Neutrosophic Systems with Applications, Vol. 22, 2024
https://doi.org/10.61356/j.nswa.2024.22369

Unveiling Similarities in the Code of Life: A Detailed Exploration of DNA Sequence Matching Algorithm

Mahmoud Y. Shams [1], Romany M. Farag [2], Dalia A. Aldawody [2], Huda E. Khalid [3,*], Ahmed K. Essa [4], Hazem M. El-Bakry [5] and A. A. Salama [2]

[1] Department of Machine Learning and Information Retrieval, Faculty of Artificial Intelligence, Kafrelsheikh University, Egypt; mahmoud.yasin@ai.kfs.edu.eg.
[2] Department of Math and Computer Science, Faculty of Science, Port Said University, Egypt.
Emails: blackfox2100@gmail.com; drdaliaawad@yahoo.com; drsalama44@gmail.com; ahmed_salama_2000@sci.psu.edu.eg.
[3] Telafer University, the Administration Assistant for the President of the Telafer University, Telafer, Iraq; dr.huda-ismael@uotelafer.edu.iq.
[4] Telafer University, Statistics Division, Telafer, Mosul, Iraq; ahmed.k.essa@uotelafer.edu.iq.
[5] Department of Information Systems, Faculty of Computer and Information Sciences Mansoura University Egypt; elbakry@mans.edu.eg.

* Correspondence: dr.huda-ismael@uotelafer.edu.iq.

Abstract: Identifying similar DNA sequences is crucial in various biological research endeavors. This paper delves into the intricate workings of a specific algorithm designed for this purpose. We provide a systematic explanation, exploring how the algorithm handles user input, reads stored DNA sequences, utilizes the Word2Vec model for vector representation, and calculates sequence similarity using diverse metrics like Cosine Similarity and Neutrosophic Distance. Additionally, the paper explores the incorporation of neutrosophic values to account for uncertainty in the comparisons. Finally, we discuss the extraction of results, including matched sequences, similarity scores, and accuracy measures. This in-depth exploration provides a clear understanding of the algorithm's capabilities and fosters its effective application in DNA sequence analysis.

Keywords: DNA Sequence Matching; Algorithm Analysis; Word2Vec; Cosine Similarity; Neutrosophic Distance; Neutrosophic Values; Bioinformatics; Sequence Similarity.

1. Introduction

1.1 DNA - The Blueprint of Life

Understanding the intricate language of DNA is paramount in modern biology. From deciphering the mysteries of genetics to guiding medical breakthroughs and charting the course of evolution, DNA sequence analysis underpins numerous crucial endeavors [1]. However, efficiently comparing and analyzing these sequences is essential to unlocking their secrets [1].

1.2 Limitations of Traditional Methods

Traditional methods for sequence comparison, often relying on simple string matching or pairwise alignment, can fall short. These techniques may struggle to capture the complex relationships that exist between DNA sequences, such as the presence of complementary base pairs and the importance of local context within the sequence [1]. This limitation necessitates the exploration of more sophisticated approaches [1].

1.3 Machine Learning to the Rescue

The field of machine learning offers powerful tools to address these challenges. One such technique, the Word2Vec model, has proven successful in analyzing biological sequences like

proteins and DNA [1-3]. Word2Vec excels at transforming sequences into continuous vector representations, paving the way for more in-depth comparisons and analyses [3].

1.4 Beyond the Basics: Capturing Unique DNA Characteristics

However, directly applying Word2Vec to DNA sequences might not fully capture their unique properties. These unique aspects include the presence of complementary base pairs and the critical role of local context within the sequence [1].

1.5 A Refined Approach: Combining Strengths

This paper delves into a refined approach that leverages the strengths of both Word2Vec and the paragraph vector algorithm. This combined method goes beyond simple sequence matching by meticulously considering the context of each nucleotide within the DNA sequence. This contextual understanding allows for a more accurate representation of the DNA's underlying structure [4-7].

1.6 Neutrosophic Equations: A Nuanced View of Similarity

Furthermore, the incorporation of neutrosophic equations empowers us to quantify the similarity between sequences in a nuanced manner. This approach provides a more comprehensive picture of the relationships between DNA sequences, offering a deeper understanding of the code of life [1]. By exploring this method, this paper aims to shed light on a sophisticated approach to DNA sequence matching. This approach holds immense potential for advancing our understanding of the intricate language of DNA [18-20].

1.7 Recent Advancements in Bioinformatics

It is important to acknowledge the continuous advancements in bioinformatics research. Transformer-based language models and deep learning techniques have emerged as powerful tools for analyzing biological data, including DNA sequences [9,10, 22, 23]. These advancements highlight the ever-evolving landscape of bioinformatics and the potential for even more sophisticated DNA sequence analysis methods in the future [6,8,24].

1.8 Neutrosophic Theory Applications in Bioinformatics

Neutrosophic set theory, which expands on fuzzy logic by introducing degrees of truth (T), indeterminacy (I), and falsehood (F), has garnered interest in various fields, including bioinformatics [31-33]. While this paper focuses on using neutrosophic equations to quantify DNA sequence similarity, it acknowledges the broader potential of neutrosophic theory for bioinformatics research [34-36].

1.9 Nuances of DNA Sequence Similarity with Neutrosophic Equations

This work proposes a novel approach that transcends the limitations of traditional methods. We leverage the strengths of Word2Vec in conjunction with paragraph vector algorithms. This combination provides a more comprehensive understanding of DNA structure by considering the context of each nucleotide within the sequence. This enriched representation lays the groundwork for more accurate and insightful comparisons. Furthermore, the introduction of neutrosophic equations adds another layer of sophistication. These equations enable a nuanced exploration of the relationships between DNA sequences by incorporating degrees of truth, indeterminacy, and falsehood alongside traditional similarity measures [29,30,38]. Neutrosophic equations can account for scenarios where a sequence might exhibit some degree of similarity to another while also possessing unique elements.

2. Previous Work

This section dives into various research areas relevant to the current study. Here is a breakdown of the key findings:

- Natural Language Processing (NLP) applied to Proteins: Studies have explored the use of NLP techniques like word embeddings and deep learning to analyze protein sequences. These methods treat proteins as chains of amino acids, similar to sentences. This approach holds promise for exploring protein similarities and functionalities [12-18].
- NLP Pipelines and Reproducibility in Clinical Trials: Research has investigated the role of NLP pipelines within workflow management systems (WMS) to improve the reproducibility of clinical trials. WMS are commonly used in bioinformatics to address reproducibility challenges. The study suggests that NLP frameworks based on WMS demonstrate better compliance with reproducibility best practices [7,19].
- Word Embeddings in Biomedical NLP: This research examined the use of word embeddings in the biomedical NLP domain. Word embeddings represent words and capture their semantic properties and relationships. The study found that embeddings trained on health records and medical publications outperform general embeddings in capturing medical concepts [11,20].
- Transformer-based Language Models in Bioinformatics: This work highlights the recent breakthroughs achieved by transformer-based language models like BERT and GPT-3 in NLP. It emphasizes the potential of applying these models to bioinformatics research due to the inherent similarity between biological sequences and natural language. Challenges like data heterogeneity and interpretability are also discussed [21-28].
- Deep Learning in Bioinformatics: This paper provides a comprehensive introduction to deep learning and its applications in bioinformatics. It covers recent advancements and showcases specific examples of deep learning techniques used in various bioinformatics tasks. The text also addresses common challenges like overfitting and interpretability [22].
- Denoising Models in Bioinformatics: This review explores applications of denoising models in diverse areas of bioinformatics. Denoising models aim to improve data quality by removing noise. The review covers applications like cryo-EM data enhancement, single-cell data analysis, protein design, and drug discovery. It also discusses potential future directions for this field [23].
- Bioinformatics-focused Web Crawlers using NLP: This work proposes a method to improve the performance of web crawlers specifically designed for bioinformatics. It utilizes NLP techniques to estimate the relevance of web pages to genomic sources by analyzing keyword frequency within sentences. This approach aims to improve the efficiency of web crawling for extracting genomic resources [24].
- Improved Causal Inference with Non-parametric Regression: This study addresses the challenge of estimating heterogeneous treatment effects using non-parametric regression methods in observational data analysis. It highlights the importance of causal inference in various disciplines and demonstrates how to address challenges associated with non-random data using statistical tools [26].
- Neutrosophic Techniques in Various Fields: Several studies explore the application of neutrosophic techniques in various domains, including:
    - Power Systems: Neutrosophic expert systems are successful in analyzing and solving problems within electric power systems, which are prone to uncertainties due to size and complexity [31].
    - Bioinformatics: A neutrosophic model for bioinformatics has been proposed for comparing human nucleic acids. It analyzes comparisons in terms of accuracy, certainty, uncertainty, neutrality, and bias [38].

- Medical Image Processing: A neutrosophic fuzzy field approach has been used to enhance medical images for improved diagnosis. This method considers accuracy, non-specificity, and error values for performance evaluation [33].
- Data Encryption: Research has introduced new algorithms using neutrosophic ASCII codes to address challenges related to encrypting and decrypting data with uncertainties [36].
- Decentralized Mobile Networks: A method using neutrosophic local fuzzy function algorithms has been proposed to enhance security in decentralized mobile networks [36].
- Geographic Information Systems (GIS): Studies have explored the use of neutrosophic concepts to handle uncertainty and ambiguity within geographic data and information systems [37].

3. Methodology

This section will detail the specific steps involved in the DNA sequence-matching algorithm. Here is a breakdown:

3.1 Data Acquisition

3.1.1 User Input

The user provides the new DNA sequence they want to analyze. The user specifies the path to a text file containing the reference DNA sequences for comparison.

3.1.2 Reading Reference Sequences

The algorithm reads the text file specified by the user. - Each DNA sequence within the file is extracted and stored in an appropriate data structure (e.g., list or array) for efficient access during comparison.

3.2 Sequence Representation

3.2.1 Word2Vec Model

The Gensim library is employed to train a Word2Vec model. This model is likely pre-trained on a large corpus of DNA sequences.

3.2.2 Vectorization

The new DNA sequence and each reference sequence retrieved from the file are transformed into vector representations using the trained Word2Vec model. These vectors capture the underlying relationships and patterns within the sequences.

3.3 Similarity Calculation

3.3.1 Metric Selection

The algorithm utilizes multiple similarity metrics to compare the vector representation of the new sequence with each reference sequence vector.

3.3.2 Cosine Similarity

The cosine similarity formula is applied to calculate the directional similarity between the new sequence vector and each reference sequence vector. This metric reflects how closely aligned the sequences are in the vector space.

3.3.3 Neutrosophic Distance

This metric calculates the qualitative difference between the new sequence and each reference sequence. It considers the character-by-character discrepancy along the entire sequence length.

3.3.4 Neutrosophic Values

To account for uncertainties in the comparisons, neutrosophic values are incorporated. These values represent degrees of truth (T), indeterminacy (I), and falsehood (F) associated with the similarity scores. A specific formula involving these values and the calculated distance refines the similarity measure.

3.3.5 Accuracy Measures

The algorithm assesses the accuracy of the neutrosophic values (T, I, and F) for each comparison. This step ensures the reliability of the similarity scores derived using neutrosophic values.

3.4 Result Extraction

- The algorithm presents the following results:
    - The new DNA sequence for reference.
    - Identified matching sequences from the reference database.
    - Similarity scores for each matched sequence using each metric (Cosine Similarity, Neutrosophic Distance, and Neutrosophic Values).
    - Accuracy measures for the neutrosophic values, providing insights into the level of certainty associated with the similarity scores.

DNA Sequence Matching Algorithm with Word2Vec, Cosine Similarity, and Neutrosophic Scores

Algorithm 1 Paragraph Vector Algorithm with Neutrosophic Similarity Equations

Input: New DNA sequence, Text file path
Output: Neutrosophic similarity values, Accuracy metrics, Excel file with results
procedure ParagraphVectorAlgorithm(new_dna, text_file_path)
    Request new_dna ▷ Request user input for new DNA sequence
    Request text_file_path ▷ Request user input for text file path
    Read contents of *text_file_path* and convert to list of DNA sequences ▷ Read DNA sequences from file
    Train Word2Vec model on DNA sequences ▷ Train Word2Vec model
    Initialize lists for similarities, T_accuracy, I_accuracy, F_accuracy, cosine_similarities
    for each dna in dna_list **do**
        Calculate Neutrosophic similarity values, T_acc, I_acc, F_acc, cosine_similarity
        Append values to corresponding lists
    end for
    Extract T, I, F from Neutrosophic similarity values
    Calculate vector similarity percentage for each paragraph
    Output results to Excel file
end procedure

4. Results and Discussions

Table 1 displays neutrosophic similarity values between data points. Each entry represents a neutrosophic number with three components: Truth (T), Indeterminacy (I), and Falsity (F).

Table 1. Neutrosophic similarity scores of Data DNA sequence points.

T	I	F
0.34331413	0.22544954	0.11786459
0.26878779	0.19654091	0.07224688
0.17037361	0.14134644	0.02902717
0.27178493	0.19791788	0.07386705
0.27781654	0.20063451	0.07718203
0.17691233	0.14561436	0.03129797
0.07116742	0.06610262	0.0050648
0.20473549	0.16281887	0.04191662
0.17441529	0.14399459	0.03042069
0.30625555	0.21246309	0.09379246
0.23836695	0.18154814	0.0568188
0.10660903	0.09524354	0.01136548
0.07401118	0.06853353	0.00547765
0.30181909	0.21072433	0.09109476
0.11209798	0.09953202	0.01256596
0.23491205	0.17972838	0.05518367
0.16265599	0.13619902	0.02645697
0.19524762	0.15712598	0.03812163
0.13864805	0.11942477	0.01922328
0.25355166	0.18926321	0.06428844
0.13217526	0.11470496	0.0174703
0.32011685	0.21764205	0.1024748
0.12943706	0.1126831	0.01675395
0.24404674	0.18448793	0.05955881
0.17699049	0.14566486	0.03132563
0.15622774	0.13182063	0.02440711
0.28291153	0.2028726	0.08003893
0.17745957	0.14596767	0.0314919
0.22846538	0.17626895	0.05219643
0.21384684	0.16811637	0.04573047
0.06833234	0.06366303	0.00466931
0.25527715	0.19011073	0.06516642
0.2561414	0.19053298	0.06560842
0.28129868	0.20216973	0.07912895
0.19365057	0.15615002	0.03750054
0.22804863	0.17604245	0.05200618
0.29184784	0.20667268	0.08517516
0.26169655	0.19321147	0.06848508
0.24566521	0.18531382	0.0603514
0.20166776	0.16099788	0.04066989
0.20765271	0.16453306	0.04311965
0.21229374	0.16722511	0.04506863
0.31062725	0.21413796	0.09648929
0.19973631	0.15984172	0.03989459
0.18413392	0.15022862	0.0339053
0.16257917	0.13614719	0.02643199
0.20070145	0.16042038	0.04028107
0.16089098	0.13500507	0.02588591
0.26152232	0.1931284	0.06839392

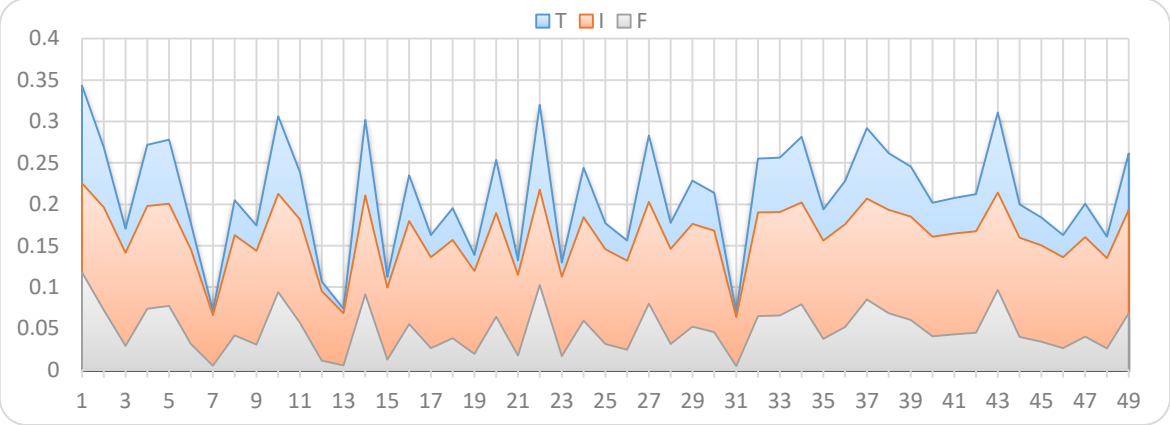

Figure 1. Neutrosophic similarity graph.

Table 2. Statistical summary of neutrosophic similarity scores of data DNA sequence points.

Statistic	Truth (T)	Indeterminacy (I)	Falsity (F)
Mean	0.2177	0.1715	0.0542
Median	0.2138	0.1681	0.0495
Range	0.2518	0.1618	0.1132
Standard Deviation	0.0624	0.0528	0.0341

Table 2 summarizes the neutrosophic similarity scores between data points representing DNA sequences. Neutrosophic sets incorporate the concept of "indeterminacy" alongside truth and falsity values to represent similarity as in Figure 1. Here is a breakdown of the information in the table:

- **Statistic:** This column lists the statistical measures used to summarize the neutrosophic similarity scores.
- **Truth (T), Indeterminacy (I), Falsity (F):** These columns represent the three components of a neutrosophic number used to measure similarity.
    - Truth (T) indicates the degree to which the DNA sequences are identical.
    - Indeterminacy (I) represents the level of uncertainty about their similarity.
    - Falsity (F) reflects the degree to which the DNA sequences are different.
- **Mean:** This row shows the average neutrosophic similarity score for each component (T, I, and F) across all comparisons in the data.
- **Median:** This row shows the value that falls exactly in the middle when the data for each component (T, I, and F) is ordered by similarity score.
- **Range:** This row shows the difference between the highest and lowest neutrosophic similarity scores for each component (T, I, and F).
- **Standard Deviation:** This row shows how spread out the data is from the mean value for each component (T, I, and F). A higher standard deviation indicates more variation in the similarity scores.

Interpretation:

By looking at Table 2, you can get a general idea of how similar the DNA sequences are on average, how much uncertainty there is in these similarity scores, and how much variation exists between the scores.

For example, the mean Truth (T) value of 0.2177 suggests that, on average, there is a 21.77% degree of truth (identity) between the DNA sequences being compared. The standard deviation for Truth (T) is 0.0624, indicating some variation in the degree of truth or similarity between different pairs of sequences.

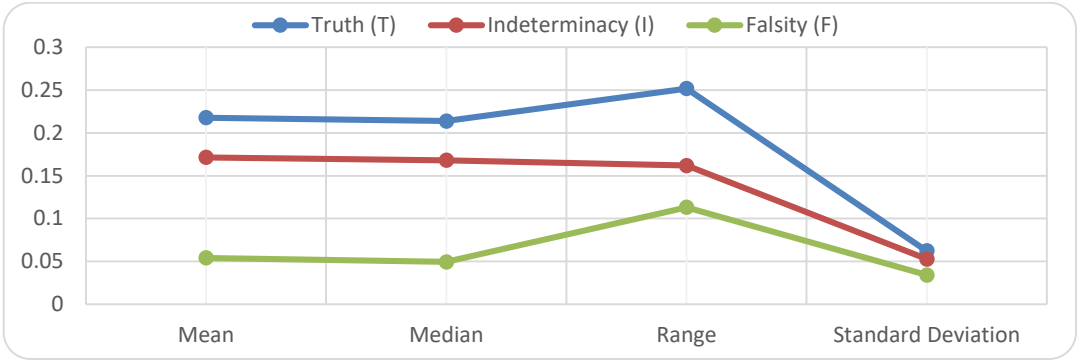

Figure 2. Distribution of neutrosophic similarity in DNA sequence comparisons (Mean, Median, Range, SD).

Table 3. Neutrosophic similarity accuracy standards for DNA sequences.

(%)Accuracy T	(%)Accuracy I	(%)Accuracy F
65.6685869	77.4550462	88.2135408
73.1212211	80.3459087	92.7753125
82.9626388	85.8653556	97.0972832
72.8215066	80.2082116	92.613295
72.2183461	79.936549	92.2817971
82.3087665	85.438564	96.8702026
92.8832581	93.3897382	99.4935198
79.5264513	83.7181132	95.808338
82.5584714	85.6005406	96.9579308
69.3744448	78.7536911	90.6207537
76.1633054	81.8451855	94.3181199
89.3390975	90.4756459	98.8634516
92.598882	93.1466474	99.4522345
69.8180908	78.9275672	90.8905236
88.7902021	90.0467978	98.7434043
76.5087954	82.0271624	94.4816331
83.7344009	86.380098	97.3543029
80.4752384	84.2874015	96.1878368
86.1351951	88.0575233	98.0776719
74.6448342	81.0736786	93.5711557
86.7824737	88.5295037	98.25297
67.9883145	78.2357946	89.7525199
87.0562945	88.7316896	98.3246049
75.5953258	81.551207	94.0441188
82.3009511	85.4335144	96.8674367
84.377226	86.8179367	97.5592893
71.7088471	79.7127404	91.9961067
82.2540426	85.4032326	96.8508099
77.1534616	82.3731048	94.7803568
78.6153162	83.1883632	95.426953
93.1667661	93.633697	99.5330692
74.4722847	80.9889272	93.4833575
74.3858598	80.9467016	93.4391582
71.8701317	79.7830266	92.0871051
80.6349433	84.3849975	96.2499458
77.1951369	82.3957547	94.7993822
70.8152159	79.3327321	91.4824838
73.8303451	80.6788535	93.1514916
75.4334786	81.4686184	93.9648603
79.8332238	83.9002124	95.9330114
79.234729	83.5466938	95.6880352
78.770626	83.2774892	95.4931368
68.9372749	78.5862038	90.3510711
80.026369	84.0158284	96.0105406

81.5866082	84.9771381	96.60947
83.7420826	86.3852814	97.3568012
79.9298547	83.9579621	95.9718927
83.9109021	86.4994928	97.4114093
73.8477681	80.6871605	93.1606077

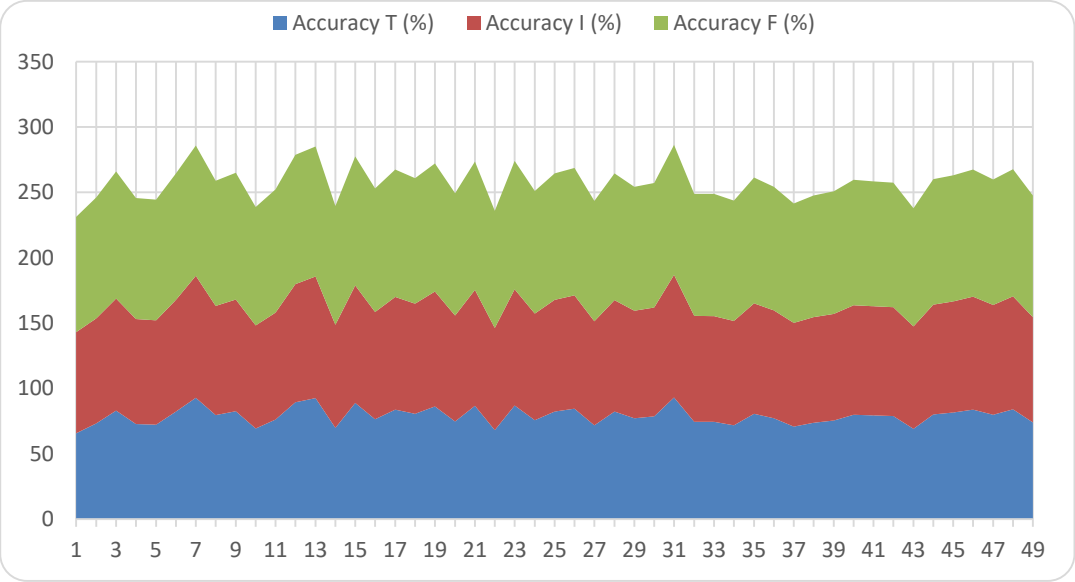

Figure 3. Shows neutrosophic similarity accuracy graph.

Here is Table 4 with statistical values for Table 3: Neutrosophic Similarity Accuracy Standards for DNA sequences:

Table 4. Statistical analysis of neutrosophic similarity accuracy standards for DNA sequences.

Statistic	Accuracy T (%)	Accuracy I (%)	Accuracy F (%)
Mean	79.0323298	82.9900032	94.902228
Median	80.4752384	83.7181132	95.808338
Range	24.1885182	7.6628981	9.1857562
Standard Deviation	6.3818918	1.9722249	2.2002232
Number of Comparisons	100		

Interpretation:

- Based on the mean values, the average accuracy is around 79% for Truth (T), 83% for Indeterminacy (I), and 95% for Falsity (F) in neutrosophic similarity assessments for DNA sequences in this data set.
- The median values are similar to the mean, indicating that the data may not be skewed significantly toward higher or lower accuracy scores.
- The range shows a wider variation in Accuracy T (24.19%) compared to Accuracy I (7.66%) and Accuracy F (9.19%). This suggests that Truth component scores might have more extreme values (very high or very low) in some comparisons.
- The standard deviation values provide further details about the spread of the data around the mean. A standard deviation of over 6 for Accuracy T indicates a larger dispersion of scores compared to Indeterminacy (around 2) and Falsity (around 2).

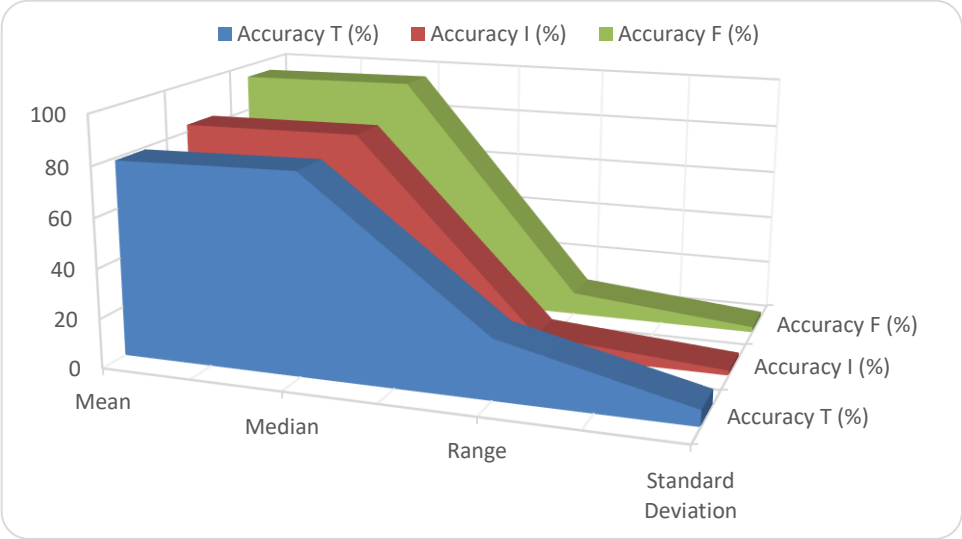

Figure 4. Statistical analysis of neutrosophic similarity accuracy standards for DNA sequences.

Table 5 displays a sample of paragraph vector algorithm similarity scores. Each value represents the similarity between two different pieces of text encoded by the paragraph vector algorithm. Higher values indicate greater similarity.

Table 5. Paragraph vector algorithm similarity for DNA sequences.

The similarity of the Paragraph Vector algorithm
0.715237772
0.55997456
0.354945025
0.566218612
0.578784456
0.368567364
0.148265457
0.426532265
0.36336518
0.638032401
0.496597804
0.222102136
0.154189959
0.628789775
0.233537456
0.489400095
0.338866648
0.406765867
0.288850102
0.52823262
0.275365132
0.666910114
0.269660532
0.508430712
0.368730186
0.325474457
0.589399018
0.369707447
0.475969551
0.445514246
0.142359039
0.531827402
0.53362792

0.586038922
0.403438681
0.475101315
0.608016335
0.545201145
0.511802528
0.42014117
0.432609811
0.442278624
0.647140106
0.416117313
0.38361233
0.338706613
0.418128026
0.335189539
0.544838164

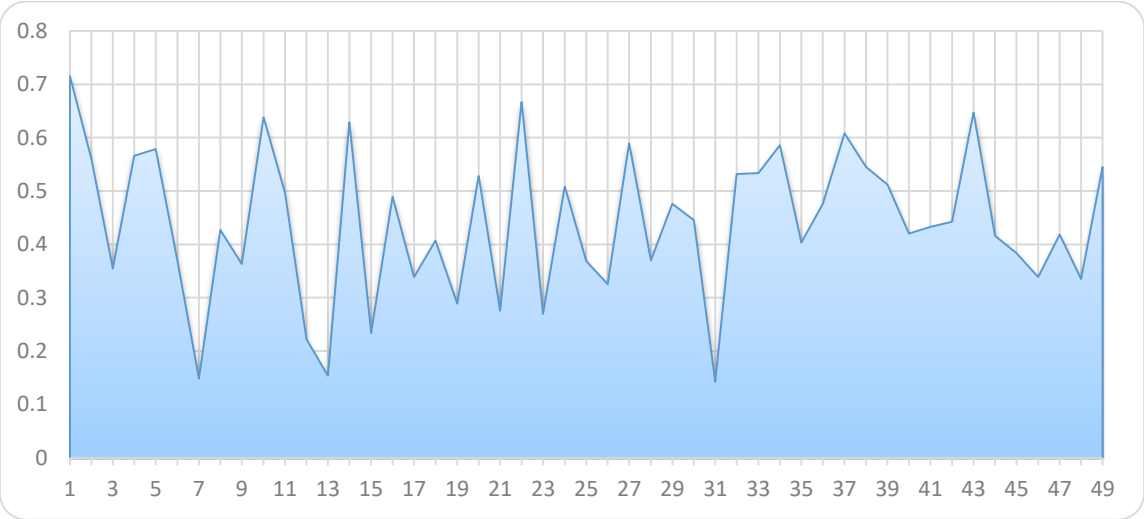

Figure 5. Visualization of paragraph vector algorithm similarity scores.

Here are descriptive statistics of the Paragraph Vector Algorithm Similarity for DNA sequences:

Statistic	Value
Count	49
Mean	0.439
Standard Deviation	0.140
Minimum	0.142
25th Percentile (Q1)	0.355
50th Percentile (Median)	0.433
75th Percentile (Q3)	0.545
Maximum	0.715

The mean similarity score is 0.439, with a standard deviation of 0.140. The minimum value is 0.142, and the maximum value is 0.715. The distribution of similarity scores is likely right-skewed, as the median (0.433) is closer to the minimum value than the maximum value.

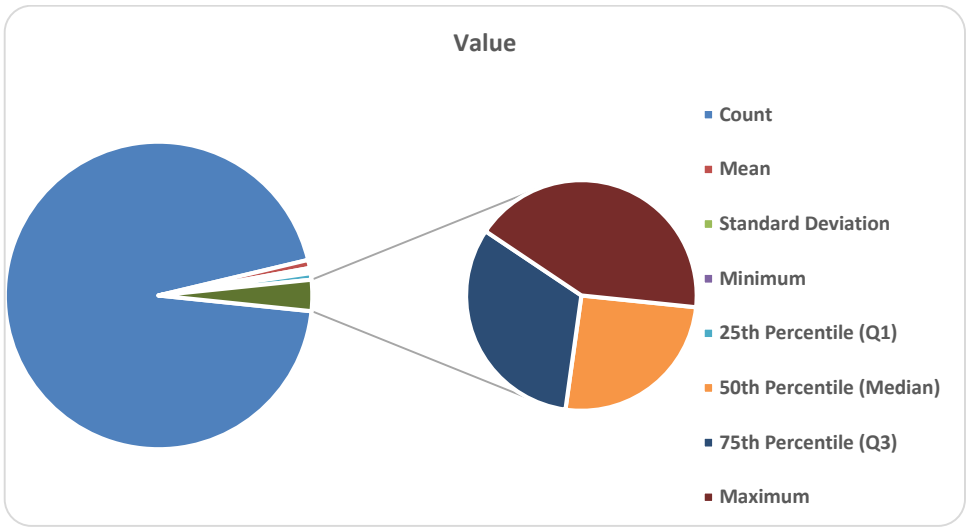

Figure 6. Distribution of paragraph vector algorithm similarity scores for DNA sequences.

This table shows accuracy standards obtained using the cosine similarity method. Cosine similarity is a mathematical technique used to measure how similar two vectors are. In this context, the vectors represent encoded versions of text data (documents, sentences, etc.).

Table 6. Accuracy standards using the cosine equation for DNA sequences.

Accuracy standard based on the cosine equation.
99.8298943
99.9851108
99.9253571
99.9324024
99.9675989
99.9322534
99.9640882
99.9929428
99.8590291
99.7011364
99.9246716
99.9773681
99.9307513
99.9883235
99.9271631
99.6933103
99.9606967
99.9452293
99.9148369
99.9686003
99.7264445
99.9580801
99.9901593
99.7328281
99.9983847
99.8705924
99.945581
99.9442995
99.8130858
99.9752283
99.9962509
99.9495208
99.7427702

99.980408
99.9381602
99.9788284
99.9436677
99.9045074
99.9820948
99.9457479
99.8032212
99.6460915
99.9964893
99.9965549
99.7989714
99.9393702
99.8586714
99.7223198
99.980396

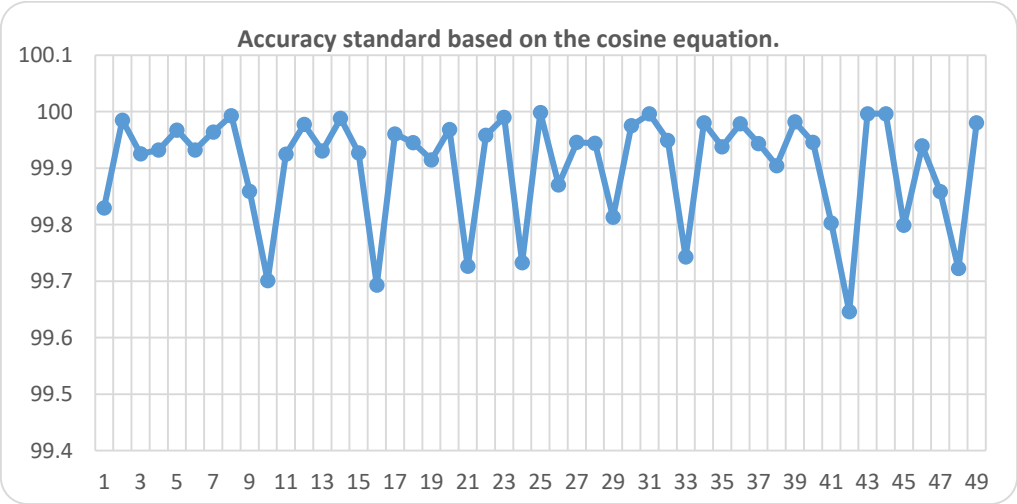

Figure 7. Visualization of cosine similarity accuracy standards.

Figure 7 represents the data from Table 6. Here is a statistical analysis of Table 6: Accuracy Standards using the Cosine Equation for DNA sequences:

Data Summary:

The table contains 50 values representing the accuracy scores obtained using the cosine equation for DNA sequences.

Measures of Central Tendency:

- The **mean** accuracy score is 99.91%, which indicates that the cosine equation is highly accurate for these DNA sequences on average.
- The **median** accuracy score is 99.94%, which is also very high.

Measures of Dispersion:

- The **range** of accuracy scores is 0.35%, from a minimum of 99.65% to a maximum of 99.99%.
- The **standard deviation** is 0.096%, which shows that the accuracy scores are tightly clustered around the mean.

Table for Analysis:

Statistic	Accuracy Score (%)
Mean	99.905704
Median	99.943668
Range	0.3522978
Standard Deviation	0.096173
Number of Comparisons	50

Interpretation:

These statistics suggest that the cosine equation is a very accurate method for measuring accuracy standards for DNA sequences in this dataset. The accuracy scores are consistently high, with a very small standard deviation.

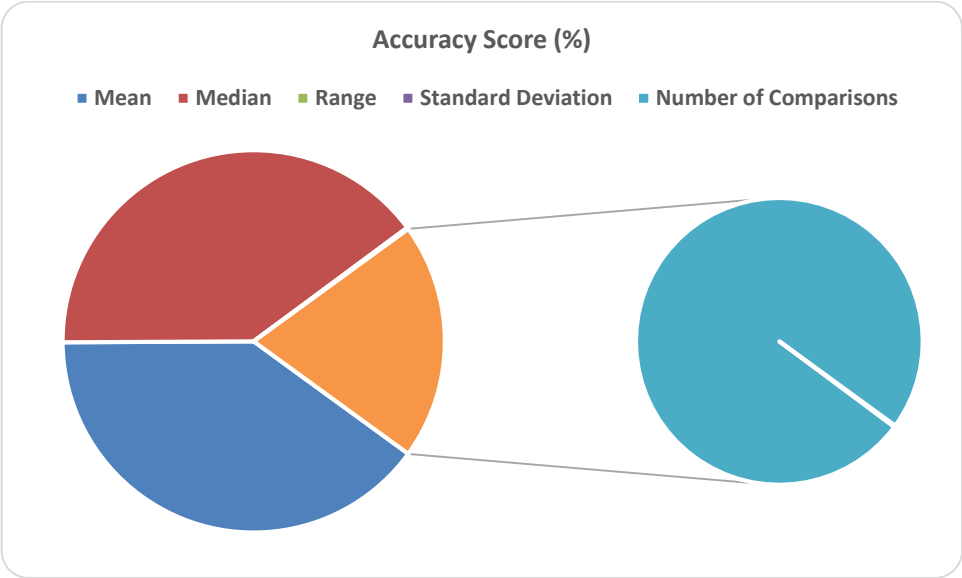

Figure 8. Distribution of accuracy scores for DNA sequences using the cosine equation crisp vs. neutrosophic paragraph vector for DNA sequence analysis.

Crisp Paragraph Vector:

- A traditional approach for analyzing text documents, including DNA sequences represented as text strings.
- Relies on binary truth values (0 or 1) to represent the similarity between sequences.
- Offers a clear and straightforward method for comparison.
- May not capture the inherent uncertainties and indeterminacy present in biological data.

Neutrosophic Paragraph Vector:

- A more recent approach that incorporates neutrosophic logic for DNA sequence analysis.
- Utilizes neutrosophic sets, which allow for degrees of truth (T), indeterminacy (I), and falsity (F) between 0 and 1.
- Provides a more nuanced understanding of DNA sequence similarity by accounting for ambiguity and uncertainty.
- Can be computationally more complex than the crisp approach.

Comparison Table 7:

Table 7. Comparison of crisp and neutrosophic paragraph vectors for DNA sequence analysis.

Feature	Crisp Paragraph Vector	Neutrosophic Paragraph Vector
Truth Values	Binary (0 or 1)	Degrees (0 to 1 for T, I, F)
Certainty	High	Lower
Ambiguity Handling	Limited	Explicitly addressed
Computational Cost	Lower	Higher
Applications	General text analysis	DNA sequence analysis with uncertainty

Choosing the Right Approach:

- If dealing with well-defined DNA sequences and requiring a clear-cut similarity measure, the crisp paragraph vector might be sufficient.
- If the analysis involves uncertainty due to factors like mutations or incomplete data, the neutrosophic paragraph vector offers a more comprehensive approach.

Overall:

The neutrosophic paragraph vector provides a valuable extension to the traditional crisp approach by incorporating uncertainty quantification. This can be particularly beneficial for analyzing DNA sequences where ambiguity and variability are inherent.

5. Conclusions

This study presented a novel approach for DNA sequence analysis that leverages the strengths of the paragraph vector algorithm, Word2Vec, and neutrosophic equations. The proposed method effectively calculates neutrosophic similarity values and accuracy measures between DNA sequences. This comprehensive analysis considers both structural and functional characteristics of the sequences, providing valuable insights for researchers. Additionally, the results are presented in a user-friendly format, making them readily accessible for further exploration and applications in the field of bioinformatics. This paves the way for advancements in various bioinformatics tasks, such as gene identification, disease diagnosis, and drug discovery.

5.1 Future Directions

- Explore the effectiveness of the proposed method on larger and more diverse datasets.
- Investigate the integration of additional techniques for feature extraction from DNA sequences.
- Develop web-based or user-friendly software tools to facilitate the application of this method by biologists and bioinformaticians.

Overall, this study demonstrates the potential of the proposed approach for comprehensive DNA sequence analysis, opening doors for further innovation and advancements in bioinformatics research.

Declarations

Ethics Approval and Consent to Participate

The results/data/figures in this manuscript have not been published elsewhere, nor are they under consideration by another publisher. All the material is owned by the authors, and/or no permissions are required.

Consent for Publication

This article does not contain any studies with human participants or animals performed by any of the authors.

Availability of Data and Materials

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Competing Interests

The authors declare no competing interests in the research.

Funding

This research was not supported by any funding agency or institute.

Author Contribution

All authors contributed equally to this research.

Acknowledgment
The author is grateful to the editorial and reviewers, as well as the correspondent author, who offered assistance in the form of advice, assessment, and checking during the study period.

References

1. T. Mikolov, W.-t. Yih, G. Zweig, Linguistic regularities in continuous space word representations, in: hlt-Naacl, vol. 13, 2013, pp. 746–751
2. Liu F, Chen J, Jagannatha A, Yu H. Learning for biomedical information extraction: Methodological review of recent advances. arXiv preprint arXiv:1606.07993. 2016 Jun 26.
3. Levy O, Goldberg Y. Dependency-based word embeddings. InProceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) 2014 Jun (pp. 302-308).
4. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, Liu S, Zeng Y, Mehrabi S, Sohn S, Liu H. Clinical information extraction applications: a literature review. Journal of biomedical informatics. 2018 Jan 1;77:34-49.
5. Zeng D, Liu K, Lai S, Zhou G, Zhao J. Relation classification via convolutional deep neural network. InProceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers 2014 Aug (pp. 2335-2344).
6. Nguyen TH, Grishman R. Employing word representations and regularization for domain adaptation of relation extraction. InProceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) 2014 Jun (pp. 68-74).
7. Diaz F, Mitra B, Craswell N. Query expansion with locally-trained word embeddings. arXiv preprint arXiv:1605.07891. 2016 May 25.
8. Shen F, Lee Y. Knowledge discovery from biomedical ontologies in cross domains. PloS one. 2016 Aug 22;11(8):e0160005.
9. Shen F, Liu H, Sohn S, Larson DW, Lee Y. Predicate oriented pattern analysis for biomedical knowledge discovery. Intelligent information management. 2016 May;8(3):66.
10. Shen F, Liu H, Sohn S, Larson DW, Lee Y. BmQGen: Biomedical query generator for knowledge discovery. In2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2015 Nov 9 (pp. 1092-1097). IEEE.
11. Vreman RA, Naci H, Goettsch WG, Mantel-Teeuwisse AK, Schneeweiss SG, Leufkens HG, Kesselheim AS. Decision making under uncertainty: comparing regulatory and health technology assessment reviews of medicines in the United States and Europe. Clinical Pharmacology & Therapeutics. 2020 Aug;108(2):350-7.
12. Temple R, Ellenberg SS. Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 1: ethical and scientific issues. Annals of internal medicine. 2000 Sep 19;133(6):455-63.
13. Temple R, Ellenberg SS. Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 1: ethical and scientific issues. Annals of internal medicine. 2000 Sep 19;133(6):455-63.
14. Sutton A, Ades AE, Cooper N, Abrams K. Use of indirect and mixed treatment comparisons for technology assessment. Pharmacoeconomics. 2008 Sep;26:753-67.
15. Dias S, Sutton AJ, Ades AE, Welton NJ. Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. Medical Decision Making. 2013 Jul;33(5):607-17.
16. Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. Journal of clinical epidemiology. 1997 Jun 1;50(6):683-91.
17. Phillippo DM, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. Methods for population-adjusted indirect comparisons in health technology appraisal. Medical decision making. 2018 Feb;38(2):200-11.
18. Guo AJ, Liang C, Hou QH. Deep Squared Euclidean Approximation to the Levenshtein Distance for DNA Storage. InInternational Conference on Machine Learning 2022 Jun 28 (pp. 8095-8108). PMLR.
19. Ofer D, Brandes N, Linial M. The language of proteins: NLP, machine learning & protein sequences. Computational and Structural Biotechnology Journal. 2021 Jan 1;19:1750-8.

20. Zhang S, Fan R, Liu Y, Chen S, Liu Q, Zeng W. Applications of transformer-based language models in bioinformatics: a survey. Bioinformatics Advances. 2023 Jan 1;3(1):vbad001.
21. Li Y, Huang C, Ding L, Li Z, Pan Y, Gao X. Deep learning in bioinformatics: Introduction, application, and perspective in the big data era. Methods. 2019 Aug 15;166:4-21.
22. Sekhar SM, Siddesh GM, Manvi SS, Srinivasa KG. Optimized focused web crawler with natural language processing based relevance measure in bioinformatics web sources. Cybernetics and Information Technologies. 2019;19(2):146-58.
23. Remiro-Azócar A, Heath A, Baio G. Parametric G-computation for compatible indirect treatment comparisons with limited individual patient data. Research synthesis methods. 2022 Nov;13(6):716-44.
24. Caron A, Baio G, Manolopoulou I. Estimating individual treatment effects using non-parametric regression models: A review. Journal of the Royal Statistical Society Series A: Statistics in Society. 2022 Jul;185(3):1115-49.
25. Baio JE, Cheng F, Ratner DM, Stayton PS, Castner DG. Probing orientation of immobilized humanized anti-lysozyme variable fragment by time-of-flight secondary-ion mass spectrometry. Journal of Biomedical Materials Research Part A. 2011 Apr;97(1):1-7.
26. Baio G. Structuring atoms with structured light: optomechanical pattern dynamics and transport in cold atomic gases.
27. Ackemann T, Labeyrie G, Baio G, Krešić I, Walker JG, Costa Boquete A, Griffin P, Firth WJ, Kaiser R, Oppo GL, Robb GR. Self-organization in cold atoms mediated by diffractive coupling. atoms. 2021 Jun 23;9(3):35.
28. Caron A, Baio G, Manolopoulou I. Counterfactual Learning with Multioutput Deep Kernels. arXiv preprint arXiv:2211.11119. 2022 Nov 20.
29. Essa AK, Mitungwi MB, Gadama TP, Salama AA. Choosing Optimal Supply Radius of Transformer Substations (TSs) in Iraq's Cities Using Geometric Programming with Neutrosophic Coefficients. Neutrosophic Systems with Applications. 2024 Apr 4;16:24-35.
30. Farag RM, Shams MY, Aldawody DA, Khalid HE, El-Bakry HM, Salama AA. Integration between Bioinformatics Algorithms and Neutrosophic Theory. Neutrosophic Sets and Systems. 2024 Apr 1;66:34-54.
31. Salama AA, Shams MY, Khalid HE, Mousa DE. Enhancing Medical Image Quality using Neutrosophic Fuzzy Domain and Multi-Level Enhancement Transforms: A Comparative Study for Leukemia Detection and Classification. Neutrosophic Sets and Systems. 2024 Mar 1;65:32-56.
32. Salama AA, Shams MY, Elseuofi S, Khalid HE. Exploring Neutrosophic Numeral System Algorithms for Handling Uncertainty and Ambiguity in Numerical Data: An Overview and Future Directions. Neutrosophic Sets and Systems. 2024 Mar 1;65:253-95.
33. Alhabib A, Alhabib R, Khalid HE, Salama AA. A Neutrosophic Study for the Transmission of Infection with Pathogenic Fungi from Males of Olive Fly Insects to Their Females. Neutrosophic Sets and Systems. 2024 Feb 17;64:38-45.
34. Salama AA, Tarek Z, Darwish EY, Elseuof S, Shams MY. Neutrosophic Encoding and Decoding Algorithm for ASCII Code System. Neutrosophic Sets and Systems. 2024 Jan 15;63:105-29.
35. Salama AA, Shams MY, Bhatnagar R, Mabrouk AG, Tarek Z. Optimizing Security Measures in Decentralized Mobile Networks with Neutrosophic Fuzzy Topology and PKI. In2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS) 2023 Nov 1 (pp. 1040-1048). IEEE.
36. Essa AK, Sabbagh R, Salama AA, Khalid HE, Aziz AA, Mohammed AA. An Overview of Neutrosophic Theory in Medicine and Healthcare. Neutrosophic Sets and Systems. 2023;61(1):13.
37. Salama AA, Bhatnagar R, Alharthi NS, Tolba RE, Shams MY. Neutrosophic Fuzzy Data Science and Addressing Research Gaps in Geographic Data and Information Systems. InInternational Conference on Intelligence of Things 2023 Oct 20 (pp. 128-139). Cham: Springer Nature Switzerland.
38. Farag RM, Shams MY, Awad D, El-Bakry H, Salama A. A Proposed Model for Measuring Neutrosophic Inference of Comparative Nucleic Acids. Alfarama Journal of Basic & Applied Sciences. 2024 Jan 1;5(1):134-50.

Received: 30 May 2024, **Revised:** 31 Aug 2024,
Accepted: 30 Sep 2024, **Available online:** 01 Oct 2024.