Review

From Nucleotides to Numbers: A Comprehensive Review of RNA Feature Extraction Methods for Computational Modelling

Fatemeh Safari , Jai J Tree , Fatemeh Vafaee *

Posted Date: 25 August 2025

doi: 10.20944/preprints202508.1739.v1

Keywords: RNA bioinformatics; non-coding RNA (ncRNA); feature extraction; machine learning; sequence; representation

 doi:10.20944/preprints202508.1739.v1

Review

From Nucleotides to Numbers: A Comprehensive Review of RNA Feature Extraction Methods for Computational Modelling

Fatemeh Safari [1,2], Jai J Tree [1] and Fatemeh Vafaee [1,2,3,*]

[1] School of Biotechnology and Biomedical Sciences, Faculty of Science, University of New South Wales, Sydney, NSW 2052, Australia

[2] UNSW Biomedical AI, University of New South Wales, Sydney NSW 2052, Australia

[3] UNSW AI Institute, University of New South Wales, Sydney NSW 2052, Australia

* Correspondence: f.vafaee@unsw.edu.au; Tel.: +61-(2)-9065-2699

Abstract

Machine learning is a powerful approach for analysing RNA sequences, particularly for understanding the function and regulation of non-coding RNAs. A critical step in this process is feature extraction, which transforms biological sequences into numerical representations that allow computational models to capture and interpret complex biological patterns. Despite its central role, the field of RNA feature extraction remains broad and fragmented, with limited standardization and accessibility hindering consistent application. In this comprehensive review, we address the fragmentation of the field by systematically organizing over 25 feature extraction strategies into sequence- and structure-based approaches. We further conduct a comparative analysis highlighting how the choice of feature sets impacts model performance, reinforcing the importance of integrated feature engineering. To facilitate practical adoption, it also provides a curated list of publicly available tools and software packages. By consolidating methodologies and resources, this work seeks to improve reproducibility, scalability, and interpretability in machine learning-driven RNA research.

Keywords: RNA bioinformatics; non-coding RNA (ncRNA); feature extraction; machine learning; sequence representation

1. Introduction

RNA sequencing (RNA-seq) has revolutionized transcriptomics by enabling the comprehensive analysis of RNA expression across various cell types, tissues, and biological conditions [1,2]. Beyond quantifying gene expression, RNA-seq data support diverse applications such as the discovery of novel transcripts, annotation of non-coding RNAs (ncRNAs), and exploration of transcriptomic diversity [1]. RNA molecules, including messenger RNAs (mRNAs) and various classes of non-coding RNAs such as microRNAs (miRNAs), long non-coding RNAs (lncRNAs), small RNAs (sRNAs), and circular RNAs (circRNAs), play essential roles in gene regulation, RNA processing, epigenetic control, and molecular interactions in both prokaryotic and eukaryotic organisms [3–6]. Determining the sequence and structural properties of these RNAs is therefore critical for understanding cellular behaviour, genetic regulatory regions, and identifying biomarkers or therapeutic targets [2].

With an increasing number of RNA-seq datasets, one of the key challenges is the transformation of raw sequence data into meaningful, quantifiable features suitable for computational modelling [7,8]. Machine learning (ML) algorithms cannot interpret nucleotide sequences in their original form and therefore require the data to be converted into informative numerical representations. This transformation is achieved through feature extraction, a vital preprocessing step that encodes sequence and structural properties into numerical formats that retain relevant biological patterns

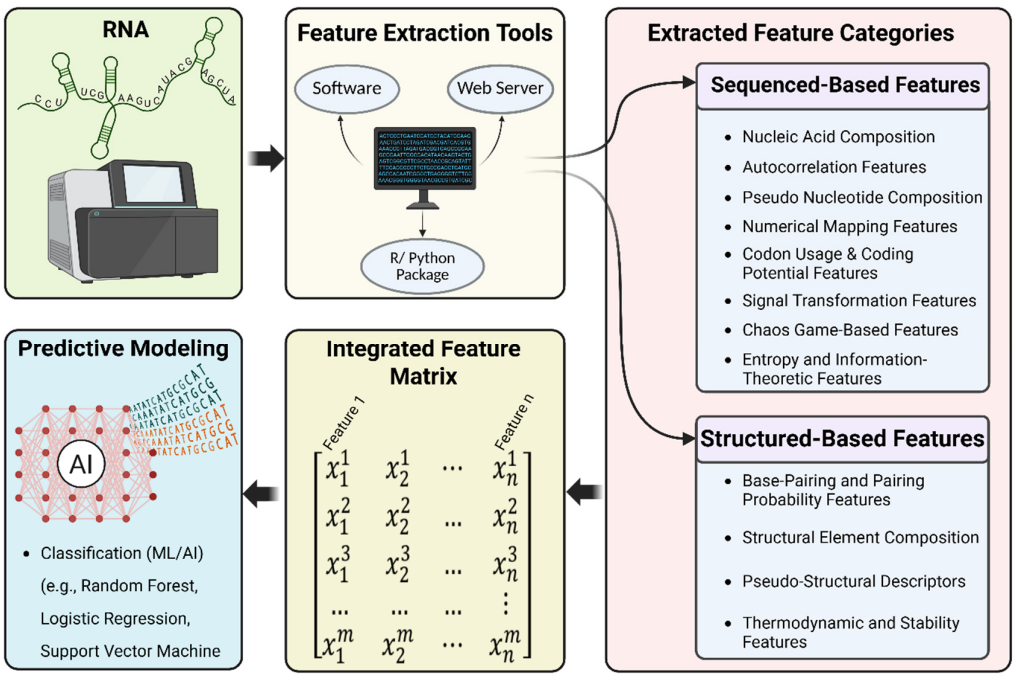
 doi:10.20944/preprints202508.1739.v1

while minimizing noise and redundancy [9–11]. These extracted features facilitate the development of predictive models based on machine learning, which can be applied to various domain-specific applications in molecular biology and biomedical research. These applications include but are not limited to: RNA classification (e.g., non-coding vs. coding RNAs), RNA-protein and RNA-RNA interaction prediction, transcript stability analysis, prediction of subcellular localization, functional annotation, and the design of therapeutic RNAs, including small interfering RNAs (siRNAs), RNA aptamers, and CRISPR guide RNAs. The quality and consistency of the extracted features are critical to the effectiveness of these applications, as they influence model accuracy, generalizability, and interpretability [7,8]

In parallel with traditional feature extraction methods, deep learning-based representation learning approaches have emerged as a promising direction in computational biology [12]. Representation learning aims to automatically extract meaningful features directly from raw sequence data, thereby eliminating the need for manual feature design [13]. However, despite its potential, representation learning faces challenges such as its reliance on large datasets, susceptibility to overfitting when applied to small datasets, significant computational requirements, and function as black-box models, limiting transparency in decision-making [14–17]. For small to medium-sized datasets, traditional machine learning methods such as support vector machines, random forests, and gradient boosting remain effective alternatives. Although they require structured feature engineering, which involves additional preprocessing, this process enables a more interpretable and systematically controlled modelling approach [18].

Despite increasing interest in ML for RNA analysis, there is no consolidated overview of feature extraction techniques tailored to RNA sequences and structures. Existing approaches are scattered across domains, vary in implementation, and lack standardized documentation, hindering reproducibility and accessibility, particularly for researchers with limited programming expertise.

To address this gap, this review provides a structured, accessible overview of established feature extraction strategies for RNA, categorized into sequence-based and structure-based methods. **Figure 1** outlines the complete workflow, from raw RNA sequences through feature extraction and integration into predictive modelling frameworks. In addition to methodological categorization, we compile publicly available tools and software packages to support practical implementation. By organizing and contextualizing existing methods, this work aims to advance reproducibility, accessibility, and interpretability in ML-driven RNA biology.



 doi:10.20944/preprints202508.1739.v1

Figure 1. Overview of RNA feature extraction and machine learning workflow. RNA sequences are processed using software tools, web servers, or programming packages to extract informative numerical features. These features are categorized as sequence-based or structure-based, assembled into a feature matrix, and used to train machine learning models for various RNA-related predictive tasks.

2. Foundations of RNA Feature Extraction

2.1. Sequence-Based Features

Feature extraction from RNA sequences is a critical step in machine learning based RNA analysis, transforming raw nucleotide strings into structured quantitative representations suitable for predictive modelling. A wide range of feature extraction techniques have been developed to encode RNA sequences, spanning from simple frequency-based representations to advanced network-theoretic approaches [19], many of which are implemented in open-source toolkits [20]. Broadly, sequence-derived features can be categorised into the following groups: nucleotide composition-based features, numerical mapping and signal transformation methods, Fourier and Chaos-based features, entropy and information-theoretic measures, autocorrelation-based descriptors, pseudo nucleotide compositions, and similarity or instance-based features [19]. This categorization reflects an increasing level of computational and biological sophistication, progressing from the capture of local nucleotide patterns to the modelling of long-range dependencies, physicochemical properties, and structural complexities embedded within RNA sequences.

2.1.1. Nucleic Acid Composition

These methods capture short range or local sequence order by counting the occurrence frequencies of adjacent or non-contiguous residues include:

One-hot encoding: One hot encoding is a widely adopted feature extraction technique that represents each nucleotide in the RNA sequence as a unique binary vector. An RNA sequence composed of the four bases A, U, C, and G can be represented by a 4-dimensional vector for each base. For example, A is represented as [1,0,0,0], U as [0,1,0,0], C as [0,0,1,0], and G as [0,0,0,1]. Therefore, an RNA sequence of length L can be expressed as a 4 × L dimensional binary matrix in which each column corresponds to a sequence position and each row represents a specific nucleotide [21].

K-mer composition: The k-mer feature counts the frequency of distinct nucleotide subsequences of length k within the RNA sequence. This is achieved by sliding a window of length k along the sequence and counting how often each possible k-mer appears. The process considers all contiguous subsequences of size k from position 1 to position (L − k + 1). The frequency ($\boldsymbol{f_{cs}}$) is calculated as:

$$f_{cs} = \frac{\left(\frac{C_k}{4^{K-k}}\right)}{L-k+1} \quad \text{Eq. 1}$$

where C_k is the count of a specific k-mer, L is sequence length, K is the maximum k value, and 4 denotes the four nucleotide types [22]. K-mer features have been widely applied in the analysis of RNA sequence properties, including classification of coding and non-coding RNAs, identification of structural motifs, and functional annotation tasks [23–25].

Enhanced Nucleic Acid Composition (ENAC): Local nucleic acid composition can be calculated using the Enhanced Nucleic Acid Composition encoding, which applies a fixed length sliding window that moves sequentially from the 5′ to the 3′ end of the nucleotide sequence. This method is generally applied to nucleotide sequences of equal length. The sliding window size and sequence length determine the ENAC encoding dimension, calculated as $(\boldsymbol{sequence\ length} - \boldsymbol{window\ size} + \mathbf{1}) \times \mathbf{4}$. The ENAC encoding is defined as follows [26]:

$$E = (b_1, b_2, \ldots, b_n), \quad \text{Eq. 2}$$

$$b(i) = \frac{N_{(i)}}{N} \quad \text{Eq. 3}$$

 doi:10.20944/preprints202508.1739.v1

$$i \in \{A, C, G, T/U\}$$

where $b(i)$ represents the frequency of nucleotide i within a given window (N) of the sequence and $N_{(i)}$ is the count of nucleotide i within that window.

Reverse complement k-mer: The reverse complement k-mer (k-RevcKmer) is a variation of the standard k-mer feature used in RNA sequence analysis. In this approach, both the original k-mers in the sequence and their reverse complements are considered during feature extraction. First, all possible k-mers are generated from the RNA sequence. Any k-mer that is identical to its reverse complement is removed to avoid redundancy. The remaining k-mers are then used to construct a feature vector, with each feature representing the frequency of a specific k-mer in the sequence. This method reduces the dimensionality of the k-mer space while retaining information from complementary strand orientations [27].

Mismatch Profile: The mismatch profile approach is an extension of traditional k-mer counting that allows up to m mismatches within each k-mer, where $m < k$. For example, if $m = 1$ and $k = 3$, the notation (3, 1) refers to a 3-length subsequence with at most one mismatch. Considering a 3-mer "AAC" with one allowed mismatch, the count would include not only "AAC" itself, but also variants such as "AAG," "AAA," "AAU," "GAC," "CAC," and "UAC" that appear in the sequence. The mismatch profile of a sequence x can be expressed as:

$$f_{k,m}^{mis}(x) = \left(\sum_{j=0}^{m} C_{1,j}, \sum_{j=0}^{m} C_{2,j}, \ldots, \sum_{j=0}^{m} C_{4^k,j}\right) \quad \text{Eq. 4}$$

Here, C_{ij} indicates the frequency of the i-th k-mer variant in sequence x with j mismatches, where i ranges from 1 to 4^k and j from 0 to m. By incorporating both exact matches and near matches, the mismatch profile captures a broader spectrum of sequence patterns, potentially revealing biologically significant variations that standard k-mer counts may miss [28,29].

xxKGAP Encoding: The xxKGAP composition is a key approach employed in PyFeat package ***[7]***, considering kgaps in RNA sub-sequences. A sliding window is utilized to count the occurrences of discontinuous bases with g gaps ($\boldsymbol{C_g}$), and the frequency ($\boldsymbol{f_{ds}}$) is calculated as:

$$f_{ds} = (C_g / 4^{G+2-g}) / (L - g - 1) \quad \text{Eq. 5}$$

where G represents the maximum value of g [22]. For example, the sequence can be encoded into X_X frequencies for mMKGap features with a kgap of 1, producing 16-dimensional features ($4 \times 1 \times 4$). If kgap = 2, the sequence can be characterised by 32 features ($4 \times 2 \times 4$). For dMKGap, the total number of features is calculated as $4^2 \times n \times 4$ [20]. This representation allows the capture of dependencies between nonadjacent nucleotides, which can reflect structural or functional patterns in RNA sequences.

GC content: GC content indicates the proportion of guanine and cytosine nucleotides within an RNA sequence. This metric is often employed to differentiate protein-coding regions from non-coding sequences. Generally, non-coding elements such as 5′ untranslated regions (UTRs) and introns have a lower percentage of GC bases compared to protein-coding sequences. The GC content is calculated as follows [30]:

$$GC\ Content = \frac{N(G)+N(C)}{L_t} \quad \text{Eq. 6}$$

where $N(C)$ and $N(G)$ refer to the numbers of G and C nucleotides respectively, and L_t is the overall transcript length.

Accumulated nucleotide frequency: The accumulated nucleotide frequency (ANF) encoding system represents the density and distribution of each nucleotide within a sequence ***[26]***. To capture the nucleotide frequency and the distribution of each nucleotide in the RNA sequence, the density ($\boldsymbol{d_i}$) of any nucleotide ($\boldsymbol{S_i}$) at position $\boldsymbol{i}$ in the RNA sequence is defined using the following formula ***[31]***,

 doi:10.20944/preprints202508.1739.v1

$$d_i = \frac{1}{|S_i|}\sum_{j=1}^{l} f(s_j), \text{ Eq. 7}$$

$$f(q) = \begin{cases} 1 \; if \; s_j = q \\ 0 \; other \; cases \end{cases} \text{ Eq. 8}$$

Here, l represents the length of the sequence, $|S_i|$ denotes the length of the i -th prefix string $\{s_1, s_2, \ldots s_i\}$ within the sequence, and $q \in \{A, C, G \; or \; U\}$. For the example sequence "UCGUUCAUGG", the density of each nucleotide is as follows: For 'U', the density is 1 (1/1) at position 1, 0.5 (2/4) at position 4, 0.6 (3/5) at position 5, and 0.5 (4/8) at position 8. For 'C', the density is 0.5 (1/2) and 0.33 (2/6) at positions 2 and 6, respectively. The density of 'A' is 0.14 (1/7) at position 7. Finally, the density of 'G' is 0.33 (1/3) at position 3, 0.22 (2/9) at position 9, and 0.3 (3/10) at position 10 [31].

AUGC Ratio: The AU/GC ratio is a simple compositional feature that generates a single scalar value for each RNA sequence. It measures the relative abundance of adenine and uracil bases compared to guanine and cytosine bases. The ratio is calculated as *[*32*]*.

$$AU/_{GC}\, Ratio = \frac{\sum A + \sum U}{\sum G + \sum C} \text{ Eq. 9}$$

GC Skew: GC skew, calculated as $(G - C/(G + C)$, measures strand-specific nucleotide asymmetry and is commonly used to determine replication origin and terminus in bacterial genomes [33,34]. Although originally developed as a genome level measure, GC skew can also be applied to RNA sequences to provide additional compositional information that may be relevant for distinguishing functional classes or structural properties [32,35].

2.1.2. Autocorrelation Descriptors

These approaches look for correlations between two di- or trinucleotides based on their physicochemical properties for RNA sequence analysis. Unlike simple compositional features, which only quantify nucleotide frequencies, autocorrelation descriptors preserve sequence-order information and can reveal periodic or long-range dependencies, making them useful for complex sequence analysis tasks. Two widely used approaches are autocovariance, which measures correlations of the same physicochemical property across nucleotide groups at a defined distance, and cross-covariance, which assesses correlations between different physicochemical indices [36]. According to the approaches applied in several studies for RNA, the autocorrelation module is divided into several categories based on different properties and correlation types. These include dinucleotide-based autocorrelation (DAC), dinucleotide-based Moran autocorrelation (DMAC), dinucleotide-based Geary autocorrelation (DGAC), and normalised Moreau-Broto autocorrelation (NMBAC). Similarly, for cross-correlation and auto-cross-correlation modules, two methods exist for RNA: dinucleotide-based cross-correlation (DCC) and dinucleotide-based auto-cross-correlation (DACC) [37,38].

2.1.3. Pseudo Nucleotide Composition

The third category of sequence-derived features includes pseudo k-tuple nucleotide composition (PseKNC) methods, which are designed to capture both global and long-range sequence-order information, as well as physicochemical properties of nucleotides. Due to their strong performance across various predictive tasks, several versatile web servers and software tools have been developed to generate pseudo nucleotide composition features [39–41]. A comprehensive overview of pseudo nucleotide composition approaches can be found in a recent review [42]. Within this category, pseudo dinucleotide composition (PseDNC) encoding is one of the most widely used methods in RNA sequence analysis. PseDNC takes into account not only the sequential arrangement of nucleotides but also the physicochemical properties of dinucleotide pairs within the RNA molecule, resulting in a numerical feature set for each analysed sequence. The total number of PseDNC features is given by 16+λ. The initial 16 features are derived from pairs of adjacent

 doi:10.20944/preprints202508.1739.v1

dinucleotides. The remaining λ features are calculated based on dinucleotide pairs that are separated by different distances along the sequence. λ denotes the greatest possible separation between any two dinucleotides considered in the analysis [43]. Several publicly available packages have been developed to extract PseDNC features such as Pse-in-One 2.0 [44], repRNA [45], and UltraPse [46].

2.1.4. Numerical Mapping Features

Real, integer, and complex number mappings: In sequence analysis, numerical mapping methods such as integer, complex, and real number representations are widely used to convert symbolic nucleotide sequences into numerical form suitable for computational analysis ***[***19***]***. Integer mapping assigns simple whole numbers to nucleotides, for example A = 0, C = 1, G = 2, and T/U = 3 ***[***47***]***. Complex number mapping places nucleotides as points in the complex plane, such as A = 1 + i, T/U = 1 − i, C = −1 − i, and G = −1 + i ***[***48***]***. Real number mapping, on the other hand, uses continuous real values such as A = −1.5, T/U = 1.5, C = 0.5, and G = −0.5. This representation has the useful property that complementary sequences can be derived by reversing the sequence order and changing the sign of each value ***[***49***]***.

EIIP: EIIP encoding transforms RNA sequences into numerical feature vectors by assigning each nucleotide a specific electron-ion interaction Pseudopotentials (EIIP) value: A = 0.1260, C = 0.1340, G = 0.0806, and U = 0.1335 ***[***50***]***. To represent trinucleotide composition, this method constructs a 64-dimensional feature vector in which each element corresponds to a specific trinucleotide. For a trinucleotide sequence $\boldsymbol{mno}$, the EIIP value is calculated as:

$$\boldsymbol{EIIP_{mno} = EIIP_m + EIIP_n + EIIP_o} \quad \text{Eq. 10}$$

where m, n, o ∈ {A, C, G, U} and f_{mno} is the frequency of that trinucleotide in the sequence. The resulting vector is:

$$D = [EIIP_{AAA} \times f_{AAA}, EIIP_{AAC} \times f_{AAC}, EIIP_{AAA} \times f_{AAA}, \ldots\ldots, EIIP_{UUU} \times f_{UUU}] \quad \text{Eq. 11}$$

Z-Curve: The Z-curve theory, originally developed for DNA sequence analysis, is a three-dimensional representation of a sequence's base distribution ***[***51***]***. This method can be effectively adapted for RNA sequence analysis due to its distinct geometrical properties and the similarity between RNA and DNA nucleotide structures, with the primary difference being the substitution of uracil (U) for thymine (T) ***[***11,52,53***]***. The Z curve is formed by a series of nodes, P_0, P_1,...P_N, where N is the sequence length and each node has coordinates $\boldsymbol{X_n}$, $\boldsymbol{Y_n}$, $\boldsymbol{Z_n}$ defined as:

$$x_n = (A_n + G_n) - (C_n + U_n) \quad \text{Eq. 12}$$

$$y_n = (A_n + C_n) - (G_n + U_n) \quad \text{Eq. 13}$$

$$z_n = (A_n + U_n) - (C_n + G_n) \quad \text{Eq. 14}$$

$$n = 0, 1, 2, \ldots, N$$

where A_n, G_n, C_n, U_n denote the cumulative counts of each nucleotide from the first position up to position n in the sequence.

Nucleotides are classified into six categories based on their properties: purine (R = A, G) versus pyrimidine (Y = C, U), amino (M = A, C) versus keto (K = G, U), and hydrogen bond strength, strong (S = G, C) versus weak (W = A, U). The x-component of the Z-curve represents the distribution of purines and pyrimidines, the y-component corresponds to amino and keto distribution, and the z-component reflects the distribution of strong and weak hydrogen bonds in the nucleotide sequence [54]. As a result, three numerical features can be generated from the Z-curve representation for downstream analysis.

2.1.5. Codon Usage and Coding Potential Features

Fickett Score: The Fickett score is a feature extraction method designed to differentiate coding from non-coding RNAs by integrating nucleotide composition with codon usage bias. It evaluates

 doi:10.20944/preprints202508.1739.v1

four position values and four content values for each sequence followed by a weighted summation. The position values capture the preference of each nucleotide (A, C, G, U) for specific positions within codons, offering insights into positional biases within the transcript. For each nucleotide, its position value within the RNA transcript is determined using the following formula:

$$A_1 = N(base\ A\ in\ Position\ 0,3,6,\dots) \quad \text{Eq. 15}$$

$$A_2 = N(base\ A\ in\ Position\ 1,4,7,\dots) \quad \text{Eq. 16}$$

$$A_3 = N(base\ A\ in\ Position\ 2,5,8,\dots) \quad \text{Eq. 17}$$

$$A_{pos} = \frac{Max\ (A_1,A_2,A_3)}{Min\ (A_1,A_2,A_3)+1} \quad \text{Eq. 18}$$

Here, $N()$ represents the total count of nucleotides under the specified condition. The values for U_{pos}, G_{pos}, C_{pos} are derived in the same way as A_{pos}. The overall content-based metrics for each nucleotide in the transcript are then computed as follows:

$$A_{content} = \frac{N(base\ A\ in\ an\ RNA\ transcript)}{L_t} \quad \text{Eq. 19}$$

The calculation methods for $U_{content}$, $G_{content}$, and $C_{content}$ are identical. Ultimately, a lookup table is employed to transform the four positional attributes and four compositional attributes into probabilities indicative of coding potential. The Fickett score is then derived by multiplying these eight probability values (p) by their respective weighting factors (w). These weights reflect the effectiveness of each positional or compositional feature in distinguishing between coding and non-coding sequences [30,55].

$$Fickett\ score = \sum_{i=1}^{8} p_i w_i \quad \text{Eq. 20}$$

Relative Codon Bias (RCB): Relative codon bias serves as a metric to quantify the non-uniform usage of codon triplets within the open reading frames (ORFs) of an RNA transcript. It measures how much the observed codon usage deviates from what would be expected based on the independent nucleotide composition at each codon position. To derive the RCB value for an ORF, the product of the individual usage biases for all its codon triplets is computed. The codon usage bias $\boldsymbol{d_{xyz}}$ for a specific triplet $(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})$ is determined as follows:

$$d_{xyz} = \frac{f(x,y,z) - f_1(x)f_2(y)f_3(z)}{f_1(x)f_2(y)f_3(z)} \quad \text{Eq. 21}$$

$$f(x,y,z) = \frac{N(x,y,z)}{L_{codon}} \quad \text{Eq. 22}$$

$$f_1(x) = \frac{N(base\ x\ in\ position\ 0\ of\ each\ codon)}{L_{codon}} \quad \text{Eq. 23}$$

Here, $N(x,y,z)$ is the count of the codon triplet (x,y,z) found in the ORF, and L_{codon} is the ORF length in codons. The calculations for $f_2(y)$ and $f_3(z)$ at the second and third positions are analogous to the calculation of $f_1(x)$ at the first nucleotide position. Subsequently, the complete RCB value for the RNA transcript's ORF is computed as shown below [30,56]:

$$RCB = \left(\prod_{i=1}^{L_{codon}} (1 + d_{xyz}^{i})\right)^{1/L} - 1 \quad \text{Eq. 24}$$

ORF Related Features (Max ORF length, Max ORF coverage, Average ORF length, Average ORF coverage): An open reading frame (ORF) is a segment within an RNA transcript that has the potential to encode a protein. Analyses of ORF characteristics are commonly used to distinguish protein-coding transcripts from long non-coding RNAs (lncRNAs), although these features are generally less effective for differentiating among various lncRNA subtypes [57]. To capture protein-coding information more comprehensively, analyses may extend beyond the conventional definition of an ORF as the region between a start codon and a stop codon. Alternative definitions include ORFs that begin with a start codon and extend to the transcript's end, or segments that span from any non-stop codon to a stop codon. An integrated approach can also select the longer sequence between these

 doi:10.20944/preprints202508.1739.v1

start-codon-focused and stop-codon-focused variants. For each ORF type, the maximum length across all three reading frames can be determined and extracted as the max ORF.

In addition to their absolute length, the max ORF coverage can be computed by dividing the max ORF length by the total transcript length. Furthermore, for conventionally defined ORFs (bounded by start and stop codons), both the average ORF length and the average ORF coverage are determined. These features provide informative measures of coding potential and have been widely applied in computational transcript classification [30].

2.1.6. Signal Transformation Features (Fourier-Based)

This category encompasses feature extraction techniques that convert RNA sequences into numerical signals and then apply methods from genomic signal processing (GSP) to derive informative features. Among these, the Fourier Transform (FT) is one of the most widely used approaches in biological sequence analysis [10,58,59]. A detailed mathematical formulation of the Fourier-based approach for nucleotide sequences is provided in [10].

2.1.7. Chaos Game-Based Features

Chaos Game Representation (CGR) is a method that visually encodes RNA sequences as two-dimensional fractal patterns derived from nucleotide composition. For machine learning applications, CGR can be quantified using Frequency Chaos Game Representation (FCGR), in which the fractal image is divided into a grid and the frequencies of subsequences falling into each grid cell are counted. This process generates a numerical matrix that can be flattened into a fixed-length vector, providing an alignment-free feature representation for RNA sequence analysis. The detailed methodology and applications of CGR and FCGR are described in [60].

2.1.8. Entropy and Information-Theoretic Features

Several studies have applied concepts from information theory to extract meaningful features from biological sequences, with Shannon entropy (SE) being one of the most widely used measures [61,62]. SE quantifies the uncertainty or diversity in the distribution of nucleotides or k-mers within a sequence, providing insights into its complexity. In addition to SE, Tsallis entropy (TE) [63,64] has been successfully employed as an alternative or complementary descriptor in sequence analysis. TE generalizes the concept of entropy by introducing a parameter that can adjust the sensitivity of the measure to rare or frequent events. Both SE and TE capture important statistical properties of RNA sequences and can be applied at different k-mer levels to highlight sequence variability and compositional bias [30].

2.2. Structural Feature Extraction

Understanding the structural configuration of an RNA molecule is an essential first stage in uncovering its functional mechanisms [65]. Among structural characteristics, the secondary structure is particularly critical in diverse biological processes and is often more conserved than the primary sequence [66]. The set of base pairs formed through hydrogen bonding between nucleotides defines the RNA secondary structure. The main challenge in secondary structure prediction lies in determining which nucleotides are paired with each other in a given sequence [67]. Thermodynamic principles can be used to predict the secondary structure of an RNA sequence [68]. These thermodynamics-based methods employ nearest-neighbour parameters to estimate structural stability, which is quantified by the change in folding free energy [69–71]. Structure prediction is commonly achieved by determining the conformation with the lowest free energy [65]. Minimum free energy (MFE) acts as a fundamental structural indicator, reflecting the stability of the RNA structure [66]. The assumption is that a lower free energy implies greater stability of the RNA secondary structure [30]. Alternative prediction strategies include sampling from the Boltzmann

 doi:10.20944/preprints202508.1739.v1

ensemble to identify a representative centroid structure [72] or selecting the structure with the highest sum of base-pairing probabilities, known as the maximum expected accuracy (MEA) structure [73].

A widely used tool for RNA secondary structure prediction is RNAfold, part of the ViennaRNA Package, which applies MFE calculations to identify the most probable configurations [74]. RNAfold decomposes RNA secondary structures into elements such as interior loops, hairpin loops, multiloops, bulge loops, and stacking pairs, with each contributing to the total free energy. The total free energy of RNA's secondary structure is determined by summing the free energy values of its constituent substructures. The most stable predicted structure is generated for each RNA transcript and used for downstream feature extraction [74]. A comprehensive list of available RNA secondary structure prediction tools is available at [75].

In a study conducted by Kang et al. [22], RNA secondary structures were predicted using the RNAfold package, which represents structural features using a system of brackets ("(" or ")" = paired nucleotide) and dots ("." = unpaired). These approaches can extract both continuous and discontinuous structural patterns. However, unlike sequence-based analyses that consider the four nucleotide types, structural analysis is constrained to two symbol types (brackets and dots), necessitating adjustments in calculation parameters. Accordingly, several structural features can be derived from the dot-bracket notations produced by these tools, as described in the following paragraphs.

2.2.1. Paired Ratio

This is a metric based on the secondary structure of an RNA transcript, representing the proportion of nucleotides involved in Watson-Crick base pairing compared to those that remain unpaired. This ratio is used to assess structural stability; RNA molecules with a higher percentage of paired nucleotides have more stable secondary structures **[**30**]**. The formula is as follows:

$$Paired\ Ratio = \frac{N(Paired\ Nucleutide\ Bases)}{L_t} \quad \text{Eq. 25}$$

2.2.2. Triplet

This method integrates both sequence and structure information and has shown superior performance in tasks such as microRNA identification **[**45,76**]**. Using dot/bracket notation, there are 8 (2^3) possible structural configurations for a set of three adjacent nucleotides: '(((', '((.', '(..', '(.(', '.((', '.(.', '..(', and '...'. By focusing on the middle nucleotide within each group of three, 32 possible structure-sequence combinations (4 × 8) can be obtained, denoted as f_A ('((('), f_G ('(((('), etc. These combinations define the triplet structure-sequence elements, which integrate both nucleotide sequence and corresponding structural information, allowing for comprehensive analysis **[**21,45**]**.

2.2.3. Pseudo-Structure Status Composition (PseSSC) & Pseudo-Distance Structure Status Pair Composition (PseDPC)

Liu et al. proposed PseSSC and PseDPC methods for capturing the compositional and sequential information of RNA sequences by efficiently representing RNA secondary structures like stem loops. These approaches approximate the sequential information of RNA sequences employing a correlation function based on secondary structure status, considering both the distance between structural status pairs and the minimum free energy **[**77**]**. Details can be found in the referenced sources **[**78,79**]**.

2.2.4. Number of Distinct Loop Structures

This metric counts different loop types in the secondary structure, including interior loops (N(I)), hairpin loops (N(H)), bulge loops (N(B)), and multibranch loops (N(M)) **[**30**]**. The typical loop structures are illustrated in Figure 2.

 doi:10.20944/preprints202508.1739.v1

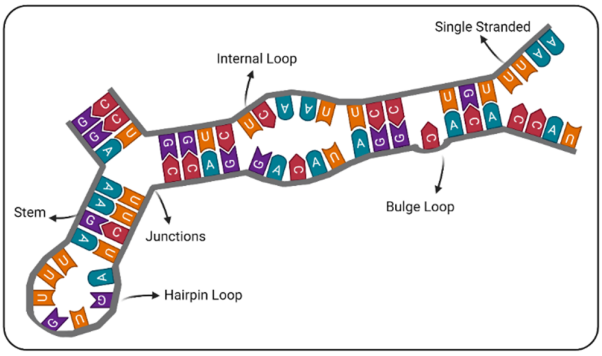


Figure 2. Example of an RNA secondary structure illustrating various types of structural elements. (Created in BioRender.com).

2.2.5. Coverage of Different Loop Structures

For each loop type, coverage is computed as the number of loops divided by the transcript length **[**30**]**.

$$C(H) = \frac{N(H)}{L_t} \text{ Eq. 26}$$

$$C(I) = \frac{N(I)}{L_t} \text{ Eq. 27 } C(B) = \frac{N(B)}{L_t} \text{ Eq. 28}$$

$$C(M) = \frac{N(M)}{L_t} \text{ Eq. 29}$$

2.2.6. GC Content of Paired Nucleotides

This attribute is calculated as the proportion of guanine-cytosine (G-C) base pairs in the secondary structure of an RNA transcript. G-C bonds are stronger and more stable than adenine-thymine/uracil (A-T/U) bonds, so a higher GC content in paired nucleotides typically reflects a more stable secondary structure for the RNA transcript **[**30**]**.

$$GC\ content\ paired\ nucleotides = \frac{N(Paired\ G) + N(Paired\ C)}{N\ (paired\ nucleotides)} \text{ Eq. 30}$$

3. Comparative Impact of Feature Set Choice on Model Performance

The predictive performance of machine learning models in RNA analysis is strongly influenced by the composition of the input feature set [9,79]. While the choice of algorithm plays a role, the diversity and informativeness of the features are equally critical in determining model accuracy. Evidence from four independent studies [81–84], each employing different combinations of RNA-derived features for various RNA classification tasks, highlights this effect (**Figure 3**). Across these examples, a consistent pattern is observed: models trained on integrated feature sets, combining multiple descriptor types, often outperform those relying on a single feature category. This reinforces the view that strategic feature engineering is not a preliminary or optional step but a core element in building reliable predictive models in RNA biology. Consequently, the careful composition of complementary feature types can yield substantial performance gains, often independent of the specific model architecture employed.

 doi:10.20944/preprints202508.1739.v1

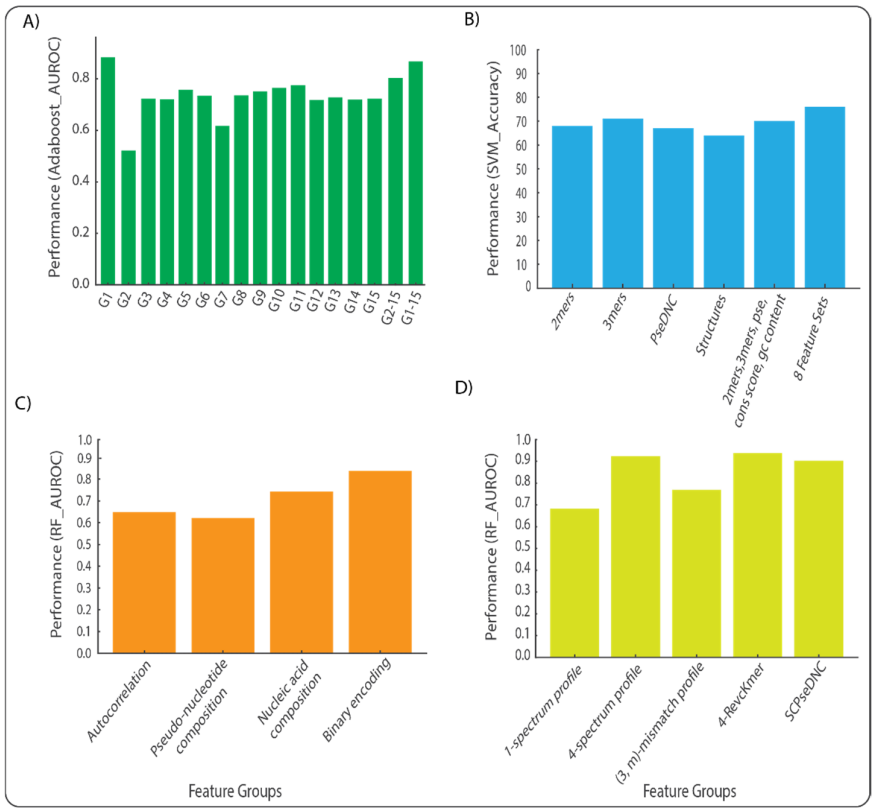


Figure 3. Performance comparison across different feature sets and machine learning models in four independent RNA-related studies. A) Area Under the Receiver Operating Characteristic (AUROC) curve for an AdaBoost model predicting bacterial small RNAs. Performance is evaluated for 15 individual feature groups (G1–G15)* and their combinations, demonstrating that combined feature sets generally yield higher predictive power [81]. B) Accuracy of a Support Vector Machine (SVM) model predicting disease-related lncRNAs, where the integration of multiple feature groups improves performance [82]. C) AUROC values from a Random Forest model predicting lncRNA localization, comparing the effectiveness of four descriptor categories [83]. D) AUROC values for a Random Forest model predicting bacterial small RNAs, illustrating differences in predictive ability across five distinct sequence encoding strategies [84].*The 15 individual feature groups (G1–G15) in panel A are, respectively: Biological features; 1-mer to 5-mer frequencies; 1-mer to 5-mer reverse complement k-mer (RCkmer) frequencies; PCPseDNC, parallel correlation pseudo-dinucleotide composition; PCPseTNC, parallel correlation pseudo-trinucleotide composition; SCPseDNC, series correlation pseudo-dinucleotide composition; SCPseTNC, series correlation pseudo-trinucleotide composition.

4. Feature Extraction Tools

Over the years, a variety of computational tools have been developed to facilitate the extraction of RNA sequence and structure features. These tools implement a broad spectrum of methodologies, enabling the derivation of descriptors such as k-mer frequencies, physicochemical properties, structural stability metrics, entropy-based measures, and other specialized attributes discussed in previous sections. Table 1 presents a summary of widely used tools in the literature, outlining their primary functionalities and feature categories. The availability of these resources has significantly streamlined the process of generating high-dimensional, informative feature sets for downstream machine learning applications in RNA biology.

 doi:10.20944/preprints202508.1739.v1

Table 1. Publicly available tools for extracting sequence-based and structure-based features from RNA sequences.

Tool/Package Name	Access Type	Type of Feature Categories	Published Year	Ref
RepRNA	Web server	Oligonucleotide composition; pseudo-nucleotide composition; structure composition	2016	[45]
PseKNC	Web server	Pseudo-dinucleotide composition (PseDNC); pseudo-trinucleotide composition (PseTNC)	2014	[39]
PseKNC-General	Web server	K-tuple nucleotide composition; autocorrelation descriptors; pseudo-nucleotide composition	2015	[40]
BioTriangle	Web server	Nucleic acid composition; autocorrelation descriptors; pseudo-nucleotide composition	2016	[85]
BioSeq-Analysis2.0	Web server	Residue-level composition; sequence-level physicochemical and structural descriptors	2019	[86]
BioSeq-Analysis	Standalone program & web server	Nucleic acid composition; autocorrelation descriptors; pseudo-nucleotide composition; predicted structure composition	2019	[36]
Nfeature	R/Python package & web server	Nucleic acid composition; distance distribution of nucleotides; nucleotide repeat index; pseudo-composition; entropy	2021	[37]
iLearn	Python toolkit	Nucleic acid composition; binary encoding; position-specific trinucleotide tendencies; autocorrelation; pseudo-composition	2019	[87]
iLearnPlus	R/Python package & web server	Nucleic acid composition; residue composition; position-specific trinucleotide tendencies; autocorrelation; physicochemical; mutual information; similarity-based; pseudo-composition	2021	[38]
ftrCOOL	R/Python package	Nucleic acid composition; substitution matrices; k-nearest-neighbor RNA; local position-specific k-frequency; maxORF-based	2022	[88]

 doi:10.20944/preprints202508.1739.v1

PyFeat	Python toolkit	Z-curve; GC content; AT/GC ratio; cumulative skew; Chou's pseudo-composition; k-gap statistics	2019	[32]
MathFeature	R/Python package & web server	Numerical mapping; chaos game descriptors; Fourier transform; entropy and graph descriptors; pseudo-composition	2022	[19]
Pse-In-One	Web server	Nucleic acid composition; autocorrelation descriptors; pseudo-nucleotide composition	2015	[41]
Pse-in-One 2.0	Web server	Nucleic acid composition; autocorrelation; triplet sequence-structure elements; pseudo-structure status composition; PseDPC	2017	[44]
UltraPse	Software platform	Nucleic acid composition; autocorrelation descriptors; pseudo-nucleotide composition	2017	[46]

5. Discussion and Conclusions

The rapid expansion of publicly available RNA sequence data has created new opportunities for computational approaches to uncover their biological roles. Despite this availability, many RNA sequences remain poorly characterized with respect to their functional and structural properties. Machine learning has emerged as a powerful framework for addressing this gap, but its success depends heavily on how effectively raw sequences are transformed into informative numerical representations. This review has provided a detailed overview of descriptor categories and feature extraction strategies for encoding RNA sequences and structures into numerical form. We have also discussed available tools and platforms that implement these methods and highlighted how the choice and diversity of feature sets can influence the predictive performance and interpretability of machine learning models in RNA-related applications.

Most prediction tasks in biological sequence analysis are framed as binary or multi-class classification problems. Numerous efficient computational approaches have been developed using machine learning algorithms to predict or analyse sequence-related characteristics solely from sequence information [87]. However, most existing machine-learning techniques, such as SVM (support vector machine) and KNN (k-nearest neighbour), are designed to handle numerical vectors rather than raw sequences [45]. Consequently, feature extraction plays a pivotal role in converting sequences into mathematical representations that preserve their intrinsic relationship with the target variable, thereby directly influencing model performance [89]. To facilitate this process, a range of web-based servers and stand-alone software tools have been developed, enabling the extraction of diverse sequence, structural, and physicochemical features [37,44,45,86]. Nevertheless, significant challenges remain. Many existing tools focus on a narrow subset of features, limiting their ability to integrate both sequence- and structure-based information in a unified framework. This limitation reduces their effectiveness for complex RNA analyses that require comprehensive feature representations.

Traditional feature extraction approaches, such as nucleic acid composition, pseudo-nucleotide composition, and autocorrelation have been widely used because of their effectiveness and simplicity

 doi:10.20944/preprints202508.1739.v1

in capturing sequence features. Tools like MathFeature expand on these foundations by incorporating innovative mathematical descriptors, such as chaos game theory, genomic signal processing, and entropy, which enable high accuracy across a variety of classification tasks [19]. On the other hand, some feature extraction pipelines, such as iLearnPlus [38] and the recently introduced R-based package ftrCOOL [88], have expanded the range of features by integrating physicochemical and structural descriptors alongside the aforementioned traditional approaches. ftrCOOL remarkably outperforms iLearnPlus in processing speed, making it a preferred choice for analysing large RNA datasets [88]

Although the field has made substantial progress, further efforts are required to address persistent challenges. Future feature extraction platforms should prioritise user-friendliness and computational efficiency, enabling both expert bioinformaticians and researchers with limited programming experience to perform advanced analyses. In addition, expanding the range of available feature descriptors and integrating broader analytical capabilities will be essential for improving model performance and reproducibility in RNA-focused machine learning studies.

Author Contribution: F.S.: Investigation, Formal analysis, Methodology, Data curation, Writing—Original Draft. F.V.: Conceptualization, Supervision, Investigation, Funding acquisition, Writing—Review & Editing. J.J.T.: Supervision, Funding acquisition, Writing—Review & Editing.

Acknowledgments: The authors acknowledge funding from the Australian Research Council Discovery Project (DP220101938) and the Australian Government Research Training Program (RTP) Scholarship.

References

1. Kukurba KR, Montgomery SB. RNA Sequencing and Analysis. Cold Spring Harb Protoc. 2015 Apr 13;2015(11):951–69.
2. Deshpande D, Chhugani K, Chang Y, Karlsberg A, Loeffler C, Zhang J, et al. RNA-seq data science: From raw data to effective interpretation. Front Genet. 2023;14:997383.
3. Mitić T, Caporali A. Emerging roles of non-coding RNAs in endothelial cell function. Current Opinion in Physiology. 2023 Aug 1;34:100672.
4. Chauvier A, Walter NG. Regulation of bacterial gene expression by non-coding RNA: It is all about time! Cell Chem Biol. 2024 Jan 18;31(1):71–85.
5. Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. Annu Rev Biochem. 2012;81:145–66.
6. Schmitt AM, Chang HY. Long noncoding RNAs in cancer pathways. Cancer cell. 2016 Apr 11;29(4):452-63.
7. van der Sluis F, van den Broek EL. Model interpretability enhances domain generalization in the case of textual complexity modeling. Patterns (N Y). 2025 Feb 6;6(2):101177.
8. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences. 2019 Oct 29;116(44):22071–80.
9. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. Nat Rev Mol Cell Biol. 2022 Jan;23(1):40–55.
10. Bonidia RP, Sampaio LDH, Domingues DS, Paschoal AR, Lopes FM, de Carvalho ACPLF, et al. Feature extraction approaches for biological sequences: a comparative study of mathematical features. Briefings in Bioinformatics. 2021 Sep 1;22(5):bbab011.
11. Gross B, Dauvin A, Cabeli V, Kmetzsch V, El Khoury J, Dissez G, et al. Robust evaluation of deep learning-based representation methods for survival and gene essentiality prediction on bulk RNA-seq data. Sci Rep. 2024 Jul 24;14(1):17064.
12. Hwang H, Jeon H, Yeo N, Baek D. Big data and deep learning for RNA biology. Exp Mol Med. 2024 Jun;56(6):1293–321.
13. Dias AL, Bustillo L, Rodrigues T. Limitations of representation learning in small molecule property prediction. Nat Commun. 2023 Oct 13;14(1):6394.
14. Ericsson L, Gouk H, Loy CC, Hospedales TM. Self-Supervised Representation Learning: Introduction, advances, and challenges. IEEE Signal Processing Magazine. 2022 May;39(3):42–62.

 doi:10.20944/preprints202508.1739.v1

15. Pan X, Yang Y, Xia CQ, Mirza AH, Shen HB. Recent methodology progress of deep learning for RNA–protein interaction prediction. WIREs RNA. 2019;10(6):e1544.
16. Pérez-Núñez JR, Rodríguez C, Vásquez-Serpa LJ, Navarro C. The Challenge of Deep Learning for the Prevention and Automatic Diagnosis of Breast Cancer: A Systematic Review. Diagnostics (Basel). 2024 Dec 23;14(24):2896.
17. Ding Z, Wang Z, Zhang Y, Cao Y, Liu Y, Shen X, Tian Y, Dai J. Trade-offs between machine learning and deep learning for mental illness detection on social media. Scientific Reports. 2025 Apr 25;15(1):14497.
18. Bonidia RP, Domingues DS, Sanches DS, De Carvalho AC. MathFeature: feature extraction package for DNA, RNA and protein sequences based on mathematical descriptors. Briefings in bioinformatics. 2022 Jan;23(1):bbab434.
19. Dou L, Li X, Ding H, Xu L, Xiang H. Prediction of m5C Modifications in RNA Sequences by Combining Multiple Sequence Features. Molecular Therapy—Nucleic Acids. 2020 Sep 4;21:332–42.
20. Guan ZX, Li SH, Zhang ZM, Zhang D, Yang H, Ding H. A Brief Survey for MicroRNA Precursor Identification Using Machine Learning Methods. Curr Genomics. 2020 Jan;21(1):11–25.
21. Kang Q, Meng J, Luan Y. RNAI-FRID: novel feature representation method with information enhancement and dimension reduction for RNA–RNA interaction. Briefings in Bioinformatics. 2022 May 1;23(3):bbac107.
22. Arceda VM. An Analysis of k-Mer Frequency Features with Machine Learning Models for Viral Subtyping of Polyomavirus and HIV-1 Genomes. InProceedings of the Future Technologies Conference 2020 Oct 31 (pp. 279-290). Cham: Springer International Publishing.
23. Kirk JM, Kim SO, Inoue K, Smola MJ, Lee DM, Schertzer MD, et al. Functional classification of long non-coding RNAs by k-mer content. Nat Genet. 2018 Oct;50(10):1474–82.
24. Lorenzi C, Barriere S, Villemin JP, Dejardin Bretones L, Mancheron A, Ritchie W. iMOKA: k-mer based software to analyze large collections of sequencing data. Genome biology. 2020 Oct 13;21(1):261.
25. Xu H, Jia P, Zhao Z. Deep4mC: systematic assessment and computational prediction for DNA N4-methylcytosine sites by deep learning. Brief Bioinform. 2020 Jun 24;22(3):bbaa099.
26. Zhang W, Shi J, Tang G, Wu W, Yue X, Li D. Predicting small RNAs in bacteria via sequence learning ensemble method. In2017 IEEE international conference on bioinformatics and biomedicine (BIBM) 2017 Nov 13 (pp. 643-647). IEEE.
27. Luo L, Li D, Zhang W, Tu S, Zhu X, Tian G. Accurate Prediction of Transposon-Derived piRNAs by Integrating Various Sequential and Physicochemical Features. PLoS One. 2016;11(4):e0153268.
28. Leslie C, Eskin E, Cohen A, Weston J, Noble W. Mismatch String Kernels for Discriminative Protein Classification. Bioinformatics (Oxford, England). 2004 Apr 1;20:467–76.
29. Li M, Liang C. LncDC: a machine learning-based tool for long non-coding RNA detection from RNA-Seq data. Sci Rep. 2022 Nov 9;12:19083.
30. Chen W, Tran H, Liang Z, Lin H, Zhang L. Identification and analysis of the N6-methyladenosine in the Saccharomyces cerevisiae transcriptome. Sci Rep. 2015 Sep 7;5(1):13859.
31. Muhammod R, Ahmed S, Md Farid D, Shatabda S, Sharma A, Dehzangi A. PyFeat: a Python-based effective feature generation tool for DNA, RNA and protein sequences. Bioinformatics. 2019 Oct 1;35(19):3831–3.
32. Hubert B. SkewDB, a comprehensive database of GC and 10 other skews for over 30,000 chromosomes and plasmids. Scientific Data. 2022 Mar 22;9(1):92.
33. Lu J, Salzberg SL. SkewIT: The Skew Index Test for large-scale GC Skew analysis of bacterial genomes. PLoS computational biology. 2020 Dec 4;16(12):e1008439.
34. Yuan GH, Wang Y, Wang GZ, Yang L. RNAlight: a machine learning model to identify nucleotide features determining RNA subcellular localization. Briefings in Bioinformatics. 2023 Jan 1;24(1):bbac509.
35. Liu B. BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. Briefings in Bioinformatics. 2019 Jul 19;20(4):1280–94.
36. Mathur M, Patiyal S, Dhall A, Jain S, Tomer R, Arora A, Raghava GP. Nfeature: A platform for computing features of nucleotide sequences. BioRxiv. 2021 Dec 16:2021-12.
37. Chen, Zhen, Pei Zhao, Chen Li, Fuyi Li, Dongxu Xiang, Yong-Zi Chen, Tatsuya Akutsu et al. iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization. Nucleic acids research 49, no. 10 (2021): e60-e60.

 doi:10.20944/preprints202508.1739.v1

38. Chen W, Lei TY, Jin DC, Lin H, Chou KC. PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. Anal Biochem. 2014 Jul 1;456:53–60.
39. Chen W, Zhang X, Brooker J, Lin H, Zhang L, Chou KC. PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. Bioinformatics. 2015 Jan 1;31(1):119–20.
40. Liu B, Liu F, Wang X, Chen J, Fang L, Chou KC. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic acids research. 2015 Jul 1;43(W1):W65-71.
41. Chen W, Lin H, Chou KC. Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. Mol BioSyst. 2015 Sep 15;11(10):2620–34.
42. 44. Liu B, Wu H, Chou KC. Pse-in-One 2.0: An Improved Package of Web Servers for Generating Various Modes of Pseudo Components of DNA, RNA, and Protein Sequences. Natural Science. 2017 Apr 28;9(4):67–91.
43. 45. Liu B, Liu F, Fang L, Wang X, Chou KC. repRNA: a web server for generating various feature vectors of RNA sequences. Mol Genet Genomics. 2016 Feb;291(1):473–81.
44. 46. Du PF, Zhao W, Miao YY, Wei LY, Wang L. UltraPse: A Universal and Extensible Software Platform for Representing Biological Sequences. Int J Mol Sci. 2017 Nov 14;18(11):2400.
45. 47. Tsonis AA, Elsner JB, Tsonis PA. Periodicity in DNA coding sequences: Implications in gene evolution. Journal of Theoretical Biology. 1991 Aug 7;151(3):323–31.
46. 48. Anastassiou D. Genomic signal processing. IEEE Signal Processing Magazine. 2001 Jul;18(4):8–20.
47. 49. Chakravarthy N, Spanias A, Iasemidis LD, Tsakalis K. Autoregressive modeling and feature analysis of DNA sequences. EURASIP Journal on Advances in Signal Processing. 2004 Jan 21;2004(1):952689.
48. 50. Harun-Or-Roshid Md, Pham NT, Manavalan B, Kurata H. Meta-2OM: A multi-classifier meta-model for the accurate prediction of RNA 2′-O-methylation sites in human RNA. PLoS One. 2024 Jun 26;19(6):e0305406.
49. 51. Zhang R, Zhang CT. Z curves, an intutive tool for visualizing and analyzing the DNA sequences. Journal of Biomolecular Structure and Dynamics. 1994 Feb 1;11(4):767-82.
50. 52. Yang YL. Study on the Specific ncRNAs Based on Z-curve Method. In2008 International Conference on MultiMedia and Information Technology 2008 Dec 30 (pp. 790-793). IEEE.
51. 53. Yang Y ling, Wang J, Yu JF, Liu G zhong. An Analysis of Non-Coding RNA Using Z-Curve Method. In 2008. p. 129–32.
52. 54. Zhang R, Zhang CT. A Brief Review: The Z-curve Theory and its Application in Genome Analysis. Curr Genomics. 2014 Apr;15(2):78–94.
53. 55. Wang L, Park HJ, Dasari S, Wang S, Kocher JP, Li W. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res. 2013 Apr 1;41(6):e74.
54. 56. Roymondal U, Das S, Sahoo S. Predicting Gene Expression Level from Relative Codon Usage Bias: An Application to Escherichia coli Genome. DNA Research. 2009 Feb 1;16(1):13–30.
55. 57. Bonidia RP, Sampaio LDH, Domingues DS, Paschoal AR, Lopes FM, de Carvalho ACPLF, et al. Feature extraction approaches for biological sequences: a comparative study of mathematical features. Briefings in Bioinformatics. 2021 Sep 1;22(5):bbab011.
56. 58. Hoang T, Yin C, Yau SST. Numerical encoding of DNA sequences by chaos game representation with application in similarity comparison. Genomics. 2016 Oct;108(3–4):134–42.
57. 59. Yin C, Chen Y, Yau SST. A measure of DNA sequence similarity by Fourier Transform with applications on hierarchical clustering. J Theor Biol. 2014 Oct 21;359:18–28.
58. 60. Löchel HF, Heider D. Chaos game representation and its applications in bioinformatics. Computational and Structural Biotechnology Journal. 2021 Jan 1;19:6263–71.
59. 61. Akhter S, Bailey BA, Salamon P, Aziz RK, Edwards RA. Applying Shannon's information theory to bacterial and phage genomes and metagenomes. Sci Rep. 2013;3:1033.
60. 62. Tenreiro Machado JA, Costa AC, Quelhas MD. Shannon, Rényie and Tsallis entropy analysis of DNA using phase plane. Nonlinear Analysis: Real World Applications. 2011 Dec 1;12(6):3135–44.

 doi:10.20944/preprints202508.1739.v1

61. 63. Tsallis C, Mendes RenioS, Plastino AR. The role of constraints within generalized nonextensive statistics. Physica A: Statistical Mechanics and its Applications. 1998 Dec 15;261(3):534–54.
62. 64. Yamano T. Information theory based on nonadditive information content. Phys Rev E. 2001 Mar 23;63(4):046105.
63. 65. Reuter JS, Mathews DH. RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics. 2010 Mar 15;11(1):129.
64. 66. Han S, Liang Y, Ma Q, Xu Y, Zhang Y, Du W, et al. LncFinder: an integrated platform for long non-coding RNA identification utilizing sequence intrinsic composition, structural information and physicochemical property. Briefings in Bioinformatics. 2019 Nov;20(6):2009.
65. 67. Sato K, Hamada M. Recent trends in RNA informatics: a review of machine learning and deep learning for RNA secondary structure prediction and RNA drug discovery. Brief Bioinform. 2023 May 25;24(4):bbad186.
66. 68. Mathews DH, Turner DH. Prediction of RNA secondary structure by free energy minimization. Current opinion in structural biology. 2006 Jun 1;16(3):270-8.
67. 69. Mathews DH, Sabina J, Zuker M, Turner DH. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol. 1999 May 21;288(5):911–40.
68. 70. Xia T, SantaLucia J, Burkard ME, Kierzek R, Schroeder SJ, Jiao X, et al. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry. 1998 Oct 20;37(42):14719–35.
69. 71. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proc Natl Acad Sci U S A. 2004 May 11;101(19):7287–92.
70. 72. DING Y, CHAN CY, LAWRENCE CE. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. RNA. 2005 Aug;11(8):1157–66.
71. 73. Lu ZJ, Gloor JW, Mathews DH. Improved RNA secondary structure prediction by maximizing expected pair accuracy. RNA. 2009 Oct;15(10):1805–13.
72. 74. Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms for Molecular Biology. 2011 Nov 24;6(1):26.
73. 75. List of RNA structure prediction software. In: Wikipedia [Internet]. 2025 [cited 2025 Aug 11]. Available from: title=List_of_RNA_structure_prediction_software&oldid=1305069840
74. 76. Xue C, Li F, He T, Liu GP, Li Y, Zhang X. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. BMC Bioinformatics. 2005 Dec 29;6(1):310.
75. 77. Wang M, Ali H, Xu Y, Xie J, Xu S. BiPSTP: Sequence feature encoding method for identifying different RNA modifications with bidirectional position-specific trinucleotides propensities. Journal of Biological Chemistry
76. 78. Liu B, Fang L, Liu F, Wang X, Chen J, Chou KC. Identification of Real MicroRNA Precursors with a Pseudo Structure Status Composition Approach. PLoS One. 2015 Mar 30;10(3):e0121501.
77. 79. Liu B, Fang L, Liu F, Wang X, Chou KC. iMiRNA-PseDPC: microRNA precursor identification with a pseudo distance-pair composition approach. Journal of Biomolecular Structure and Dynamics. 2016 Jan 2;34(1):223–35.
78. 80. de ON Lopes I, Schliep A, de LF de Carvalho AC. The discriminant power of RNA features for pre-miRNA recognition. BMC Bioinformatics. 2014 May 2;15(1):124.
79. 81. Jha T, Mendel J, Cho H, Choudhary M. Prediction of Bacterial sRNAs Using Sequence-Derived Features and Machine Learning. Bioinform Biol Insights. 2022 Jan 1;16:11779322221118335.
80. 82. Khalid R, Naveed H, Khalid Z. Computational prediction of disease related lncRNAs using machine learning. Sci Rep. 2023 Jan 16;13(1):806.
81. 83. Li J, Ju Y, Zou Q, Ni F. lncRNA localization and feature interpretability analysis. Mol Ther Nucleic Acids. 2024 Dec 12;36(1):102425.
82. 84. Tang G, Shi J, Wu W, Yue X, Zhang W. Sequence-based bacterial small RNAs prediction using ensemble learning strategies. BMC Bioinformatics. 2018 Dec 21;19(20):503.

 doi:10.20944/preprints202508.1739.v1

83. 85. Dong J, Yao ZJ, Wen M, Zhu MF, Wang NN, Miao HY, et al. BioTriangle: a web-accessible platform for generating various molecular representations for chemicals, proteins, DNAs/RNAs and their interactions. Journal of Cheminformatics. 2016 Jun 21;8(1):34.
84. 86. Liu B, Gao X, Zhang H. BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. Nucleic Acids Research. 2019 Nov 18;47(20):e127.
85. 87. Chen Z, Zhao P, Li F, Marquez-Lago TT, Leier A, Revote J, et al. iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. Briefings in Bioinformatics. 2020 May 18;21(3):1047–57.
86. 88. Amerifar S, Norouzi M, Ghandi M. A tool for feature extraction from biological sequences. Briefings in Bioinformatics. 2022 May 1;23(3):bbac108.
87. 89. Chou KC. Some remarks on protein attribute prediction and pseudo amino acid composition. Journal of theoretical biology. 2011 Mar 21;273(1):236-47.