JMSCR Vol||12||Issue||02||Page 66-77||February 2024

http://jmscr.igmpublication.org/home/
ISSN (e)-2347-176x ISSN (p) 2455-0450
crossref DOI: https://dx.doi.org/10.18535/jmscr/v12i02.10

IGM Publication
Journal Of Medical Science And Clinical Research
An Official Publication Of IGM Publication

Digital Signal Processing Approaches in the field of Genomics: A Recent Trend

Authors
Shivani Saxena[1], Ahsan Z, Rizvi[2]
[1]Dept. of Computer Sciences and Engineering Institute of Advanced Research Gandhinagar, India
[2]Dept. of Computer Sciences and Engineering Institute of Advanced Research Gandhinagar, India

Abstract
Digital signal processing (DSP) techniques have emerged as powerful tools in the field of genomics, enabling researchers to extract valuable insights from complex genetic data. This research paper presents a comprehensive analysis of the recent trends and advance- ments in applying DSP approaches to genomics. The objective is to provide an overview of the transformative role of DSP in genomic data analysis, variant calling, and interpretation. By leveraging DSP methods such as filtering, feature extraction, time-frequency analysis, and machine learning algorithms, researchers can enhance the quality of genetic signals, identify genetic variants, and gain a deeper understanding of genomic processes. The paper highlights key applications of DSP in genomics, including DNA sequence analysis, RNA expression pro- filing, epigenetics, and genome-wide association studies. Additionally, the challenges associated with applying DSP techniques in genomics, such as signal noise, data in- tegration, and computational complexity, are discussed. This research paper serves as a valuable resource for researchers, bioinformaticians, and geneticists seeking to harness the power of DSP in genomics, advancing our knowledge of genetic diseases and paving the way for personalized medicine and precision healthcare.
Keywords: *Digital signal processing, Genome analysis, Feature extraction, DNA sequence analysis, RNA expression profiling.*

Introduction

Digital signal processing techniques have been widely applied in various fields, including biomedical research and analysis [1,2]. In recent years, there has been a notable shift towards utilizing these approaches in the field of genome analysis [2,3]. The advancements in high-throughput sequencing technologies have generated vast amounts of genomic data,

making it necessary to develop efficient computational methods for extracting meaningful information from the data. Digital signal processing (DSP) techniques offer valuable tools and algorithms for processing and analyzing genomic signals, enabling researchers to uncover valuable insights into the structure, function, and variation of genomes. One prominent application of digital signal processing in genome analysis is in the identification and annotation of functional elements within the genome [2 - 5]. Techniques such as wavelet analysis, Fourier analysis, and other spectral analysis methods have been employed to detect patterns, motifs, and regions of interest in DNA and protein sequences. These methods allow for the identification of protein-coding regions, regulatory elements, and other functional elements, aiding in the understanding of gene expression, regulation, and protein function. Furthermore, digital signal processing approaches have been utilized in the analysis of genetic variation and genomic alterations. Copy number variations, single nucleotide polymorphisms, and structural variations in the genome can be detected and characterized using signal processing algorithms. This enables the identification of genetic markers associated with diseases, population studies, and personalized medicine applications.

Signal denoising and smoothing techniques have also found utility in genome analysis. With the advent of high-throughput sequencing technologies, genomic data is often plagued by noise and artifacts. Digital signal processing methods[6], such as wavelet denoising, filtering, and deconvolution, can enhance the quality of the genomic signals, improving downstream analysis and interpretation. Moreover, the integration of digital signal processing with other computational approaches, such as machine learning and data mining, has further enhanced the analysis of genomic data. Classification, clustering, and pattern recognition algorithms derived from signal processing principles have been employed to classify diseases, predict outcomes, and identify genetic markers associated with specific phenotypes.

In summary, the application of digital signal pro- cessing approaches in genome analysis has emerged as a recent trend in the field of computational bi- ology. These techniques provide powerful tools for processing, analyzing, and interpreting genomic sig- nals, enabling researchers to unravel the complex biological mechanisms underlying health and disease. As genomic data continues to grow in volume and complexity, digital signal processing will continue to play a crucial role in extracting valuable knowledge and insights from the vast genomic datasets. In this paper, our focus is on exploring the recent trends, ad- vancements, challenges, and potential applications of digital signal processing (DSP) approaches specifically in the field of genome analysis. We aim to examine the impact of DSP techniques in various aspects of genome analysis, including data preprocessing, variant calling, and data analysis. By investigating how DSP can enhance our understanding of genetic diseases and drive innovation in personalized health- care, we provide valuable insights for researchers, clinicians, and geneticists.

Digital Signal Processing Techniques

In this section, we delve into the fundamentals of Digital Signal Processing (DSP) techniques and their connection to

the Wavelet Transform. We explore the utilization of wavelets in the analysis of biological sequences, highlighting their significance in this field. Furthermore, we present an overview of various wavelet families that are commonly employed in sequence analysis. In the realm of signal analysis, a significant breakthrough was made by the French mathematician Jean Baptiste Joseph Fourier (1768- 1830), who developed a method to represent any periodic function as a weighted sum of cosine and sine functions[1]. This groundbreaking technique, known as Fourier analysis, laid the foundation for understanding the frequency components of a signal. However, Fourier analysis assumes that signals are stationary, meaning their properties do not change over time. In reality, many signals encountered in real-world applications are non-stationary, exhibiting variations in both the time and frequency domains. To address this limitation, alternative methods have been developed that offer a more localized representation of signals.

One such approach is the Haar Wavelet, introduced by Alfred Haar in 1909[6]. Haar wavelets provide a simple wavelet set that can be used to analyze signals. Unlike the harmonic functions used in Fourier analysis, Haar wavelets are finite in time, making them well-suited for capturing localized features in effectively analyze non-stationary signals and extract localized information in both the time and frequency domains. The versatility and adaptability of wavelet analysis make it a valuable tool in various domains, including image processing, audio signal analysis, biomedical signal processing, and many others.

In the real world, signals are often non-stationary, meaning their properties change over time and fre- quency domains. Unlike stationary signals, which maintain consistent properties, non-stationary signals exhibit variations in characteristics such as amplitude, frequency, and phase. Fourier Transform (FT) is a widely used mathematical tool that allows us to ana- lyze signals in the frequency domain. It decomposes a signal into its constituent frequency components, providing valuable insights into the signal's spectral content. The FT of a signal x (t) is defined as:

$$FT[x(t)] = X(\omega) = \int_{-\infty}^{\infty} x(t)e^{-j\omega t}\,dt \qquad (1)$$

where, $\omega = 2\pi f$. f is the frequency in Hertz and ω is the phase in radians: The integral captures the contribution of each frequency component in the signal, revealing the amplitude and phase information associated with each frequency. By performing FT, we can effectively analyze the frequency content of a signal, enabling us to identify dominant frequencies, harmonics, and other spectral characteristics. How- ever, FT assumes that signals are stationary over the entire duration of analysis, which may not hold true for many real-world signals. To address the limitations of FT in analyzing non-stationary signals, Gabor pro- posed the STFT technique[6]. In this technique, the windowing method is used. The analysis is done on each separate window. Let g(t) be the sliding window of constant length. then STFT of the signal is defined as

$$STFT[x(t)] = X(\omega) = \int_{-8}^{?8} x(t)g(t - b)e^{-j\omega t}\,dt \qquad (2)$$

The Haar wavelet's scaling properties have been proven to enhance its accuracy in signal analysis[6]. Another significant development in signal analysis is the concept of Multiresolution Analysis (MRA), introduced by S.G. Mallat in 1989 [6]. MRA enables the analysis of signals at

multiple scales, allowing for a more comprehensive understanding of signal properties. Building upon MRA, Ingrid Daubechies introduced the Daubechies wavelet family for signal analysis [6]. The Daubechies wavelets offer a versatile set of functions that can accurately represent signals with varying characteristics.

In addition to the Haar and Daubechies wavelets, other wavelet families have been devised to address specific signal analysis needs. For example, the Morlet Wavelets, developed by Jean Morlet, emerged as an alternative to the Gabor window used in Short-Time

But due to its fixed window size, there are some limitations. By using a narrow window size, there is poor frequency resolution, whereas using a wider window size results in poor time resolution. To solve the resolution problem, alternative approaches such as wavelet analysis have emerged [6]. Wavelet analy- sis offers a time-frequency representation of signals, allowing for localized analysis of signal properties in both the time and frequency domains. This provides a more accurate depiction of signal characteristics, es- pecially for signals with time-varying frequency com- ponents. The Continuous Wavelet Transform (CWT) of the signal x(t) is defined as

$$CWT^{\psi}(a,b) = X(a,b) = \int_{-\infty}^{\infty} x(t)\,\psi\left(\frac{t-b}{a}\right)\frac{dt}{\sqrt{a}} \quad (3)$$

where a and b are the scaling and translation parameters, respectively. $\psi^*_{a,b}$ is the mother wavelet (base function), used to generate other window func- tions. Wavelet analysis techniques outperform the traditional FT. A summary of this is presented in Table [I].

Table I: Comparison of Wavelet Transform and Fourier Transform.

Properties	FT	WT
Stationary Signal	Yes	Yes
Non-Stationary Signal	No	Yes
Time Domain	No	Yes
Frequency Domain	Yes	Yes
Scaling	Yes	Yes
Shifting	No	Yes

In the following sections, we will delve into the concept of wavelet analysis and its application in the analysis of biological sequences, providing a compre- hensive understanding of the benefits and applica- tions of this powerful signal-processing technique.

Wavelet Analysis on Biological Sequence

In the last few decades, studying the human genome has been a crucial task done by scientists[7]. As an example, taking the DNA sequence of any species and studying it is a tedious task. To solve this problem, Digital Signal Processing (DSP) techniques are used. Basically, the DNA sequence consists of 4 types of nucleotides, namely A, T, G, and C. Con- verting these nucleotides into a mathematical form is a prerequisite for processing DSP techniques to

*A **Wavelet Families for Biological Sequence***

Commonly used wavelet families in biological se- quences are described here. Basically, there are 4 types of Wavelet families[6]:

- **Orthogonal wavelets with scaling finite impulse responses (FIR) filters**
  Wavelets that fall under this category are defined by the low-pass scaling filter. Example: Haar, Daubechies, Coiflets, and Symlet
- **Biorthogonal wavelets with scaling finite im- pulse responses filters**
  These wavelets have two scaling filters one for reconstruction and another for

decomposition. The Bior Splines wavelet family is an example of this type.

- **Wavelets with scaling function**
  These wavelets are defined by the mother wavelet, father wavelet, and scaling function. A perfect example of this wavelet is Meyer Wavelet.
- **Wavelets without scaling filters and withoutscaling function**
  This wavelet has a time-domain representation only. Example: e Morlet and Mexican Hat.

A summary of Wavelets Used in Biological Sequence Analysis is shown below:

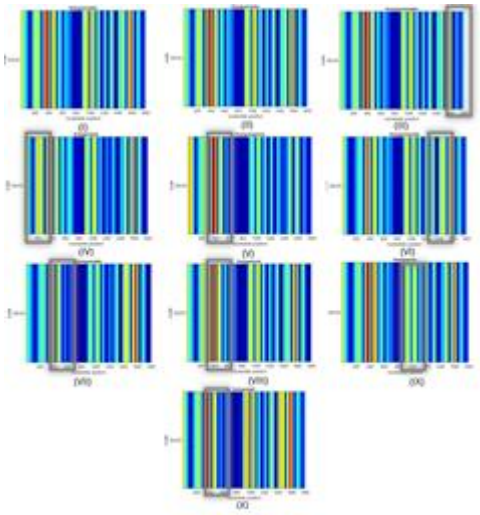
$$\psi(t) = \begin{cases} 1 & 0 = t < 1/2 \\ -1 & 1/2 = t < 1 \\ 0 & otherwise \end{cases} \quad (4)$$

Scaling function:

$$\varphi(t) = \begin{cases} 1 & 0 = t < 1 \\ 0 & otherwise \end{cases} \quad (5)$$

Applications: Feature Extraction for spice sites identification[9], Speed DNA sequence search and gene sequence analysis[10],[11], Structure Analysis in conserved Protein motif detection[12].

- **Daubechies**
  Defined by wavelet coefficients. Belongs to orthogonal wavelets defining Discrete Wavelet Transform (DWT). It is characterized by the num-ber of vanishing moments.
  Applications: Noise reduction for gene identifi- cation[13],[14], CpG island identification[15], Feature extraction and noise reduction in exons and introns prediction[14], Structure analysis in conserved protein motif detection[12].



- **Meyer**
  Belongs to an orthogonal Continuous wavelet defined in the frequency domain. It is indefinitely differentiable with infinite support.

$$\begin{cases} \frac{1}{2\pi}\sin\left(\frac{\pi}{2}v\left(\frac{3|\omega|}{2\pi}-1\right)\right)e^{\frac{i\omega}{2}} & \frac{2\pi}{3} < |\omega| < \frac{4\pi}{3} \\ 0 & other\ wise \end{cases} \quad (6)$$

where,

$$v(x) = \begin{cases} 0 & x < 0 \\ x & 0 < x < 1 \\ 1 & x > 1 \end{cases} \quad (7)$$

Application: Noise reduction for exomic regions identification [12], [12].

- **Mexican Hat**
  A special case of the family of continuous wavelets. It is the negative normalized second derivative of the Gaussian filter function.

$$\psi(t) = \frac{2}{\sqrt{2\sigma\pi}}\left(1 - \frac{t^2}{\sigma^2}\right)e^{\frac{t^2}{2\sigma^2}} \quad (8)$$

Application: Noise reduction for protein coding region prediction[16], Noise reduction for re- peating motifs detection[17].

- **Morlet**

Also known as the Gabor wavelet and belongs to the continuous wavelet. It is composed of complex exponential multiplied by a Gaussian envelope. There is a trade-off between time and frequency resolution.

$$\psi_\sigma(t) = c_\sigma \pi^{-\frac{1}{4}} exp^{-\frac{t^2}{2}} e^{igt} - k_\sigma \quad (9)$$

wjere, $k_\sigma = exp^{-\frac{\sigma}{2}}$ is defined by the admissi-

bility criterion and the normalized constant c_σ is defined as:

$$c \quad \pm e^{-\sigma^2} - 2e^{-\frac{3}{2}} \quad (10)$$

Application: Identification of protein coding re- gions [18], [19], Analysis of Human DNA [20], Protein secondary structure prediction [21].

- **Shannon**

Belongs to the family of Continuous wavelets obtained from the frequency B-Spline wavelets. It is indefinitely differentiable with infinite support

Real Shannon Wavelet:

$$\psi(t) = \frac{\sin\frac{\pi t}{2}\cos\frac{3\pi t}{2}}{\frac{\pi t}{2}} \quad (11)$$

Complex Shannon Wavelet:

$$\psi(t) = \frac{\sin(\pi t)}{\pi t} exp(-j\omega t) \quad (12)$$

Application: Analysis of Human DNA [20].

Numerical Representation Of The Biological Sequences

Biological sequences in genome evaluations are an- alyzed by converting them into a numerical form, en- abling the application of various mathematical tools. In the field of genomics, there are two main types of biological sequences: DNA sequences and protein sequences.

Biological Sequence Analysis Using Wavelets

In this section, we will explore the advancements made in the analysis of biological sequences using wavelet mathematical tools.

Sequence analysis. In DNA sequence analysis, wavelet transforms are used to decompose the DNA strand and identify specific regions of interest, such as protein-coding regions. Wavelet functions allow for a localized analysis of the DNA sequence, uncovering hidden patterns and aiding in the detection of structural characteristics. This approach has been applied in various areas, including gene prediction, sequence alignment, functional annotation, and evolutionary studies. Different wavelet functions, such as Gabor wavelets, discrete wavelets (e.g., Daubechies and Meyer wavelets), and Hidden Markov Tree approaches, have been employed to analyze DNA sequences and identify important features.

In protein sequence analysis, wavelets have been utilized for motif search, protein structure compari- son, secondary structure prediction, and protein clus- tering. Wavelet analysis captures both time and fre- quency information, making it well-suited for analyz- ing the complex nature of protein sequences. It has been used to identify motifs in protein sequences and compare protein structures, even when the se- quential identity is low. Wavelet analysis has also been employed in predicting the secondary struc- ture of proteins, providing valuable insights into pro- tein characteristics and aiding in clustering protein sequences to understand evolutionary relationships. Overall, wavelet analysis offers a powerful mathe- matical toolset for the analysis of DNA and protein sequences, enabling the identification of important features, uncovering hidden

patterns, and gaining in- sights into the functional elements and characteristics embedded in these sequences.

RNA Sequence Analysis

RNA sequence analysis is a vital area of research that helps in understanding gene expression and protein synthesis. RNA molecules, including mRNA, tRNA, and rRNA, have distinct roles in cellular processes. The RNA sequence is composed of four nucleotides: Adenine (A), Uracil (U), Guanine (G), and Cytosine (C). In a study[30], RNA sequence analysis was conducted using the TV-Curve representation, which captures the secondary structure information of RNA molecules. Wavelet transform and fractal dimension analysis were then applied to compare the secondary structures of different RNA sequences.

By utilizing wavelet analysis in RNA sequence anal-ysis, researchers can uncover important structural characteristics of RNA molecules. This approach pro- vides insights into the complexity and functional properties of different RNA molecules, contributing to the understanding of gene expression and protein synthesis. Wavelet analysis aids in capturing both lo- cal and global features of the RNA sequence, allowing for the detection of hidden patterns and structural variations. It enables researchers to compare and analyze the secondary structures of RNA molecules, facilitating further investigations into their biological functions and molecular interactions. Overall, wavelet analysis serves as a valuable tool in RNA sequence analysis, offering a comprehensive approach to study the structural properties and functional implications of RNA molecules. It enhances our understanding of gene expression and protein synthesis processes, contributing to advancements in molecular biology and related fields.

Cancer Genome Analysis

cancer genome analysis is a vital field that helps in understanding the molecular basis of cancer. Wavelet analysis has proven to be a valuable tool in analyzing the cancer genome, providing insights into various types of mutations such as point mutations, copy number alterations, and translocations. Wavelet analysis offers the advantage of simultaneous localization of time and frequency information, allowing researchers to study genomic alterations at different resolutions. By examining the patterns and characteristics of these mutations, researchers can gain insights into the underlying mechanisms driving cancer development.

For substitution mutations, wavelet analysis can help in identifying specific substitution patterns and investigating their association with cancer subtypes, prognosis, and treatment response. By analyzing insertion or deletion mutations, wavelet analysis enables the identification of specific patterns and their potential functional consequences. Copy number alterations can be analyzed using wavelet analysis to identify amplifications or deletions, providing insights into their location, extent, and functional impact. Additionally, wavelet analysis can be employed to investigate translocation mutations, studying the genomic regions involved and their effects on gene expression patterns.

By leveraging the power of wavelet analysis, re- searchers aim to improve early cancer detection, enhance personalized treatment strategies, and con- tribute to advancements in cancer research and pa- tient care. The unique properties of wavelets, such as multi-scale analysis and the ability to capture variations in the signal, make them well-suited for mutation detection and

characterization in cancer genomics. Overall, wavelet analysis plays a significant role in cancer genome analysis, aiding in the understanding of the molecular mechanisms under- lying cancer development and progression. It helps in identifying important genomic patterns associated with mutations and can potentially lead to improved diagnostic markers and targeted therapies for cancer.

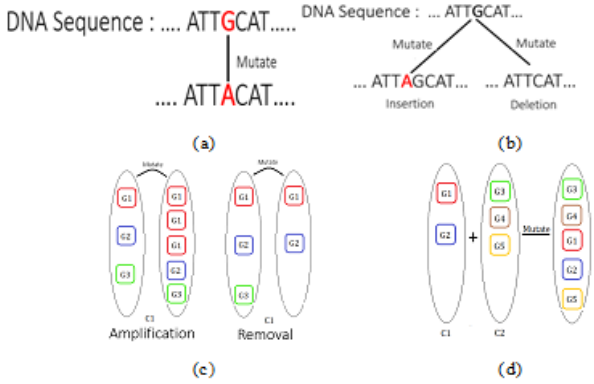


Fig. 2: (a) Substitution Mutation, (b) Insertion / Dele-tion Mutation, (c) Copy Number Alteration Mutation,
(d) Translocation Mutation

Challenges In Genome Signal Analysis

Despite the significant advancements and potential benefits of digital signal processing (DSP) approaches in genome analysis, there are several challenges that researchers face in effectively analyzing genomic sig- nals. These challenges stem from the unique char- acteristics and complexities of genomic data. In this section, we discuss some of the key challenges en- countered in genome signal analysis.

- Data Volume and Complexity: Genomic data is inherently vast and complex. With the advent of high-throughput sequencing technologies, the amount of genomic data being generated has increased exponentially. Analyzing and interpreting this large-scale data requires efficient storage, computational resources, and advanced algorithms capable of handling big data challenges.
- Noise and Artifacts: Genomic signals are often contaminated with various sources of noise and artifacts. These can arise from experimental variations, sequencing errors, or technical biases introduced during data acquisition and processing. Accurately identifying and removing noise while preserving the underlying signal is a critical challenge in genome signal analysis.
- Variability and Heterogeneity: Genomic signals exhibit inherent variability and heterogeneity due to genetic variations across individuals, cell types, and biological conditions. This variability poses challenges in accurately detecting and characterizing genomic features and patterns, especially in the presence of confounding factors and biological noise.
- Interpretability and Validation: Interpreting and validating the findings from genome signal analysis can be challenging. While DSP techniques enable the extraction of meaningful features and patterns, attributing biological significance to these findings requires careful interpretation and validation. Integrating additional biological knowledge, functional annotations, and experimental validations are essential to ensure the reliability and reproducibility of the results.
- Computational Complexity: Implementing complex DSP algorithms for genome signal analysis can be computationally demanding. Processing and analyzing large-scale genomic datasets require efficient algorithms, parallel computing

architectures, and optimized software implementations to achieve reasonable computation times.

- Integration of Multi-omics Data: Genomic research often involves integrating multi-omics data, such as genomics, transcriptomics, epigenomics, and proteomics, to gain a comprehensive understanding of biological processes. Integrating and analyzing these heterogeneous data types pose significant challenges in terms of data integration, data harmonization, and developing suitable computational models to extract meaningful insights.

Overcoming these challenges requires interdisci- plinary collaborations, advancements in algorithm development, integration of domain knowledge, and continuous efforts to improve data quality, analysis pipelines, and validation strategies. Addressing these challenges will unlock the full potential of genome signal analysis and facilitate advancements in person- alized medicine and precision healthcare.

Conclusions

Digital signal processing (DSP) techniques have revolutionized genomics research by enabling the analysis of complex genetic data. These methods en- hance signal quality, extract relevant features, and im- prove variant calling accuracy. DSP approaches such as filtering, feature extraction, and time-frequency analysis are valuable in DNA sequence analysis, RNA expression profiling, epigenetics, and genome-wide association studies. Integrating machine learning al- gorithms with DSP techniques allows for decipher- ing vast amounts of genomic data and uncovering meaningful patterns, paving the way for personalized medicine and precision healthcare. However, chal- lenges persist in handling large-scale genomic data, addressing computational complexities, and ensuring interpretability and reproducibility. Future research should focus on developing robust and scalable DSP algorithms, integrating multi-omics data, and advancing data visualization techniques. DSP's contributions to genomics research have facilitated breakthroughs in understanding genetic diseases, drug discovery, and personalized healthcare, with further advancements expected to enhance genomics research and its im- pact on human health.

References

1. 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 Quarter2009.
2. cancer paper
3. Sudipta Acharya, Laizhong Cui, Yi Pan, "A Re- fined 3-in-1 Fused Protein Similarity Measure: Application in Threshold-Free Hub Detection", IEEE/ACM Transactions on Computational Biol- ogy and Bioinformatics (Volume: 19, Issue: 1, Jan.-Feb. 1 2022), 13 February 2020 , D.O.I.
4. Yu Tian, Ruiqing Zheng, Zhenlan Liang, Sun- ing Li, Fang-Xiang Wu, Min Li , "A data-driven clustering recommendation method for single-cell RNA-sequencing data", Tsinghua Science and Technology (Volume: 26, Issue: 5, Oct. 2021), 20 April 2021, D.O.I. (10.26599/TST.2020.9010028).
5. Alexandre Gondeau, Zahia Aouabed, Mohamed Hijri, Pedro R. Peres-Neto, Vladimir Makarenkov, "Object Weighting: A New Clustering Approach to Deal with Outliers and

Cluster Overlap in Computational Biology", IEEE/ACM Transactions on Computational Biology and Bioinformatics (Volume: 18, Issue: 2, March-April 1 2021). 10 June 2019, D.O.I. (10.1109/TCBB.2019.2921577).

6. A. Jensen, Anders la Cour-Harbo, "Ripples in Mathematics: The Discrete Wavelet Transform", Springer Science Business Media, 2001.
7. A.K. Nagar and D. Sokhi, "On Wavelet-Based Adaptive Approach for Gene Comparison", Int'l J. Intelligent Systems Technologies and Applica- tions, vol. 5, pp. 104-114, 2008.
8. Saxena, S., Nair, A.M., Rizvi, A.Z., "Analysis of COVID-19 Genome Using Continuous Wavelet Transform," Networks and Systems, vol 662, no. 1, pp. 1047-1077, 2023.
9. 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.
10. C. Cattani, "Fractals and Hidden Symmetries in DNA," Math. Problems in Eng., vol. 2010, pp. 1- 32, 2010.
11. C. Cattani, "On the Existence of Wavelet Symme- tries in Archaea DNA," Computational and Math. Methods in Medicine, vol. 2012, pp. 1-21, 2012.
12. J.K. Meher, M.K. Raval, P.K. Meher, and G.N. Nash, "Wavelet Transform for Detection of Con- served 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.
13. 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.
14. R. Gupta, A. Mittal, K. Singh, P. Bajpai, and S. Prakash, "A Time Series Approach for Identifi- cation of Exons and Introns," Proc. 10th Int'l Conf. Information Technology (ICIT '07), pp. 91- 93, Dec. 2007
15. 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.
16. S. Deng, Z. Chen, G. Ding, and Y. Li, "Predic- tion 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.
17. 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.
18. 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
19. T.S. Gunawan and E. Ambikairajah, "Parallel Im- plementation of Genomic Sequences Classifica- tion Using Modified Gabor Wavelet Transform on Multicore Systems," Proc. Int'l Conf. Biomedical Eng. (ICoBE), pp. 165-168, Feb. 2012
20. 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.

21. J. Qiu, R. Liang, X. Zou, and J. Mo, "Prediction of Protein Secondary Structure Based on Contin- uous Wavelet Transform," Talanta, vol. 61, no. 1, pp. 285-293, Apr. 2003.
22. K. Chou, "Prediction of Protein Cellular Attributes Using Pseudo Amino Acid Composition," Pro- teins, vol. 43, no. 3, pp. 246-255, May 2001.
23. S. Chandra and A.Z. Rizvi, "Wavelet Analysis of Hiv-1 Genome," Proc. Int'l Assoc. Computer Sci- ence and Information Technology - Spring Conf. (IACSITSC '09), pp. 559-561, Apr. 2009.
24. D.T. Jones, M. Tress, K. Bryson, and C. Hadley, "Successful Recognition of Protein Folds Using Threading Methods Biased by Sequence Similar- ity and Predicted Secondary Structure," Proteins: Structure, Function, and Bioinformatics, vol. 37, no. S3, pp. 104-111, 1999.
25. C.H. Trad, Q. Fang, and I. Cosic, "Protein Se- quence Comparison Based on the Wavelet Trans- form Approach," Protein Eng., vol. 15, pp. 193- 203, Mar. 2002.
26. J. Qiu, S. Luo, J. Huang, and R. Liang, "Using Support Vector Machine for Prediction of Pro- tein Structural Classes Based on Discrete Wavelet Transform," J. Computation Chemistry, vol. 30, no. 8, pp. 1344-1350, June 2009.
27. H. Chen, F. Gu, and F. Liu, "Predicting Protein Secondary Structure Using Continuous Wavelet Transform and ChouFasman Method," Proc. IEEE Conf. Eng. in Medicine and Biology Soc., vol. 3, pp. 2603-2606, Mar. 2005.
28. T. Boveri, "Concerning the Origin of Malignant Tumours by Theodor Boveri," J. Cell Science, vol. 121, no. Supplement 1, pp. 1-84, 2008.
29. Lina Yang, Yuan Yan Tang, Yang Lu, Huiwu Luo, "A Fractal Dimension and Wavelet Transform Based Method for Protein Sequence Similarity Analysis", IEEE/ACM Transactions on Computa- tional Biology and Bioinformatics, Vol. 12, Issue, 2, 16 October 2014.
30. Yang Liu, Lina Yang, Yuan Yan Tang, Patrick Wang, "Comparison of RNA Secondary Structure by using Discrete Wavelet Transform and Fractal Dimension", Proceedings of the 2020 interna- tional Conference on Wavelet Analysis and Pattern Recognition, 4 Dec 2020.
31. P. Dutta, S. Basu, and M. Kundu, "Assessment of semantic similarity between proteins using infor- mation content and topological properties of the gene ontology graph," IEEE/ACM transactions on computational biology and bioinformatics, vol. 15, no. 3, pp. 839–849, 2018.
32. P. Maji, E. Shah, and S. Paul, "Relsim: An inte- grated method to identify disease genes using gene expression profiles and ppin based similar- ity measure," Information Sciences, vol. 384, pp. 110– 125, 2017.
33. A. Schlicker, F. S. Domingues, J. Rahnenfuhrer, and T. Lengauer, ¨ "A new measure for functional similarity of gene products based on gene ontol- ogy," BMC bioinformatics, vol. 7, no. 1, p. 302, 2006.
34. T. Xu, L. Du, and Y. Zhou, "Evaluation of go-based functional similarity measures using s. cerevisiae protein interaction and expression profile data," BMC bioinformatics, vol. 9, no. 1, p. 472, 2008.
35. R. Cao and J. Cheng, "Integrated protein function prediction by mining function associations, se- quences, and

protein–protein and gene–gene interaction networks," Methods, vol. 93, pp. 84–91, 2016.
36. B. Li, J. Z. Wang, F. A. Feltus, J. Zhou, and F. Luo, "Effectively integrating information con- tent and structural relationship to improve the go-based similarity measure between proteins," arXiv preprint arXiv:1001.0958, 2010.
37. J. Peng, Y. Wang, and J. Chen, "Towards inte- grative gene functional similarity measurement," in BMC bioinformatics, vol. 15, no. 2. BioMed Central, 2014, p. S5.
38. Z. Tian, M. Guo, C. Wang, L. Xing, L. Wang, and Y. Zhang, "Constructing an integrated gene sim- ilarity network for the identification of disease genes," Journal of biomedical semantics, vol. 8, no. 1, p. 32, 2017.
39. P. Baldi and G.W. Hatfield, DNA microarrays and gene expression: from experiments to data anal- ysis and modeling, Cambridge University Press, 2002.
40. B.N. Li et al, "Integrating spatial fuzzy clustering with level set methods for automated medical image segmentation," Comput. Biol. Med., vol. 41, pp. 1-10, 2011.
41. J.H. Young et al, "Computational discovery of pathway-level genetic vulnerabilities in non- small-cell lung cancer," Bioinformatics, vol. 32, pp. 1373–1379, 2016.
42. D. Tamborero et al, "OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes," Bioinformatics, vol. 29, pp. 2238-2244, 2013.
43. Y. Yang et al, "SAFE-clustering: single-cell aggre- gated (from ensemble) clustering for single-cell RNA-seq data," Bioinformatics, bty793, in press, 2018.
44. X. Hao et al, "Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering," Bioinformatics, vol. 27, pp. 611-618, 2011.
45. P. Jiang and M. Singh, "SPICi: a fast clustering algorithm for large biological networks," Bioin- formatics, vol. 26, pp. 1105-1111, 2010.
46. J. Choi et al, "Improved prediction of breast cancer outcome by identifying heterogeneous biomarkers," Bioinformatics, vol. 33, pp. 3619–3626, 2017.