Wavelet Analysis in Current Cancer Genome to Identify Driver Mutation

Saiba Najeeb(PG Scholar)
Department of Computer Science and Engineering
Younus College of Engineering and Technology
Pallimukku, Kollam-691010

Prof. Nijil Raj N
Department of Computer Science and Engineering
Younus College of Engineering and Technology
Pallimukku, Kollam-691010

***Abstract*—Now a days the expanse of biological sequence data of the cancer genome upsurges exponentially, which calls for effectual and current algorithms that may recognize patterns hidden beneath the raw data which may distinguish cancer crisis. From a signal processing idea, biological units including DNA and protein sequences, have been viewed as one-dimensional signals. Therefore, the signal processing techniques are used many researches to find out the potentially important patterns within these sequences. In recent years, wavelet transforms have become a significant mathematical analysis tool, with a widespread and ever increasing choice of applications. The adaptability of wavelet analytic techniques has forged new interdisciplinary bounds by presenting common solutions to apparently different problems and providing a new unifying perspective on problems of cancer genome research. In this paper, an empirical study of how wavelet analysis is applied to cancer bioinformatics,that is to identify the driver mutation in cancer genome. We evaluate the effectiveness of the features computed using wavelet analysis and in addition to the wavelet features, the amino acid index (AAindex) features are also extracted.Proper combination of the wavelet coefficient-based features with protein physicochemical property-based features enhances the classification performance.**

***Index Terms*—Cancer genome, driver mutation, wavelet analysis, AAindex.**

I. Introduction

CANCER is one of the greatest medical causes of mortality.It is liable for one in eight deaths worldwide. Critical treads in developing systemic and local therapies for cancer have been made possible from the increasing knowledge of the human protein and the relevant genetic changes found in tumors.

Cancer is an evolutionary process, all cancers are believed to share a common pathogenesis. Each is the outcome of a process of Darwinian evolution happening among cell populations within the microenvironments provided by the tissues of a multicellular organism.Analogous to Darwinian evolution happening in the origins of species, cancer development is based on two integral processes, the continuous acquisition of heritable genetic variation in individual cells by more-or-less random mutation and natural selection acting on the resultant phenotypic diversity.

At present, a generalizable concept of cancer states that malignancies outcome from accumulated mutations in genes that upsurge 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, apparently through the development of mutations, leading to metastases and increased patient mortality.Mutations every so 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 preoccupied with in unique sequences, as are the resulting proteins.

Current struggles to understand how mutations in DNA lead to the development of cancers have been partly limited by the overall inability to examine through the massive quantities of data generated by cancer genome sequencing projects and the studies of individual investigators. As a consequence, there is a necessity for tools to parse through this large sum of data to present relevant gene changes that may be serious for either understanding how cancers develop or/and determining how they could ultimately be treated. From the view point of signal processing , 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.

Section 2 describes related works in this domain.Methodology is discussed in section 3. Section 4 gives detailed explanation of the proposed method.Paper conclude in section 5.

II. Related Work

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.

During the period of the sequencing of the human genome (19902003), cancer researchers continued to accumulate knowledge of the basic mechanisms of cancer, and using a variety of clever cloning strategies, with steadily improving sequencing capabilities, identified the majority of the most potent oncogenes and tumor suppressors. An inventory of the genes associated with cancer yielded 291 cancer genes based on mutation data available in the literature: ;1% of the coding sequence [3]. It was noted that 90% of these genes were somatically mutated, 20% germline mutated, and 10% could be found in both categories. The division between germline and somatic genes is a mysterious dichotomy that remains unexplained in the most current inventory The most common form of variation in the 2004 inventory was translocation leading to the production of oncogenic fusion proteins. Until 2004, no one had studied more than a handful of genes at any one time in a single patient.

That was the state of cancer genomic research at the threshold of the genomic era of cancer research: an era heralded by the availability of the high-quality reference genome, and the dramatic explosion of DNA sequence data fueled by the introduction of inexpensive massively parallel sequencing instruments. This year is the 10^{th} anniversary of the completion of that remarkable milestone in sciencethe completion of the reference human genome. At this juncture, we recapitulate some of the key findings and challenges that have emerged from the sequence analysis of the cancer genome.

With base-level resolution of the human reference genome in hand, cancer researchers turned to the large-scale study of mutation, with the promise of generating the entire catalog of mutations peculiar to a given disease as well as to a single patient. Figure 1 tracks the development of some of the key technologies, resources, and milestones in the development of the present-day armamentarium of cancer mutation discovery. Massively parallel sequencing was introduced by Roche 454 and Illumina in 20042006 and soon demonstrated the feasibility of sequencing complete normal and tumor genomes of exemplar human subjects on both platforms [4]. At the time, it appeared that the application of whole-genome sequencing to routine research and clinical diagnosis might be on the horizon. Although the use of whole-genome sequence (WGS) is far from routine today, the results generated so far are lending insight into the potential of WGS for diagnostic, prognostic, and therapeutic improvement in the treatment of cancer patients.

Using PCR and dye-terminator sequencing, Vogelstein and colleagues amplified and sequenced each coding exon of 18,000 genes, defined by the human genome sequence, in 11 each of breast and colorectal tumors [5]. This brute force whole-exome sequencing (WES) approach afforded for the first time a comprehensive view of the mutation profile of each patient, which, when summed across patients, revealed the cancer genes for the patients in the given cohort. In one stroke, the mutation profile, composed of recurrently mutated genes, plus a collection of one-off mutations belonging to pathways and processes known to be involved in tumorigenesis, were revealed for a cancer. The fact that the most frequently mutated genes they observed, APC, TP53, and KRAS for colon cancer and TP53 for breast cancer, recapitulated what was already known, validated the approach and paved the way for expanded application of genome-scale sequencing.

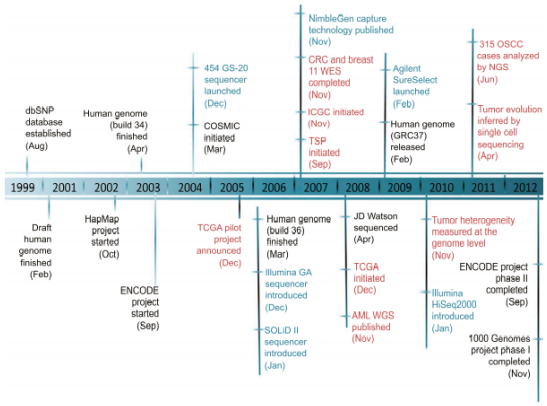

Fig. 1. MAJOR EVENTS IN A DECADE OF CANCER GENOMICS.(Dark blue) Major advances in massively parallel sequencing platforms and targeted enrichment technologies; (black) major large-scale projects designed to catalog genomic variations of normal human individuals; (red) cancer genomics. (dbSNP) Database of single nucleotide polymorphism; (HapMap) haplotype map of the human genome; (ENCODE) Encyclopedia of DNA Elements; (COSMIC) Catalog of Somatic Mutations in Cancer; (TCGA) The Cancer Genome Atlas; (GA) genome analyzer; (CRC) colorectal carcinoma; (WES) wholeexome sequencing; (ICGC) International Cancer Genome Consortium; (TSP) tumor sequencing project; (AML) acute myeloid leukemia; (WGS) wholegenome sequencing; (OSCC) ovarian small cell carcinoma.

The introduction of DNA sequence enrichment technologies from NimbleGen and Agilent [6] enabled WES on large scales. WES has additional advantages over WGS in that the average depth of coverage is about fivefold greater, and the cost of sequencing, data processing and storage are all much less. Given the relative tractability of interpreting variation in the coding sequence compared to intergenic or intronic mutations, the period between 2004 and 2013 has seen a profusion of tumor types analyzed in large cohorts (100500 patients), mainly byWES. WGS for a variety of tumors has also been reported and, in spite of the smaller numbers of patients, has led to surprising insights into cancer biology, based largely on analysis of structural variation in tumor genomes. Using WGS, genetic alterations observed in the DNA of the cancer cell span six orders of magnitude, from single-base point mutations to chromosome-scale amplification, using different modes of sequence analysis [7] available today.

III. METHODOLOGY

The main objective of this work was to classify the driver mutation.In order to fulfill our purpose, our proposed framework firstly several features are extracted and the SVM was applied in order to recognize diver mutation.

TABLE I
THE COMPLEX REPRESENTATION OF 20 AMINO ACIDS

Amino Acid Name	Symbol	Complex Number Repr.
Alanine	A	0.61 + 88.3i
Arginine	R	0.60 + 181.3i
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 + 118.2i
Theronine	T	0.05 + 118.2i
Triptophan	W	2.65 + 227.0i
Tyrosine	Y	1.88 + 193.0i
Valine	V	1.32 + 141.4i

A. Numerical Representation

For further analysis of biological sequences they need to be encoded in a suitable format. Then the input biological sequences can be processed as signals and signal processing techniques such as wavelets can be utilized to extract hidden features out of these sequences. The encoding process is a kind of numerical substitution for each character symbols that forms the biological sequence.

Using these encoding process two types of biological sequences such as DNA nucleotide sequences and protein amino acid sequences can be successfully mapped to required formats for processing. DNA sequence is easier compared with protein sequences for encoding since only 4 character symbols are there with DNA sequences where proteins are represented with 20 amino acids.It can been successfully used in cancer research to classify driver genes and passenger genes

The original amino acids in the protein sequence are converted to numerical representation 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 [8], where the hydrophobicity is the real part and the residue volume is the imaginary part. In this experiment, only the real component of the complex representation is used and the mapping scheme is shown in Table 1.

B. Wavelet Analysis

Jean Batiste Joseph Fourier, a French mathematician developed the concept of Fourier Trignometric series. Through this concept he represented a periodic function in terms of a weighted sum of cosine and sine functions. This was considered as the origin of wavelets theory. In 1909 Alfred Haar developed Haar Wavelets family which is considered the simplest wavelets. Compared with harmonic functions used in Fourier analysis, wavelets can be used to analyze a given signal in terms of functions that are more finite in time. One of the important property of the Haar wavelets which gave wide acceptance across the globe was the scaling property which give more accurate results in modeling functions. The idea of multiresolution, which is the base theory of versatile wavelets families, was proposed[9]. Using this multiresolution concept, Daubechies [10] created the most frequently used Daubechies wavelets family.This is evident from the above statements that the wavelet theory originated from Fourier Transform.

Fourier Transform is one way to find frequency content and measure the signal composition in frequency. Fourier Transform can be calculated using equation(1). Here F is the frequency in Hertz and Ωt is the phase in radians:

$$FT\{x(t)\} = x(\Omega) = \int_{-\infty}^{\infty} x(t)e^{-j\Omega t}dt, \Omega = 2\pi F. \quad (1)$$

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 signals 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. Gabor 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 (2), where $g(t-b)e^{-j\Omega t} = \ ^{*}_{\Omega,b}(t)$is the complex conjugate of $\ _{\Omega,b}(t)$:

$$STFT_{g(\Omega,b)}\{x(t)\} = \int_{-\infty}^{\infty} x(t)g(t-b)e^{-j\omega t}dt = X_g(\omega, b). \quad (2)$$

One of the limitations of STFT is due to the fixed size window used. A narrow window and wide window results in poor frequency resolution and poor time resolution respectively. Also it is really difficult to determine the time intervals where a particular frequency exists. Thus wavelet transform was proposed to get rid of these problems as an alternative to STFT. The definition of continuous wavelet transform is given below :

$$\begin{aligned} CWT_x\ (a,b) &= X\ \ (a,b) \\ &= \frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} x(t)[\ \ ^{*}(\frac{t-b}{a})]dt \\ &= \langle x(t), \psi^{*}_{a,b}(t)\rangle \end{aligned} \quad (3)$$

here a and b are the scaling and translation parameters, respectively, and $\ ^{*}_{a,b}(t) = \frac{1}{\sqrt{a}}\ ^{*}(\frac{t-b}{a})$ is the mother wavelet (base function), a prototype for generating the other window functions.

In summary, wavelet analysis techniques outrun the traditional FT in the following perspectives[11]:

- wavelets are suitable for analysis on both stationary and nonstationary signals where FT is less useful in analyzing nonstationary signals;

- wavelets are well localized in both time and frequency domains, where the standard FT is localized in frequency domain only;
- the base functions of wavelets can both be scaled and shifted, while the FT can only be scaled; and
- wavelets have solid mathematics foundation and a wider range of applications than FT such as nonlinear regression and compression.

Wavelet families generally belong to one of the following types.

- 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.

C. AAindex Features

In addition to the wavelet features, 566 amino acid index (AAindex) features[12] that represent the physicochemical properties of the proteins are also extracted from the database AAindex.

Protein structures and functions are defined by the combinations of physicochemical and biochemical properties of 20 naturally occurring amino acids that are the building-blocks of proteins. A wide variety of properties of amino acids have been investigated through a large number of experiments and theoretical studies. Each of these amino acid properties that can be represented by a set of 20 numerical values is referred to as an amino acid index.

The AAIndex currently contains 566 amino acid indices. Each entry consists of an accession number, a short description of the index, the reference information and the numerical values for the properties of 20 amino acids.

IV. Proposed Method

In this section,we discuss our work in which the wavelet analysis is applied to solve our important problem in cancer genome which is the identification of ”driver” mutation.

A. Classifying the ”Driver” and ”Passenger”

As described above section 1., genetic mutation are responsible for the cancer.These mutations could be classified into drive mutations and passenger mutations. Driver mutations confer growth benefits on the cells carrying them and have been positively selected during the evolution of the cancer. They usually donate 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 obtains one of its drivers. Therefore, the passenger mutation are generally neutral and are not eventually responsible for any pathogenic characteristics exhibited by the tumor. Since driver mutations are causally concerned 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 sys tematic resequencing of the kinome of cancer cell lines has revealed that passenger mutations are much more common equated to driver mutations [13]. 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 explanations make the differentiation a challenging research topic.

This delinquent 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 significant to prioritize experimental work with the hope that the driver mutations could be specially identified over passenger mutations. Therefore, a computational algorithm for automatically categorizing the aforementioned two types of mutations is needed.

Wavelet analysis and AAindex features 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 and the AAindex features captures the global features of the protein. Therefore, wavelet analysis combined with AAindex features ,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 along with AAindex features 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.

B. Computational Framework

Fig. 2 displays the architecture of the framework. Main, the driver and passenger genes are collected from existing knowledge and downloaded from GenBank [14]. 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 applied to the mutation samples to obtain original wavelet coefficients at different scales, which are sampled and converted to feature

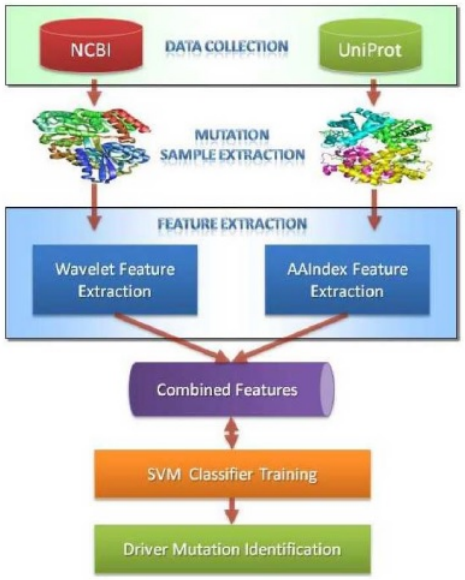

Fig. 2. THE FRAMEWORK FOR DRIVER GENE IDENTIFICATION.

vectors. Finally, a classification technique, SVM-based classifier, is applied to classify the driver and passenger mutations. The details are discussed as follows:

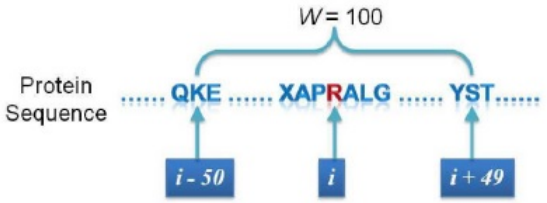

Fig. 3. MUTATION SAMPLE EXTRACTION.

- *Data collection*. We collect 29 driver genes and 58 passenger genes from the published papers and COSMIC database [14]. Based on those genes, 40 driver mutation samples and 39 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. 3.
- *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 [11], [15].) Based on the results of the 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. 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.
- Sequence-based protein features. In addition to the wavelet features, the amino acid index (AAindex) features [11] that represent the physicochemical properties of the proteins are also extracted.
- Support vector machine. The LIBSVM package is one of the most popular off-the-shelf classifiers. In this study, the LIBSVM classifier is utilized as the classification model.Here we also experimented Naive Bayes Classifier but LIBSVM is more accurate than bayes classifier .Table 2 show there accuracy of classfication.

TABLE II
CLASSIFIER AND THEIR ACCURACY.

Classifier	Accuracy
LIBSVM	0.941
Naive Bayes Classifier	0.91

- *Evaluation*. In terms of evaluation, the Accuracy, F1, and Matthews 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:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

$$F1 = \frac{2 \cdot \frac{TP}{TP + FP} \cdot \frac{TP}{TP + FN}}{\frac{TP}{TP + FP} + \frac{TP}{TP + FN}} \tag{5}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TN + FN)(TP + FN)(TN + FP)}} \tag{6}$$

TABLE III
FEATURE SET AND ITS SIZE.

Group Sl. No	Features	Feature Size
Step 1		
1	Daubechies wavelets	100
2	AAindex	566
Step 2		
3	AAindex + Daubechies wavelets	666

TABLE IV
DIFFERENT VALUES FOR FEATURE SET

Group Sl. No	Method	Accuracy	F1	MCC
Step 1				
1	3-fold cross validation	0.778302	0.618474	0.559625
2	3-fold cross validation	0.858491	0.628378	0.717492
Step 2				
3	3-fold cross validation	0.859774	0.628986	0.707831

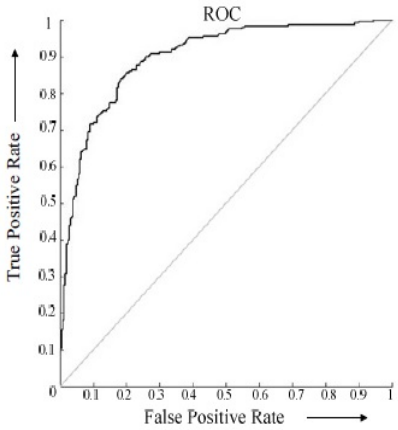

Fig. 4. ROC CURVE WHEN AAINDEX AND WAVELET FEATURES ARE SELECTED

V. Experimantal Results

One experiments are conducted to evaluate the contributions and characteristics of three different groups of features. Table 3 shows the group IDs and their corresponding features. The LIBSVM classifier is utilized to evaluate those different groups of features.The threefold cross validation is used in the experiment.

From the experimental results shown in the Table 4, it could be seen that the AAindex features (Group 2)outperform the Daubechies wavelet features (Group 1). The reasons are as follows: First, the dimension of the AAindex features is 566 but the Daubechies wavelet features are only of 100 dimensions respectively. The AAindex features contain more information. In addition, each dimension of the AAindex features represents one kind of physiochemical properties.The two SVM parameters C and γ are tuned using grid search.

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.

TABLE V
COMPARISON B/W EXISTING AND PROPOSED SYSTEM.

Method	Accuracy
Existing Method	0.8389
Proposed Method	0.859774

VI. Conclusion

We did an study,that show a proper combination of wavelet coefficient-based features with protein physicochemical property-based features enhance the classification performance. Conversely, 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 by the work, will serve as a foundation for future wavelet research in carcinogenesis.

In 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 used 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 Chous pseudo amino acid composition. In addition, another research direction is to integrate information gained from applying wavelet analysis on microarray images.

References

[1] Michael R Stratton, Peter J Campbell, and P Andrew Futreal. The cancer genome. *Nature*, 458(7239):719–724, 2009.

[2] E Premkumar Reddy, Roberta K Reynolds, Eugenio Santos, and Mariano Barbacid. A point mutation is responsible for the acquisition of transforming properties by the t24 human bladder carcinoma oncogene. 1982.

[3] P Andrew Futreal, Lachlan Coin, Mhairi Marshall, Thomas Down, Timothy Hubbard, Richard Wooster, Nazneen Rahman, and Michael R Stratton. A census of human cancer genes. *Nature Reviews Cancer*, 4(3):177–183, 2004.

[4] David A Wheeler, Maithreyan Srinivasan, Michael Egholm, Yufeng Shen, Lei Chen, Amy McGuire, Wen He, Yi-Ju Chen, Vinod Makhijani, G Thomas Roth, et al. The complete genome of an individual by massively parallel dna sequencing. *nature*, 452(7189):872–876, 2008.

[5] Laura D Wood, D Williams Parsons, Siân Jones, Jimmy Lin, Tobias Sjöblom, Rebecca J Leary, Dong Shen, Simina M Boca, Thomas Barber, Janine Ptak, et al. The genomic landscapes of human breast and colorectal cancers. *Science*, 318(5853):1108–1113, 2007.

[6] Thomas J Albert, Michael N Molla, Donna M Muzny, Lynne Nazareth, David Wheeler, Xingzhi Song, Todd A Richmond, Chris M Middle, Matthew J Rodesch, Charles J Packard, et al. Direct selection of human genomic loci by microarray hybridization. *Nature methods*, 4(11):903–905, 2007.

[7] Lynda Chin, William C Hahn, Gad Getz, and Matthew Meyerson. Making sense of cancer genomic data. *Genes & development*, 25(6):534–555, 2011.

[8] Changchuan Yin and Stephen S-T Yau. Numerical representation of dna sequences based on genetic code context and its applications in periodicity analysis of genomes. In *Computational Intelligence in Bioinformatics and Computational Biology, 2008. CIBCB'08. IEEE Symposium on*, pages 223–227. IEEE, 2008.

[9] Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7):674–693, 1989.

[10] Ingrid Daubechies et al. *Ten lectures on wavelets*, volume 61. SIAM, 1992.

[11] M Sifuzzaman, MR Islam, and MZ Ali. Application of wavelet transform and its advantages compared to fourier transform. 2009.

[12] Shuichi Kawashima, Piotr Pokarowski, Maria Pokarowska, Andrzej Kolinski, Toshiaki Katayama, and Minoru Kanehisa. Aaindex: amino acid index database, progress report 2008. *Nucleic acids research*, 36(suppl 1):D202–D205, 2008.

[13] Christopher Greenman, Philip Stephens, Raffaella Smith, Gillian L Dalgliesh, Christopher Hunter, Graham Bignell, Helen Davies, Jon Teague, Adam Butler, Claire Stevens, et al. Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132):153–158, 2007.

[14] Dennis A Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. Genbank. *Nucleic acids research*, 41(D1):D36–D42, 2013.

[15] JK Meher, MK Raval, PK Meher, and GN Dash. Wavelet transform for detection of conserved motifs inprotein sequences with ten bit physico-chemicalproperties. *International Journal of Information and ELECTRONICS Engineering*, 2(2):200, 2012.