for analyzing piecewise constant copy number signals [84]. Other similar studies belonging to this category include [85] and [86].

The next generation sequencing-based approaches provide an alternative way for analyzing the copy number alterations in relatively high resolution. However, it suffers from the concomitant relatively high-noise issue [87]. One advantage of the wavelet analysis is decomposing the signal into a spectrum of different frequencies and the high-frequency components are sometimes corresponding to noise. Therefore, the wavelet decomposition could be adopted to perform noise reduction. The authors of [88] proposed a CNAseg algorithm to identify the copy number alteration from the second-generation sequencing data. The researchers treated the count number along the chromosome coordinates as the discrete signals and utilized an undecimated discrete wavelet transform to smooth the count data, shrunk the noisy wavelet coefficients, and computed the inverse transform from the modified coefficients to reconstruct the original signal. The Daubechies wavelet transform was used as it is better to smooth signals. The number of decomposition levels was determined by the length of the window counts for each chromosome. The reconstructed signal then went through the Hidden Markov Model (HMM) model for segmentation and the chi-square statistics-based segment merging step. Experimental results showed that those proposed approaches reduced the unevenness in read depth and decreased the number of noncopy number alteration induced HMM segments. This reduction improves the performance of the system from two perspectives. First, it reduces the false-positive detection in the final segmentation results. Second, it decreases the computational complexity in the merging step. The wavelet decomposition-based noise reduction is commonly used in studies in this research direction, such as in [85] and [87].

4.2.3 Wavelet in Machine Learning Research Framework of Cancer Genome

With the rapid growth of cancer genomics and proteomics data, more and more researchers resort to machine learning-based approaches for cancer genome analysis. The assumption here is that by mining the patterns from the existing data, mathematical models can be built to learn patterns and, therefore, make predictions in unanalyzed data. To achieve this, the raw data first need to be converted into a relatively compact and meaningful representation. This process is often termed as feature extraction. Since wavelet analysis captures the global and local characteristics of sequence data, it could be utilized to extract features from a series. Wavelet analysis is used as a feature extraction approach in some applications. For example, Liu et al. [89] proposed a framework to utilize wavelets to extract features from hundreds of protein markers in survival analysis in colorectal cancer. The authors utilized the Daubechies wavelet db7 to perform the continuous wavelet transform to extract the coefficients from the protein marker expression data. These coefficients, which contain information at different scales of the original biomarker signal, were utilized as features for cancer classification. In [90], wavelet analysis was utilized to extract features from DNA microarray data to extract important features for classification.

In summary, cancer is deemed to be a genetic disease which is caused by mutations. To combat this disease, a thorough understanding of the mechanism of mutations is necessary. Wavelet analysis, which is able to perform multiscale analysis as well as capture the local and global information of a time series, has found its application in many areas of cancer genome research, such as mutation identifications and cancer biomarker identifications.

5 Experimental Analysis

In this section, we introduce an empirical study in which the wavelet analysis is applied to solve one important problem in cancer genome research which is the identification of "driver" mutations in the cancer genome. We evaluate the effectiveness of the features computed using wavelet analysis and discuss some insights based on the experimental results.

5.1 Classifying the "Driver" and "Passenger"

As described in Section 4.1, genetic mutations are responsible for the cancers. These mutations could be categorized into "drive" mutations and "passenger" mutations. Driver mutations confer growth advantages on the cells carrying them and have been positively selected during the evolution of the cancer. They usually contribute to tumorigenic potential. On the other hand, the passenger mutations do not confer growth advantage and happen to be present in the ancestor of the cancer cell when it acquires one of its drivers. Therefore, the "passenger" mutation are usually "neutral" and are not ultimately responsible for any pathogenic characteristics exhibited by the tumor. Since driver mutations are causally implicated in oncogenesis, one of the central goals of current cancer genome analysis is the identification of cancer genes that carry driver mutations. To complicate this issue, recent systematic resequencing of the kinome of cancer cell lines has revealed that passenger mutations are much more common compared to driver mutations [91]. In addition, some mutational processes are directed at specific genomic regions and, thus, generate clusters of passenger mutations that may be mistaken for drivers [1]. All of these experimental observations make the differentiation a challenging research topic.

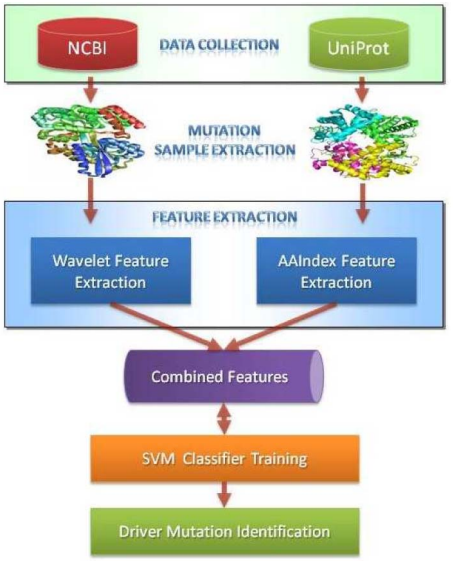


Fig. 7. The framework for driver gene identification.

This problem could be addressed by biological experiments to a certain degree, given the number of mutations is relatively small. However, with thousands of mutations in the cancer cell line, it would be important to prioritize experimental work with the hope that the driver mutations could be preferentially identified over passenger mutations. Therefore, a computational algorithm for automatically classifying the aforementioned two types of mutations is needed.

Wavelet analysis can be applied to represent the DNA sequence to generate the sequence-based features, since wavelet analysis provides multiresolution information about the sequence, which is usually missing in the primary features generated from the sequence data. Therefore, wavelet analysis combined with machine learning and data mining approaches can provide promising solutions to the problem of differentiating the genes, which harbor the driver mutations with the genes that carry passenger mutations. In this empirical study, we propose to apply wavelet analysis to the DNA sequence or protein sequence. In addition, such an analysis method does not require homology analysis. Therefore, this approach can be applied to a high-throughput system and applied to uncharacterized genes that do not show any homology to known sequences.

5.2 A Unique Computational Framework

Fig. 7 shows the architecture of the framework. First, the driver and passenger genes are collected from existing knowledge and downloaded from GenBank [92]. Next, the mutation samples are extracted according to the mutation location on the corresponding protein sequences, and those samples are represented by numerical numbers according to a certain mapping scheme. Then, wavelet transforms are

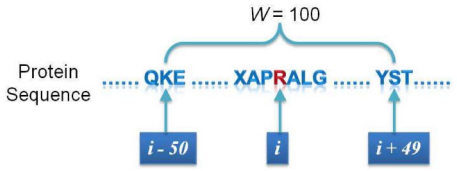


Fig. 8. Mutation sample extraction.

applied to the mutation samples to obtain original wavelet coefficients at different scales, which are sampled and converted to feature vectors. Finally, a classification technique, SVM-based classifier, is applied to classify the driver and passenger mutations. The details are discussed as follows:

- *Data collection*. We collect 29 driver genes and 58 passenger genes from the published papers and COSMIC database [92]. Based on those genes, 78 driver mutation samples and 110 passenger mutation samples are extracted.
- *Mutation sample extraction*. The mutation samples are extracted from the original protein sequence based on the mutation location using a fixed window size. To be specific, a window size of 100 is used to extract the mutation sample centered at mutation spot i. The mutation sample extraction scheme is illustrated in Fig. 8.
- *Numerical representation*. The original amino acids are converted to numerical numbers based on the mapping scheme in Table 1. In this experiment, only the real component of the complex representation is used.
- *Wavelet analysis*. The Matlab wavelet toolbox provides a powerful tool for wavelet analysis. In the current experiment, the continuous wavelet transform based on Daubechies wavelets function is used to extract wavelet coefficients from mutation samples. (The Daubechies wavelets are chosen due to their successful applications in biological sequences analysis [38], [43].) Based on the results of the empirical study, the differences between the wavelet coefficients before and after the mutation are more significant at the scale levels 2 through 100. Therefore, the scales are set to be 2:2:100, where the second 2 represents a sampling step of 2 (similar to the example illustrated in Section 3.1.2). The obtained COEFS are a 50 by 100 matrix, where each row is a coefficient sequence at a specific scale. The averages of the rows of the coefficients in the matrix are calculated to obtain a 100-dimensional feature vector.

TABLE 5
Values of Different Evaluation Criteria with Maximized F1

Feature Group ID	$C(\log_2)$	$\gamma((\log_2)$	F1	Accuracy	MCC
1	3	-7	**0.7976**	0.8500	0.6857
2	5	-7	**0.4833**	0.6889	0.3121
3	1	-5	**0.5239**	0.7056	0.3861
4	9	-13	**0.8064**	0.8389	0.6739
5	3	-9	**0.8214**	0.8667	0.7165

- *Sequence-based protein features*. In addition to the wavelet features, the amino acid index (AAindex) features [19] that represent the physicochemical properties of the proteins are also extracted.
- *Support vector machine*. The LIBSVM package [93] is one of the most popular off-the-shelf classifiers. In this study, the LIBSVM classifier is utilized as the classification model.
- *Evaluation*. In terms of evaluation, the "*Accuracy*," "*F1*," and "*Matthew's correlation coefficient*" (MCC) performance metrics are used. Here, "TP" is the total number of true-positive instances, "TN" is the total number of true-negative instances, "FP" is the total number of false-positive instances, and "FN" is the total number of false-negative instances. In addition, MCC ranges from -1 to 1. A value of $MCC = 1$ indicates the best possible prediction; while $MCC = -1$ indicates the worst possible prediction. $MCC = 0$ is expected for a random prediction scheme. The equations for different criteria are shown below:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN},$$

$$F1 = \frac{2 \cdot \frac{TP}{TP+FP} \cdot \frac{TP}{TP+FN}}{\frac{TP}{TP+FP} + \frac{TP}{TP+FN}},$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TN + FN)(TP + FN)(TN + FP)}}.$$

5.3 Experimental Results

Three experiments are conducted to evaluate the contributions and characteristics of five different groups of features. Table 4 shows the group IDs and their corresponding features. The LIBSVM classifier is utilized to evaluate those different groups of features.

The fivefold cross validation is used in all experiments. The two SVM parameters C and γ are tuned using a grid search approach to maximize one of the evaluation criteria described in the previous section. Tables 5, 6, and 7 show

TABLE 4
Feature Group ID and Features

Feature Group ID	Features	Feature Dimension
1	AAindex	544
2	Daubechies wavelets	100
3	Haar wavelets	100
4	AAindex + Daubechies wavelets	644
5	AAindex + Haar wavelets	644

TABLE 6
Values of Different Evaluation Criteria with Maximized Accuracy

Feature Group ID	$C(\log_2)$	$\gamma((\log_2)$	F1	Accuracy	MCC
1	3	-7	0.7976	**0.8500**	0.6857
2	1	-3	0.4347	**0.7220**	0.4376
3	1	-5	0.5239	**0.7056**	0.3861
4	9	-13	0.8064	**0.8389**	0.6739
5	3	-9	0.8214	**0.8667**	0.7165

TABLE 7
Values of Different Evaluation Criteria with Maximized MCC

Feature Group ID	$C(\log_2)$	$\gamma((\log_2)$	F1	Accuracy	MCC
1	13	-13	0.7934	0.8500	**0.6949**
2	1	-3	0.4347	0.7220	**0.4376**
3	1	-5	0.5239	0.7056	**0.3861**
4	9	-13	0.8064	0.8389	**0.6739**
5	3	-9	0.8214	0.8667	**0.7165**

the best performance of the cross-validation results obtained by maximizing F1, Accuracy, and MCC, respectively.

From the experimental results shown in the three figures, it could be seen that the AAindex features (Group 1) outperform the Daubechies wavelet features (Group 2) and the Haar wavelet features (Group 3). The reasons are as follows: First, the dimension of the AAindex features is 544 but both the Haar wavelet features and the Daubechies wavelet features are only of 100 dimensions. The AAindex features contain more information. In addition, each dimension of the AAindex features represents one kind of physiochemical properties, which determines the protein structure that is related to the function based on the classic biological assumption that the structure and property of the protein determine its biological functions. The wavelet transform captures relatively indirect features of the protein sequence. In terms of the Haar wavelet features and the Daubechies wavelet features, their performances are comparable and it is not clear which one outperforms the other. However, when the AAindex features are combined with the Haar wavelet features, the performance is improved compared to that using AAindex, Haar wavelet, or Daubechies wavelet features individually. It also suggests that even though the Haar wavelet features themselves do not give good performance, they could be utilized to enhance the AAindex-based features. This is relatively counterintuitive because it is easy to draw the conclusion that if a feature set with good performance is combined with the one with worse performance, an average performance is achieved. The reason is that the AAindex feature, which captures the global feature of the protein sequence, loses all the information about the sequence position. However, the sequence of the protein also determines the properties of the proteins. The wavelet-based features capture the sequence or the temporal information of the proteins and complement the AAindex features. The phenomenon that the sequence information enhances the representation of the characteristics of the protein is also observed in [94]. In that paper, the authors utilized the so-called pseudo amino acid composition, which captures the sequence information in the m-tier ($m >= 1$) correlation factor. The authors showed that by adding the correlation factors to the feature pool, the performance of the protein cellular attribute prediction could be improved. This suggests that the wavelet-based feature representation is another representation of the sequence information. In addition, combining the Daubechies wavelet and the AAIndex feature sets does not seem to consistently enhance the performance compared to using the AAIndex feature set alone. Further investigation is needed to disclose the reasons why the Haar wavelet-based features could enhance the AAindex features more than the Daubechies wavelet features from the perspective of wavelet transforms. An intuitive explanation is that the Haar wavelet features capture more the high-frequency (local) information of the original DNA sequence. However, which type of wavelets is most suitable is not an easy question to answer. Actually, the choice of the wavelet algorithm depends on the application itself and more empirical studies have to be conducted to determine the "best" wavelet function. The complete results including the SVM model parameters, all evaluation criteria, and the values obtained from the three experiments are also shown in the tables.

As shown in these tables, the SVM parameters selected to maximize the three criteria, namely F1, Accuracy, and MCC, are the same across the three runs for feature groups 3, 4, and 5. For feature group 1, the parameters selected to maximize F1 and Accuracy are the same; while the parameters are different when maximizing MCC. However, the values for the three evaluation criteria are close. Feature group 2 shows relatively big differences in Tables 5 and 6. It indicates that the classification performance using Daubechies wavelet features is not stable. In summary, for most of the feature groups, using each of the three criteria produces relatively similar results.

6 Conclusions and Future Work

In this paper, we review the current progress of using wavelets in biological sequences analysis in cancer genome. First, an overview of the framework of applying wavelet analysis in cancer genome research is given to familiarize readers with a global picture. We identify three important steps, which are numerical representations of biological sequences, wavelet transforms, and pattern recognition based on the wavelet coefficients. The numerical representation of DNA/protein sequences is crucial in the success of the overall framework and is an active research area. Different approaches are described in detail in Section 2 so that researchers in this domain could refer to these methods. Following that, different state-of-the-art wavelet analysis methods are introduced and reviewed in Section 3. The applications of wavelets in solving different biological problems are shown in that section to exemplify the pattern recognition step using the wavelet coefficients. Intuitions are also provided to bridge the gap between signal processing domain and biological research domain. Based on the foundations built in previous sections, a detailed description of applying wavelet analysis in cancer genome research is given in Section 4 to illustrate its usage in Cancer research. In Section 5, using a specific research problem, differentiating the driver mutation from the passenger mutation, we did an empirical study to illustrate the overall process. These data show that a proper combination of the wavelet coefficient-based features with protein physicochemical property-based features enhances the classification performance. However, the choice of the wavelet transform approaches could affect the performance and should be given careful attention. In summary, the application of wavelets to cancer research, as reviewed in these studies, and extended by our own empirical studies, will serve as a foundation for future wavelet research in carcinogenesis

In the future, the most imperative task is to enhance the numerical representation of the protein sequence and the

scheme of applying the wavelet transform. Other wavelet transforms, such as Morlet, Mexican Hat, and Meyer, can be considered and the detailed comparison of the performance of using different wavelet-based features shouldl be conducted. As a novel approach of representing the protein amino acid sequence information, wavelet-based features can also be compared with the existing sequence information representation methods such as the well-recognized Chou's pseudo amino acid composition [95]. In addition, another research direction is to integrate information gained from applying wavelet analysis on microarray images [96].

REFERENCES

[1] M.R. Stratton, P.J. Campbell, and P.A. Futreal, "The Cancer Genome," *Nature,* vol. 458, no. 7239, pp. 719-724, 2009.

[2] E.P. Reddy, R.K. Reynolds, E. Santos, and M. Barbacid, "A Point Mutation Is Responsible for the Acquisition of Transforming Properties by the T24 Human Bladder Carcinoma Oncogene," *Nature,* vol. 300, no. 5888, pp. 149-152, July 1981.

[3] R.R. Voss, "Evolution of Long-Range Fractal Correlations and 1/F Noise in DNA Base Sequences," *Physical Rev. Letters,* vol. 68, no. 25, pp. 3805-3808, 1992.

[4] T. George and T. Thomas, "Discrete Wavelet Transform De-Noising in Eukaryotic Gene Splicing," *BMC Bioinformatics,* vol. 11, no. Suppl. 1, article S50, 2010.

[5] M. Abo-Zahhad, S.M. Ahmed, and S.A. Abd-Elrahman, "Genomic Analysis and Classification of Exon and Intron Sequences Using DNA Numerical Mapping Techniques," *Int'l J. Information Technology and Computer Science,* vol. 4, no. 8, pp. 22-36, July 2012.

[6] R. Zhang and C.T. Zhang, "Z Curves, an Intuitive Tool for Visualizing and Analyzing the DNA Sequences," *J. Biomolecular Structure and Dynamics,* vol. 11, no. 4, pp. 767-782, 1994.

[7] C. Zhang, R. Zhang, and H. Ou, "The Z Curve Database: A Graphic Representation of Genome Sequences," *Bioinformatics/Computer Applications in the Biosciences,* vol. 19, pp. 593-599, 2003.

[8] A. Rushdi and J. Tuqan, "Gene Identification Using the Z-Curve Representation," *Proc. IEEE Int'l Conf. Acoustics, Speech and Signal Processing (ICASSP '06),* vol. 2, p. II, May 2006.

[9] B.D. Silverman and R. Linsker, "A Measure of DNA Periodicity," *J. Theoretical Biology,* vol. 118, no. 3, pp. 295-300, 1986.

[10] H.T. Chang, C.J. Kuo, N.W. Lo, and W.Z. Lv, "DNA Sequence Representation and Comparison Based on Quaternion Number System," *Int'l J. Advanced Computer Science and Applications,* vol. 3, no. 11, pp. 40-46, 2012.

[11] C. Cattani, "Complex Representation of DNA Sequences," *Comm. in Computer and Information Science,* vol. 13, pp. 528-537, 2008.

[12] S.W.A. Bergen and A. Antoniou, "Application of Parametric Window Functions to the STDFT Method for Gene Prediction," *Proc. IEEE Pacific Rim Conf. Comm., Computers and Signal Processing (PACRIM)* pp. 324-327, Aug. 2005.

[13] D. Anastassiou, "Genomic Signal Processing," *IEEE Signal Processing Magazine,* vol. 18, no. 4, pp. 8-20, July 2001.

[14] P.D. Cristea, "Genetic Signal Representation and Analysis," *Proc. Soc. of Photo-Optical Instrumentation Engineers (SPIE) Conf.,* vol. 4623, pp. 77-84, 2002.

[15] H. Zhou and H. Yan, "Autoregressive Models for Spectral Analysis of Short Tandem Repeats in DNA Sequences," *Proc. IEEE Int'l Conf. Systems, Man and Cybernetics (SMC '06),* vol. 2, pp. 1286-1290, Oct. 2006.

[16] A.S. Nair and S.P. Sreenadhan, "A Coding Measure Scheme Employing Electron-Ion Interaction Pseudopotential (EIIP)," *Bioinformation,* vol. 1, no. 6, pp. 197-202, 2006.

[17] S. Maetschke, M.W. Towsey, and M. Boden, "Blomap: an Encoding of Amino Acids Which Improves Signal Peptide Cleavage Site Prediction," *Proc. Third Asia Pacific Bioinformatics Conf.,* pp. 141-150, 2005.

[18] C.H. Wu and J.W. McLarty, *Neural Networks and Genome Informatics.* Elsevier Science Inc., 2000.

[19] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa, "AAindex: Amino Acid Index Data, Progress Report 2008," *Nucleic Acids Research,* vol. 36, pp. D202-D205, 2008.

[20] K. Chou, "Prediction of Protein Cellular Attributes Using Pseudo Amino Acid Composition," *Proteins,* vol. 43, no. 3, pp. 246-255, May 2001.

[21] K. Chou, "Using Amphiphilic Pseudo Amino Acid Composition to Predict Enzyme Subfamily Classes," *Bioinformatics,* vol. 21, no. 1, pp. 10-19, Jan. 2005.

[22] M.X. Chen, B.C. Liu, W.Y. Yan, and B.R. Shen, "Wavelet Transform Based Protein Decoy Discrimination," *Proc. Third Int'l Conf. Bioinformatics and Biomedical Eng. (ICBBE '09),* pp. 1-4, June 2009.

[23] I. Cosic, "Macromolecular Bioactivity: Is It Resonant Interaction between Macromolecules?-Theory and Applications," *IEEE Trans. Biomedical Eng.,* vol. 41, no. 12, pp. 1101-1114, Dec. 1994.

[24] R. Swanson, "A Vector Representation for Amino Acid Sequences," *Bull. of Math. Biology,* vol. 46, no. 4, pp. 623-639, 1984.

[25] C. Yin and S. Yau, "Numerical Representation of Dna Sequences Based on Genetic Code Context and Its Applications in Periodicity Analysis of Genomes," *Proc. IEEE Symp. Computational Intelligence in Bioinformatics and Computational Biology,* pp. 223-227, 2008.

[26] M. Randic, K. Mehulic, D. Vukicevic, T. Pisanski, D. VikicTopic, and D. Plavsic, "Graphic Representation of Proteins as Four-Color Maps and Their Numerical Characterization," *J. Molecular Graphics and Modelling,* vol. 27, no. 5, pp. 637-641, Jan. 2008.

[27] P.A. He, X.F. Li, J.L. Yang, and J. Wang, "A Novel Descriptor for Protein Similarity Analysis," *Match Communications in Math. and in Computer Chemistry,* vol. 65, pp. 445-458, 2011.

[28] F. Bai and T. Wang, "On Graphical and Numerical Representation of Protein Sequences," *J. Biomolecular Structure and Dynamics,* vol. 23, no. 5, pp. 537-545, 2006.

[29] S.B. Needleman and C.D. Wunsch, "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins," *J. Molecular Biology,* vol. 48, no. 3, pp. 443-453, 1970.

[30] C. Branden et al., *Introduction to Protein Structure,* vol. 2. Garland, 1991.

[31] C.A.F. Anderson and S. Brunak, "Representation of Protein-Sequence Information by Amino Acid Subalphabets," *AI Magazine,* vol. 25, no. 1, pp. 97-104, 2004.

[32] C. Gargour, M. Gabrea, V. Ramachandran, and J.M. Lina, "A Short Introduction to Wavelets and Their Applications," *IEEE Circuits and Systems Magazine,* vol. 9, no. 2, pp. 57-68, Second Quarter 2009.

[33] A. Haar, "Zur Theorie Der Orthogonalen Funktionensysteme," *Mathematische Annalen,* vol. 69, no. 3, pp. 331-371, 1910.

[34] J.B. Allen and L.R. Rabiner, "A Unified Approach to Short-Time Fourier Analysis and Synthesis," *Proc. IEEE,* vol. 65, no. 11, pp. 1558-1564, Nov. 1977.

[35] S.G. Mallat, "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 11, no. 7, pp. 674-693, July 1989.

[36] I. Daubechies, *Ten Lectures on Wavelets,* series CBMS-NSF Regional Conf. Series in Applied Math. Soc. for Industrial and Applied Math., 1992.

[37] D. Gabor, "Theory of Communication," *IEEE Radio Comm. Eng. J.,* vol. 93, no. 26, pp. 429-441, Nov. 1946.

[38] M. Sifuzzaman, M.R. Islam, and M.Z. Ali, "Application of Wavelet Transform and Its Advantages Compared to Fourier Transform," *J. Physical Sciences,* vol. 13, pp. 121-134, 2009.

[39] A.K. Nagar and D. Sokhi, "On Wavelet-Based Adaptive Approach for Gene Comparison," *Int'l J. Intelligent Systems Technologies and Applications,* vol. 5, pp. 104-114, 2008.

[40] Q. Liu, S. Wan, and Y. Sun, "Identification of Splice Sites Based on Discrete Wavelet Transform and Support Vector Machine," *Proc. Int'l Conf. Bioinformatics and Biomedical Eng.,* 2008.

[41] C. Cattani, "Fractals and Hidden Symmetries in DNA," *Math. Problems in Eng.,* vol. 2010, pp. 1-32, 2010.

[42] C. Cattani, "On the Existence of Wavelet Symmetries in Archaea DNA," *Computational and Math. Methods in Medicine,* vol. 2012, pp. 1-21, 2012.

[43] J.K. Meher, M.K. Raval, P.K. Meher, and G.N. Nash, "Wavelet Transform for Detection of Conserved Motifs in Protein Sequences with Ten Bit Physico-Chemical Properties," *Int'l J. Information and Electronics Eng.,* vol. 2, no. 2, pp. 200-204, 2012.

[44] M. Ahmad, A. Abdullah, and K. Buragga, "A Novel Optimized Approach for Gene Identification in DNA Sequences," *J. Applied Sciences,* vol. 11, no. 5, pp. 806-814, 2011.

[45] R. Gupta, A. Mittal, K. Singh, P. Bajpai, and S. Prakash, "A Time Series Approach for Identification of Exons and Introns," *Proc. 10th Int'l Conf. Information Technology (ICIT '07),* pp. 91-93, Dec. 2007.

[46] N. DasGupta, S. Lin, and L. Carin, "Sequential Modeling for Identifying CPG Island Locations in Human Genome," *IEEE Signal Processing Letters,* vol. 9, no. 12, pp. 407-409, Dec. 2002.

[47] O. Abbasi and J. Rasi, "Exonic Regions Finding on DNA Sequences Using RLS Algorithm and Denoising with Discrete Wavelet," *Proc. Int'l Symp. Artificial Intelligence and Signal Processing (AISP),* pp. 66-70, June 2011.

[48] O. Abbasi, A. Rostami, and G. Karimian, "Identification of Exonic Regions in DNA Sequences Using Cross-Correlation and Noise Suppression by Discrete Wavelet Transform," *BMC Bioinformatics,* vol. 12, pp. 1-10, 2011.

[49] S. Deng, Z. Chen, G. Ding, and Y. Li, "Prediction of Protein Coding Regions by Combining Fourier and Wavelet Transform," *Proc. Third Int'l Congress on Image and Signal Processing (CISP),* vol. 9, pp. 4113-4117, Oct. 2010.

[50] K.B. Murray, D. Gorse, and J.M. Thornton, "Wavelet Transforms for the Characterization and Detection of Repeating Motifs," *J. Molecular Biology,* vol. 316, no. 2, pp. 341-363, 2002.

[51] J.P. Mena-Chalco, H. Carrer, Y. Zana, and R.M. Cesar, "Identification of Protein Coding Regions Using the Modified Gabor-Wavelet Transform," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 5, no. 2, pp. 198-207, Apr.-June 2008.

[52] T.S. Gunawan and E. Ambikairajah, "Parallel Implementation of Genomic Sequences Classification Using Modified Gabor Wavelet Transform on Multicore Systems," *Proc. Int'l Conf. Biomedical Eng. (ICoBE),* pp. 165-168, Feb. 2012.

[53] J.A.T. Machado, A.C. Costa, and M.D. Quelhas, "Wavelet Analysis of Human Dna," *Genomics,* vol. 98, no. 3, pp. 155-163, 2011.

[54] J. Qiu, R. Liang, X. Zou, and J. Mo, "Prediction of Protein Secondary Structure Based on Continuous Wavelet Transform," *Talanta,* vol. 61, no. 1, pp. 285-293, Apr. 2003.

[55] R. Jiang and H. Yan, "Studies of Spectral Properties of Short Genes Using the Wavelet Subspace Hilbert Huang Transform (WSHHT)," *Physica A: Statistical Mechanics and Its Applications,* vol. 387, no. 16, pp. 4223-4247, 2008.

[56] A. Arneodo, Y. d'Aubenton Carafa, E. Bacry, P.V. Graves, J.F. Muzy, and C. Thermes, "Wavelet Based Fractal Analysis of DNA Sequences," *Physica D-Nonlinear Phenomena,* vol. 96, pp. 291-320, 1996.

[57] M. El-Zanaty, M. Saeb, A.B. Mohamed, and S.K. Guirguis, "Haar Wavelet Transform of the Signal Representation of Dna Sequences," *Int'l J. Computer Science and Comm. Security,* vol. 1, pp. 56-62, 2011.

[58] E. Lin and E. Linton, "Wavelet Packet Analysis of DNA Sequences," *Proc. Fifth Int'l Conf. Bioinformatics and Biomedical Eng. (ICBBE),* pp. 1-3, May 2011.

[59] S. Chandra and A.Z. Rizvi, "Wavelet Analysis of Hiv-1 Genome," *Proc. Int'l Assoc. Computer Science and Information Technology - Spring Conf. (IACSITSC '09),* pp. 559-561, Apr. 2009.

[60] A. Arneodo, E. Bacry, P.V. Graves, and J.F. Muzy, "Characterizing Long-Range Correlations in DNA Sequences from Wavelet Analysis," *Physical Rev. Letters,* vol. 74, pp. 3293-3296, 1995.

[61] E. Linton, P. Albee, P. Kinnicutt, and E. Lin, "Multiresolution Analysis of DNA Sequences," *Proc. Second Int'l Conf. Computer Research and Development,* pp. 218-222, May 2010.

[62] L. Ravichandran, A. Papandreou-Suppappola, A. Spanias, Z. Lacroix, and C. Legendre, "Waveform Mapping and Time-Frequency Processing of DNA and Protein Sequences," *IEEE Trans. Signal Processing,* vol. 59, no. 9, pp. 4210-4224, Sept. 2011.

[63] B. Weng, G. Xuan, J. Kolodzey, and K.E. Barner, "Discriminating Dna Sequences from Terahertz Spectroscopy - A Wavelet Domain Analysis," *Proc. IEEE 32nd Ann. Northeast Bioeng. Conf.,* pp. 211-212, 2006.

[64] L.S. Cheong, F. Lin, and H.S. Seah, "Frequency-Domain Algorithms for Visual Analysis on Genomic Structures in Prokaryotes," *Proc. Int'l Conf. Computer Graphics, Imaging and Visualisation,* pp. 96-103, July 2006.

[65] S.B. Arniker and H.K. Kwan, "Graphical Representation of DNA Sequences," *Proc. IEEE Int'l Conf. Electro/Information Technology (eit '09),* pp. 311-314, June 2009.

[66] K.B. Murray, D. Gorse, and J.M. Thornton, "Wavelet Transforms for the Characterization and Detection of Repeating Motifs," *J. Molecular Biology,* vol. 316, no. 2, pp. 341-363, Feb. 2002.

[67] D.T. Jones, M. Tress, K. Bryson, and C. Hadley, "Successful Recognition of Protein Folds Using Threading Methods Biased by Sequence Similarity and Predicted Secondary Structure," *Proteins: Structure, Function, and Bioinformatics,* vol. 37, no. S3, pp. 104-111, 1999.

[68] C.H. Trad, Q. Fang, and I. Cosic, "Protein Sequence Comparison Based on the Wavelet Transform Approach," *Protein Eng.,* vol. 15, pp. 193-203, Mar. 2002.

[69] A. Sabarish R. and T. Thomas, "A Frequency Domain Approach to Protein Sequence Similarity Analysis and Functional Classification," *Signal and Image Processing,* vol. 2, no. 1, pp. 36-49, Mar. 2011.

[70] J. Qiu, S. Luo, J. Huang, and R. Liang, "Using Support Vector Machine for Prediction of Protein Structural Classes Based on Discrete Wavelet Transform," *J. Computation Chemistry,* vol. 30, no. 8, pp. 1344-1350, June 2009.

[71] H. Chen, F. Gu, and F. Liu, "Predicting Protein Secondary Structure Using Continuous Wavelet Transform and Chou-Fasman Method," *Proc. IEEE Conf. Eng. in Medicine and Biology Soc.,* vol. 3, pp. 2603-2606, Mar. 2005.

[72] R. Mousa, Q. Munib, and A. Moussa, "Breast Cancer Diagnosis System Based on Wavelet Analysis and Fuzzy-Neural," *Expert Systems with Applications,* vol. 28, no. 4, pp. 713-723, 2005.

[73] M. Ramaraj and S. Raghavan, "A Survey of Wavelet Techniques and Multiresolution Analysis for Cancer Diagnosis," *Proc. Int'l Conf. Computer, Comm. and Electrical Technology (ICCCET),* pp. 109-114, 2011.

[74] T. Boveri, "Concerning the Origin of Malignant Tumours by Theodor Boveri," *J. Cell Science,* vol. 121, no. Supplement 1, pp. 1-84, 2008.

[75] X. Xu, K. Zhu, F. Liu, Y. Wang, J. Shen, J. Jin, Z. Wang, L. Chen, J. Li, and M. Xu, "Identification of Somatic Mutations in Human Prostate Cancer by RNA-Seq," *Gene,* vol. 519, no. 2, pp. 343-347, May 2013.

[76] M. Chinnam and D.W. Goodrich, "RB1, Development, and Cancer," *Current Topics in Developmental Biology,* vol. 94, pp. 129-169, 2011.

[77] M. Jongen-Lavrencic, S. Salesse, R. Delwel, and C.M. Verfaillie, "BCR/ABL-Mediated Downregulation of Genes Implicated in Cell Adhesion and Motility Leads to Impaired Migration toward Ccr7 Ligands Ccl19 and Ccl21 in Primary BCR/ABL-Positive Cells," *Leukemia,* vol. 19, no. 3, pp. 373-380, 2005.

[78] R. Dulbecco, "A Turning Point in Cancer Research: Sequencing the Human Genome," *Science,* vol. 231, no. 4742, pp. 1055-1056, 1986.

[79] L. Chin, W.C. Hahn, and G. Getz, "Making Sense of Cancer Genomic Data," *Genes and Development,* vol. 25, no. 6, pp. 534-555, 2011.

[80] E.V. Ball, P.D. Stenson, S.S. Abeysinghe, M. Krawczak, D.N. Cooper, and N.A. Chuzhanova, "Microdeletions and Microinsertions Causing Human Genetic Disease: Common Mechanisms of Mutagenesis and the Role of Local Dna Sequence Complexity," *Human Mutation,* vol. 26, no. 3, pp. 205-213, 2005.

[81] E.M. Kvikstad, F. Chiaromonte, and K.D. Makova, "Ride the Wavelet: A Multiscale Analysis of Genomic Contexts Flanking Small Insertions and Deletions," *Genome Research,* vol. 19, no. 7, pp. 1153-1164, 2009.

[82] P.L. Foster, A.J. Hanson, H. Lee, E.M. Popodi, and H. Tang, "On the Mutational Topology of the Bacterial Genome," *G3: Genes, Genomes, Genetics,* vol. 3, no. 3, pp. 399-407, 2013.

[83] L. Song, "Computational Analysis of Genome-Wide DNA Copy Number Changes," PhD dissertation, Virginia Polytechnic Inst. and State Univ., 2011.

[84] J. Shore, "On the Application of Haar Functions," *IEEE Trans. Comm.,* vol. C-21, no. 3, pp. 209-216, Mar. 1973.

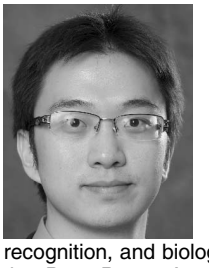
[85] L.M. Tran, B. Zhang, Z. Zhang, C. Zhang, T. Xie, J.R. Lamb, H. Dai, E.E. Schadt, and J. Zhu, "Inferring Causal Genomic Alterations in Breast Cancer Using Gene Expression Data," *BMC Systems Biology,* vol. 5, no. 1, pp. 121-134, 2011.

[86] E. Ben-Yaacov and Y.C. Eldar, "A Fast and Flexible Method for the Segmentation of aCGH Data," *Bioinformatics,* vol. 24, no. 16, pp. i139-i145, 2008.

[87] K.C. Amarasinghe, J. Li, and S.K. Halgamuge, "CoNVEX: Copy Number Variation Estimation in Exome Sequencing Data Using HMM," *BMC Bioinformatics,* vol. 14, no. Supplement 2, article S2, 2013.

[88] S. Ivakhno, T. Royce, A.J. Cox, D.J. Evers, R.K. Cheetham, and S. Tavarē, "Cnaseg - A Novel Framework for Identification of Copy Number Changes in Cancer from Second-Generation Sequencing Data," *Bioinformatics,* vol. 26, no. 24, pp. 3051-3058, 2010.

[89] Y. Liu, U. Aickelin, J. Feyereisl, and L.G. Durrant, "Wavelet Feature Extraction and Genetic Algorithm for Biomarker Detection in Colorectal Cancer Data," *Knowledge-Based Systems,* vol. 37, pp. 502-514, 2013.
[90] A.M. Sarhan, "Wavelet-Based Feature Extraction for Dna Microarray Classification," *Artificial Intelligence Rev.,* vol. 39, no. 3, pp. 237-249, 2013.
[91] C. Greenman, P. Stephens, and R. Smith, "Patterns of Somatic Mutation in Human Cancer Genomes," *Nature,* vol. 446, no. 1, pp. 153-158, 2007.
[92] D.A. Benson, I. Karsch-Mizrachi, K. Clark, D.J. Lipman, J. Ostell, and E.W. Sayers, "Genbank," *Nucleic acid research,* vol. 39, pp. D32-D37, 2011.
[93] C. Chang and C. Lin, "Libsvm : A Library for Support Vector Machines," *ACM Trans. Intelligent Systems and Technology,* vol. 2, no. 27, pp. 1-27, Apr. 2011.
[94] J.D. Hand, "Measuring Classifier Performance," *Machine Learning,* vol. 77, pp. 103-123, 2009.
[95] K.C. Chou, "Pseudo Amino Acid Composition and Its Applications in Bioinformatics, Proteomics and System Biology," *Current Proteomics,* vol. 6, no. 4, pp. 262-274, 2009.
[96] L. Prasad and S.S. Iyengar, *Wavelet Analysis with Applications to Image Processing.* CRC, 1997.



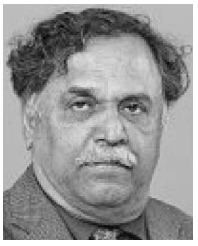
Tao Meng received the BS degree in biotechnology with honor in July 2006, from the University of Science and Technology of China (USTC), Hefei, China, He is currently working toward the PhD degree working under the supervision of Dr. Mei-Ling Shyu, at the Department of Electrical and Computer Engineering (ECE), University of Miami (UM) since January 2009. His research interests include multimedia concept analysis and mining, biomedical image pattern recognition, and biological sequence semantics discovery. He received the Best Paper Award from the IEEE International Conference on information reuse and integration in 2012. He is a student member of the IEEE.

Ahmed T. Soliman received the MS degree in computer engineering from the University of Miami in 2007, and is currently working toward the PhD degree at the College of Engineering, University of Miami. His research interests include medical image processing and data mining.

Mei-Ling Shyu (M'95-SM'03) received the PhD degree from the School of Electrical and Computer Engineering and three master's degrees, all from Purdue University, West Lafayette, Indiana. She is a full professor at the Department of Electrical and Computer Engineering (ECE), University of Miami (UM) since June 2013. Prior to that, she was an associate/assistant professor in ECE at UM from January 2000. Her research interests include multimedia data mining, management and retrieval, and security. She received 2012 Computer Society Technical Achievement Award and ACM 2012 Distinguished Scientists Award. She received the Best Paper Award in 2012, the Best Published Journal Article in IJMDEM for 2010 Award, the Best Student Paper Award with her student in 2009. She serves/served as an associate editor for several journals including *IEEE Transactions on Human-Machine Systems*, and on the editorial board of many other journals. She is a fellow of the SIRI. She is a senior member of the IEEE.

Yimin Yang received the MS degree in computer science from Florida International University (FIU), Miami, in 2012. She is currently working toward the PhD degree at the School of Computing and Information Sciences, FIU. Her research interests include multimedia data mining, multimedia systems, image and video processing, and information retrieval. She is a student member of the IEEE.

Shu-Ching Chen received the master's degrees in computer science, electrical engineering, and civil engineering and the PhD degree in electrical and computer engineering in 1992, 1995, 1996, and 1998, respectively, all from Purdue University, West Lafayette, Indiana. He is a full professor at the School of Computing and Information Sciences (SCIS), Florida International University (FIU), Miami, since August 2009. Prior to that, he was an assistant/associate professor in SCIS at FIU from 1999. His primary research interests include content-based image/video retrieval, multimedia data mining, multimedia systems, and disaster information management. He received the ACM Distinguished Scientist Award in 2011. He received the Best Paper Award from 2006 IEEE International Symposium on Multimedia. He received the IEEE Systems, Man, and Cybernetics Society's Outstanding Contribution Award in 2005 and was the corecipient of the IEEE Most Active SMC Technical Committee Award in 2006. He is a senior member of the IEEE.



S.S. Iyengar is a ryder professor and the director of the School of Computing and Information Sciences at the Florida International University, Miami. He is a pioneer in the field of distributed sensor networks/sensor fusion, and high-performance computing. He has published more than 400 research papers and has written eight texts and edited nine books published by John Wiley & Sons, Prentice Hall, CRC Press, Springer Verlag, and so on. He received the Distinguished Alumnus Award of the Indian Institute of Science, Bangalore, and received the IEEE Computer Society Technical Achievement for the contributions to sensor fusion algorithms, and parallel algorithms. He has received a Lifetime Achievement Award conferred by the International Society of Agile Manufacturing (ISAM) in research and administration and a lifelong contribution to the fields of Engineering and Computer Science at Indian Institute of Technology (BHU). He serves on the advisory board of many corporations and universities in the world. He has served on many National Science Boards such as NIH—National Library of Medicine in Bioinformatics, National Science Foundation review panel, NASA Space Science, Department of Homeland Security, Office of Naval Security, and many others. His contribution was a centerpiece of this pioneering effort to develop image analysis for our science and technology and to the Goals of the US Naval Research Laboratory. The impact of his research contributions can be seen in companies National Labs like Raytheon, Telecordia, Motorola, the United States Navy, DARPA agencies, and so on. He is also the founding editor of *International Journal of Distributed Sensor Networks*. He is currently the editor of *ACM Computing Surveys*. He is also a member of the European Academy of Sciences, a fellow of the IEEE, ACM, AAAS, Society of Design and Process Program (SPDS), and Institution of Engineers (FIE).

John S. Yordy is an assistant professor of radiation oncology at the University of Texas Southwestern. He completed the residency in radiation oncology at MD Anderson Cancer Center before joining UT Southwestern. He is a practicing radiation oncologist, focusing on head and neck cancer, lung cancer and stereotactic ablative radiation therapy, and has an active translational research program in head and neck cancer. He also has an interest in genomics and applications of genomics to clinical cancer care.

Puneeth Iyengar received the bachelor's of science degree from MIT, and the MD and PhD degrees from the Albert Einstein College of Medicine in New Yory City before completing residency in radiation oncology at MD Anderson Cancer Center in Houston, Texas. He is an assistant professor of radiation oncology at UT Southwestern Medical Center at Dallas. He is a member of the Harold Simmons Cancer Center, a coleader of the Thoracic Disease Oriented Team for the Cancer Center, and a leader of the Thoracic Radiation Oncology Disease Oriented Team. He treats lung cancer patients with radiation and runs a translational research laboratory studying lung cancer therapeutic resistance and cancer related cachexia. He has received several awards, including research grants from the National Lung Cancer Partnership, Radiological Society of North America, UT Southwestern's Presidential Research Council, Lung Cancer Research Foundation, and the Sidney Kimmel Foundation for Cancer Research.