Protein-Coding sORFs Prediction Based on U-Net and Coordinate Attention with Hybrid Encoding

Ziling Wang[1], Wenxi Yang[2], Zhijian Qu[3]
School of Computer Science and Technology, Shandong University of Technology, SDUT, Country Zibo, China[1]
School of Computer Science and Technology, Shandong University of Technology, Country Zibo, China[2, 3]

***Abstract*—Small proteins encoded by small open reading frames (sORFs) exhibit significant biological activity in crucial biological processes such as embryonic development and metabolism. Accurately predicting whether sORFs encode small proteins is a key challenge in current research. To address this challenge, many methods have been proposed, however, existing methods rely solely on biological features as the sequence encoding scheme, which results in high feature extraction complexity and limited applicability across species. To tackle this issue, we proposed a deep learning architecture UAsORFs based on hybrid coding of sORFs sequences. In contrast to mainstream prediction methods, this framework processes sORF sequences using a mixed encoding approach, including both one-hot and gapped k-mer encodings, which effectively captures global and local sequence information. Additionally, it autonomously learns to extract features of sORFs and captures both long-range and short-range interactions between sequences through U-Net and coordinate attention mechanisms. Our research demonstrates significant progress in predicting encoded peptides from eukaryotic and prokaryotic sORFs, particularly in improving the cross-species predictive MCC index on the eukaryotic dataset.**

***Keywords*—Small open reading frames; deep learning; hybrid coding; U-Net; coordinate attention**

I. INTRODUCTION

With the rapid development of transcriptomic and proteomic technologies, researchers have come to better understand the potential coding regions of the genome [1]. Open reading frames (ORFs) are widely recognized as important sequence regions for protein coding [2], but short open reading frames (sORFs), as a group of ORFs up to 300 nucleotides in length, which were previously considered unlikely to code due to their short length [3, 4]. However, recent studies have shown that sORFs have a wide range of biological functions and can be directly transcribed and translated into biologically active small proteins [5-8]. These small proteins are involved in a variety of biological processes, including embryonic development [9], muscle function [5, 10-12], the regulation of cell growth and development [13-16], and the control of metabolic pathways [17]. Researchers have identified several sORFs with ribosomal activity by using versatile histological sequencing techniques, such as mass spectrometry and ribosome profiling [7, 18-20].

A prerequisite for the search for new protein-coding sORFs is their correct identification. Due to the short length, low expression levels, and lack of experimental validation for their functionality, sORFs have long been insufficiently annotated and studied. Investigating the coding potential of sORFs for small proteins is complex. Therefore, there is an urgent need for accurate and rapid methods to predict the coding ability of sORFs.

One of the main problems in predicting the microproteins-coding sORFs is to design an effective biological sequences coding scheme. The coding features are also crucial for distinguishing between coding and non-coding sORFs. Biological sequence coding schemes can be mainly divided into two types: sequential models and discrete models. Sequential models assign numerical values to each nucleotide in the biological sequence while preserving the order of the bases [1]. A prominent example of this is one-hot encoding (also known as C4 coding) [21], where each of the four nucleotides is represented by a unique four-bit binary vector (A-[1,0,0,0], C-[0,1,0,0], etc.). Each nucleotide's binary number is orthogonal to each other and has the same Hamming distance. In contrast, discrete models aim to design a set of features based on knowledge from the biological sequence. Some commonly used biological features include the codons usage [22], codon prototype [23], hexamer usage [24] and Z curves [25].

The sequential and discrete encoding models each present distinct advantages and limitations. While the sequential model preserves the global sequences order information [26], but this approach cannot fully capture biological features. Neural networks are not easily able to learn higher-order correlations from very low-level input [27]. Additionally, one-hot coding is unable capture frequency domain features such as k-mer [28]. On the other hand, taking 3-mer as an example, it is a discrete model of biological sequences and has become one of the features used to distinguish small proteins from non-encoding ones. Although the 3-mer is effective, it can only incorporate the local sequences order information between neighboring nucleotides and cannot reflect the global sequences order information [26].

Therefore, we designed a coding-protein prediction tool, named UAsORFs, utilizing a hybrid encoding strategy. This tool incorporates U-Net and Coordinate Attention (CA) mechanisms within its deep learning framework. This method effectively utilizes global sequence information, non-overlapping gapped k-mers and deep learning to autonomously learn sORF sequence features. By employing hybrid encoding, the method effectively extracts global and local sequence information of sORFs. Additionally, neural networks are employed to automatically differentiate between encoding and non-encoding sORFs.

The main contributions of this article are summarized as follows:

- A hybrid coding scheme combining sequence model and discrete model is designed. Unlike previous prediction tools that only use discrete models to extract sequence features, UAsORFs introduces sequence coding, effectively capturing global sequence information with one-hot coding, thereby enriching the encoding scheme and expression level.
- By exclusively using gkm as the discrete biological feature, excessive manual feature extraction is reduced, and gkm features significantly improve predictive performance on prokaryotic datasets.
- U-Net is used for the first time in the prediction of coding-protein sORF, facilitating the extraction of multi-scale, long-range and short-range interaction features of sORF sequences. A combined model (UCA) of U-Net and CA is constructed, leading to improved cross-species prediction results.

II. Related Work

In numerous biological prediction tasks, combining sequence information with biological features can lead to significant performance enhancements in specific applications. For instance, iTIS-PseTNC [26] introduced a sequence-based predictor for identifying translation initiation sites in human genes and claims that using k-mer representation in DNA sequences only reflects local sequence order information while losing global sequence order information. To overcome this limitation, the approach leveraged collaborative representations known as pseudo trinucleotide assemblies, integrating physicochemical properties into DNA sequences alongside k-mer features [26]. MHCDG [29] is a hybrid sequence-based deep learning model that integrates MeDIP-seq data with histone information to predict DNA methylation CpG status. By incorporating multiple biologically relevant features and sequence information, it outperformed other methods achieves more satisfactory promoter prediction performance. These works demonstrate the importance of hybrid coding.

In the prediction work of protein-coding sORFs, many prediction tools solely utilize various biological features for coding schemes. Among them, MiPepid [30] identified protein-coding sORFs using a logistic regression model and tetramer (4-mer) features from sequences. CPPred [31] is an SVM-base classifier that uses 38-dimensional biological features such as ORF coverage, Fickett score and CTD, etc., to predict the coding potential in both regular ORFs and sORFs. CPPred-sORF [32], based on CPPred coding, incorporates additional features such as GC count and mRNN-11 codons, and evaluates sORFs using non-AUG start codons. PsORFs [33] predict protein-coding sORFs in other species using 64 codon frequencies based on a random forest model trained on sORFs from prokaryotes. CodingCapacity [34] predicts the coding potential of sORFs using the Z-curve, codon frequencies, k-mer, and all features included in CPPred-sORF. DeepCPP [35] use a 589-dimensional feature set composed of nucleotide bias information and minimal distribution similarity. Notably, DeepCPP employs a convolutional neural network based on deep learning for prediction.

All the aforementioned work gives us a strong intuition that combining global sequence order information with biological features (such as gkm [36]) can enhance the prediction of protein-coding sORFs.

III. Materials and Methods

The prediction problem of protein-coding sORFs aims to determine whether an sORF has the ability to be transcribed and translated into a small protein. Given an sORF sequence $s = s_1 s_2 ... s_n$, the label of the sequence is $y = i, i \in \{1, 0\}$, i represents coding (1) or noncoding (0). Our goal is to convert the original sequences into a computer-recognizable format and predict it as either coding or noncoding sORFs using a deep learning framework.

A. Data Description

Our study aimed to establish a model for predicting the coding potential of sORFs in multiple species, covering both prokaryotes and eukaryotes. We utilized the standard dataset from PsORFs, which was generated based on a random order strategy. The same prokaryotic training dataset Pro-1282 and five test datasets (Hum-7111, Mou-7385, Ara-2125, Pro-6318 and Bac-150) from PsORFs were employed. Prokaryotic sORFs were selected from the Ref-Seq database [37]. Whereas human and mouse sORFs were downloaded from the sORFs.org database [38], while Arabidopsis thaliana sORFs were obtained from the TAIR database [39]. An experimental validation dataset (Bac-150) published by Hemm et al [40], included 150 positive sORFs and 53 negative sORFs detected from the E. coli genome. The detailed information of the datasets is presented in Table I, where coding sORFs refer to sequences capable of being translated into small proteins.

TABLE I. Number of Datasets for Each Species

Dataset	Species	Number of coding sORFs	Number of non-coding sORFs	Number of sORFs
Hum-7111	Prokaryotic genomes	7111	7111	14222
Mou-7385	Mouse	7385	7385	14770
Ara-2125	Arabidopsis thalian	2125	2125	4150
Pro-6318	Prokaryotic genomes	6318	6318	12645
Pro-1228	Prokaryotic genomes	1228	1327	2556
Bac-150	E.coli genome	150	53	203

The length distributions of sORFs across six datasets are illustrated in Fig. 1. The length distributions of sORFs in the Hum-7111 and Mou-7385 dataset are very similar, predominantly concentrated within the range of 60 to 140 nucleotides. In contrast to mammals, Arabidopsis, which is also a eukaryotic organism, exhibits a different distribution, with sORF lengths mainly distributed between 200 and 300 nucleotides. The length distributions of sORFs in the prokaryotic datasets Pro-6318 and Pro-1282 was similar, primarily concentrated between 150 and 300 nucleotides. Thus, it is evident that there are significant differences in sORF lengths among different species.

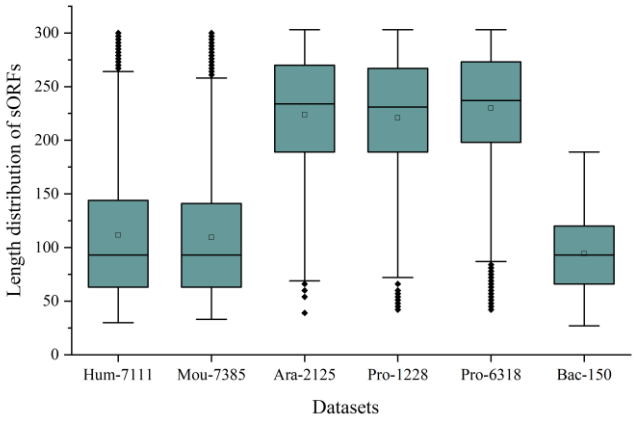

Fig. 1. Distribution of sORFs length of six datasets.

The differences in sORFs lengths among different species may pose challenges in predicting protein-coding sORFs in multiple species. sORFs with short lengths may be overlooked or misinterpreted as noise in certain species, thereby increasing the difficulty of prediction. Additionally, significant differences may exist in the sequence characteristics and preferences of sORFs across different species, making it challenging to identify universal features and patterns for predicting protein-coding sORFs across species. To address these challenges, the consideration of deep learning models is warranted, as deep learning offers advantages such as automatic feature learning, strong generalization capabilities, and flexibility.

B. Overview of the Designed Framework

As illustrated in Fig. 2, the overall workflow of UAsORFs comprises three stages. Firstly, the sORF sequences in the fasta file are preprocessed and encoded into one-hot and gkm coding formats. Subsequently, the one-hot coding is fed into the neural network UCA, which is composed of a U-Net and CA that can learn the importance of each nucleotide in the sequence. The resulting one-hot coding processed by the UCA model is concatenated with gkm coding to form a hybrid encoding representation. Finally, the hybrid encoding undergoes processing through a Multilayer Perceptron (MLP) and SoftMax activation function to obtain the final prediction results.

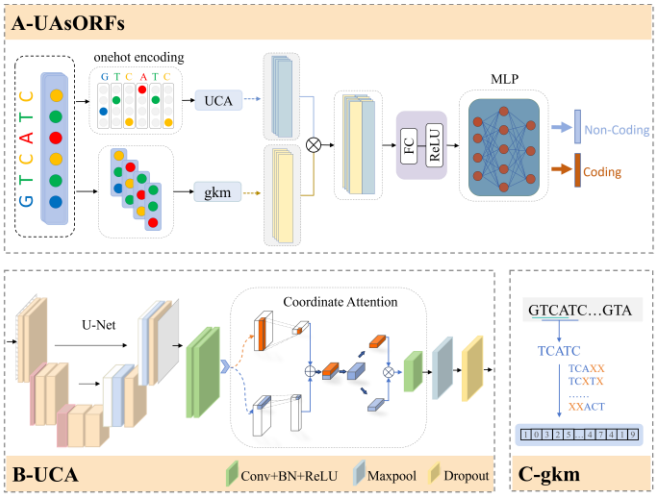

Fig. 2. The overall workflow of UAsORFs.

C. Hybrid Coding

To overcome the limitations of sequential models and discrete models, we propose a hybrid encoding scheme that combines global sequence models and biological features. The aim is to fully utilize the advantages of both, thereby enhancing the prediction of protein-coding sORFs. As illustrated in Fig. 3, For a given sequence $_S$, the hybrid encoding scheme can be formulated as follows:

$$E=[E_1,E_2] \tag{1}$$

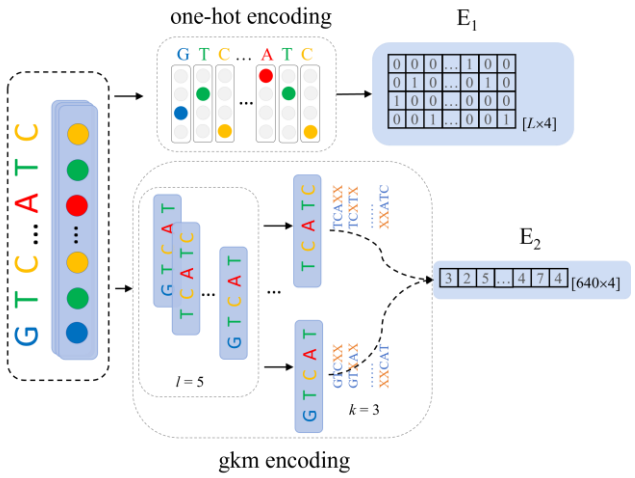

Fig. 3. Hybrid encoding scheme. E_1 is one-hot sequential model, E_2 is gkm discrete model.

Here, we utilize the one-hot sequential model [41] to capture the global sequence order information, Therefore, E_1 can be expressed as:

$$E_1=[x_1,x_2,...,x_i] \tag{2}$$

$$x_i=\begin{cases}[1,0,0,0], if\ i='A',\\ [0,1,0,0], if\ i='T',\\ [0,0,1,0], if\ i='G',\\ [0,0,0,1], if\ i='C',\\ [0,0,0,0], otherwise.\end{cases} \tag{3}$$

During one-hot encoding, we employ binary vectors of length 4 to encode the four nucleotides in biological sequences. Specifically, the position corresponding to a base is represented as 1, while the other positions are represented as 0. Additionally, when dealing with shorter sequences, we use the encoding [0,0,0,0] for sequence padding to maintain consistency in sequence length. Once a small protein sequence N of length i is inputted, a feature vector matrix is obtained to be fed into the model for training.

For the discrete model, we employ non-overlapping gkm [36] to capture local sequences order information effectively. K-mer, as a classic and effective feature representation, have been widely used in the field of bioinformatics, notably in the prediction of protein coding regions [42-44], coding potentials [31, 45], and identification of regulatory elements [29, 41].

Nevertheless, traditional k-mer methodologies are constrained by a pivotal issue that the increase of k leads to a very long and sparse feature vector [36]. To overcome this issue, we introduce the concept of gaps, which allows for certain mismatch exist in the k-mer sequence [36]. Gkm not only effectively reduce the dimensionality and sparsity of the feature vector but also demonstrates outstanding predictive prowess over conventional k-mer approaches, as evidenced by multiple studies in the biological domain [36, 46].

The gkm [36] method has two parameters: the length of the whole word l and the number of informative positions k. Therefore, the gap count is $l-k$. Combining previous work [1] and the dimension range of one-hot feature vectors, which is [100, 1200], we set $l=5$ and $k=3$. This not only effectively reduces the feature vector's dimensionality from 451,024 to $C_5^2 4^3 = 640$, which is close to the dimension of the one-hot feature vectors, but also allows using both together to enhance the ability to learn relevant patterns. This benefits the deep learning model by improving its expressive power and predictive accuracy. But also encompasses non-overlapping 3-mers information (e.g., AAAXX, ..., TTTXX). As shown in Fig. 3, when $l=5$ the length of each sORF subsequence is 5. For $k=3$, calculate the frequency of occurrence of each subsequence with three consecutive nucleotides. Thus, E_2 can be expressed as:

$$E_2 = [f(XXAAA), f(XAXAA), ..., f(TTTXX)] \quad (4)$$

Where $f(XXAAA)$ calculates the frequency of non-overlapping gapped trinucleotides (XXAAA) occurring in biological sequences. By introducing two gaps XX, the two words $GTACA$ and $CTACA$ of length 5 have the same gapped trinucleotides $XTXCA$.

D. UCA Model

Considering the hybrid coding approach used for sORF sequences, effectively integrating global sequence order information and gkm through deep learning, and autonomously learn features of sORFs of different lengths in different species is a problem that needs to be addressed. To this end, as illustrated in Fig. 2, we propose a UAsORFs architecture aimed at addressing this issue. In the UAsORFs architecture, sORF sequences are first encoded into one-hot coding and gkm feature representations. These are then processed through the UCA module and concatenated with gkm encoding.

While the issue of hybrid encoding has been discussed in previous sections, the focus of this section is to provide a detailed explanation of the UCA network module. The UCA module mainly consists of two key components: the U-Net [47-49] and the CA [50] based on convolutional neural network (CNN).

We are the first apply U-Net to the prediction of protein-coding sORFs in order to extract multi-scale, long-range and short-range interaction features from input sequences. This design is intended to ensure that the network can capture sufficient contextual information for longer sequences and maintains effective representation capabilities for shorter sequences.

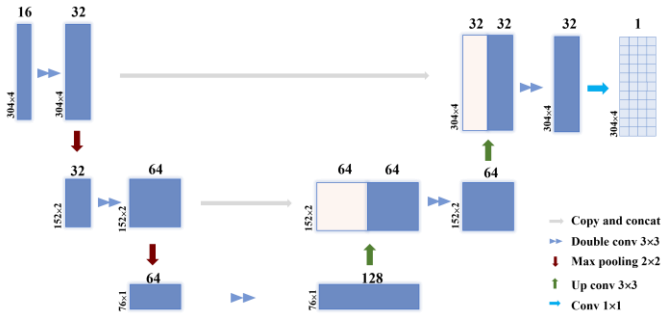

Fig. 4. U-Net model.

As shown in Fig. 4, the input of the U-Net network is denoted as E_1 after a convolution operation, while the output is represented as Y. The numbers at the top of the image represent the number of channels, while the lower-left corner shows dimensions as length × width. The kernels for Double conv and Up conv are 3 × 3, Conv uses a 1 × 1 kernel, and the max-pooling layer uses a 2 × 2 kernel. Through convolutional operations, $E_1 \in \{0,1\}^{304\times4}$ is transformed into $\in \{0,1\}^{16\times304\times4}$, where K can be regarded as an image that contains 16 color channels with a size of 304×4. Each channel signifies one of 16 possible nucleotide combinations, enabling the model to explicitly consider long-range interactions among nucleotides. Following processing by the U-Net, the output feature layer Y, $Y \in R^{16\times304\times4}$ forms a matrix of size $16 \times 304 \times 4$. Compared with the original one-hot coding, feature map Y can capture more abundant sORF sequence feature information, extract local patterns, context relations and higher-level semantic information in the sequence after multi-layer convolution and pooling layer processing of U-Net. It can provide powerful support for the next prediction work.

Next, for CA-based CNN, in image processing, CNN plays a crucial role in image processing, enabling the learning and extraction of effective image features. In the feature extraction phase, we employ a series of layer structures including Convolutional Layer, Batch Normalization (BN), Revised Linear Unit Activation Function (ReLU), Coordinate Attention, Max-Pooling and Dropout operation. The collaboration between these layer structures helps to achieve effective feature extraction and characterization of the input data.

In the traditional convolutional pooling process, applying channel attention mechanisms like Squeeze-and-Excitation (SE) attention [51] can assess the importance of each channel to learn the weights of different channel features [50]. However, the SE attention only considers encoding inter-channel information but neglects the importance of positional information, especially in cases of global sequential encoding (such as one-hot coding). The CA mechanism considers both inter-channel relationships and positional information within the feature space. Such a mechanism effectively focuses on different spatial locations of the input feature maps, enhancing the model's perception of key features and aiding in the better learning of useful features by the network model.

The CA Block is divided into two processes: Coordinate Information Embedding and Coordinate Attention Generation, as illustrated in Fig. 5.

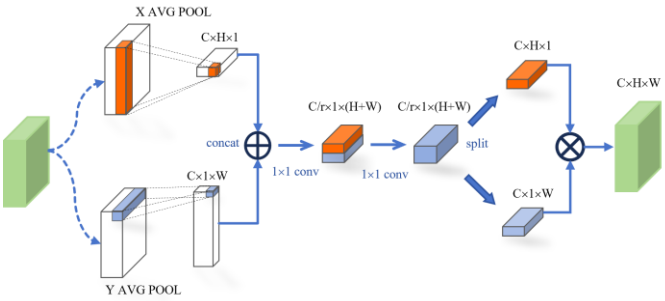

Fig. 5. Coordinate attention model.

$$y = \sigma(W_c \times f(W_g \times x)) \times x \tag{5}$$

Where x denotes the input feature map, W_c and W_g are the weight matrices of generation attention and channel attention, × denotes element-by-element multiplication, σ denotes the sigmoid function, and f denotes the weight coefficients of channel attention. Through the generative attention mechanism, a weight coefficient is learned for each spatial location. This weight coefficient is multiplied element-by-element with the input feature map to obtain a weighted feature map. Then, a weight coefficient is learned for each channel through the channel attention mechanism.

This weight coefficient is multiplied element-by-element with the weighted feature map and is used to weight the different channels of the feature map. This allows the network to pay more attention to the important channel information and suppress the unimportant channels to extract more effective feature representations and get the final output feature map.

IV. Training and Evaluation

A. Loss Function

To enhance the generalization ability and robustness against noise ability of the model, we use Label Smoothing (LS) loss as the loss function. The Label Smoothing loss function reduces the risk of overfitting and overconfidence by introducing a certain degree of smoothness. Its formula is as follows:

$$L = (1-\varepsilon) \times CE(y, y') + \varepsilon \times CE(\mu, y') \tag{6}$$

$$CE(p,q) = -\sum(p_i \times \log(q_i)) \tag{7}$$

Where y is the true label, y' is the output label probability distribution of the model, u is the smoothed label, ε is the smoothing factor, and CE is the cross-entropy loss function. In the loss function, we multiply the loss of the true labels by $(1-\varepsilon)$, multiply the loss of the smoothed labels by ε, and then weight and sum the two portions to get the final loss value. $\varepsilon > 0$ the loss portion of the smoothed labels will play a certain role of regularization, which helps to reduce the risk of overfitting.

B. Evaluation Indicators

We adopted four evaluation metrics, including Sensitivity (SN), Specificity (SP), Accuracy (ACC), and Matthews Correlation Coefficient (MCC), to evaluate the robustness of the model and its predictive performance for encoding sORFs. The formulas are as follows:

$$SN = \frac{TP}{TP+FP} \tag{8}$$

$$SP = \frac{FP}{TP+FP} \tag{9}$$

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{10}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)}\sqrt{(TP+FN)}\sqrt{(TN+FP)}\sqrt{(TN+FN)}} \tag{11}$$

Among them, TP and TN represent the number of correctly predicted coding and coding sORFs. FP and FN denote the number of incorrectly predicted coding and non-coding sORFs. SN and SP measure the model's ability to identify coding and non-coding sORFs. ACC reflects the proportion of correct predictions among all predictions. On the other hand, MCC comprehensively considers the relationships among TP, TN, FP, and FN, evaluating the correlation between predictions and annotations within the range of [-1, 1]. This metric system provides a comprehensive assessment of the model performance.

C. Training Parameter Settings

During training, we employed a learning rate decay strategy to prevent overfitting and accelerate the convergence of the learning algorithm. The initial learning rate was set to 4e-3, step size = 10 and gamma = 0.7. The Adam optimizer was chosen for parameter adjustment and optimization. In addition, set the batch size to 256, the number of epochs to 20 and the random seed to 42. Table II provides the UCA network architecture and parameters, with the U-Net and CA network architectures illustrated in Fig.4 and Fig.5. Our experimental environment consists of a CPU: AMD Ryzen 7 5800H and a GPU: NVIDIA GeForce RTX 3060.

TABLE II. UCA Network Architecture and Parameterisation

Layer	Size
Input	16×304×4
U-Net	16×304×4
CA	16×304×4
Conv	16×(3,3)
Maxpool	(2,2)
CA	16×152×2
Dropout	0.1
Conv	6×(3,3)
Maxpool	(2,1)
Dropout	0.1
Flatten	912
Concat	1096(912+640)
Linear	50
Linear	2
Softmax	2

V. RESULTS

In this section, we conducted four experiments on six datasets. The first demonstrated the importance of hybrid coding. The second and third experiments compared our approach with current popular methods on multiple species datasets. The fourth experiment demonstrated the importance of each module of UAsORFs.

A. Significance of Hybrid Coding

To validate the effectiveness of hybrid coding, we conducted ablation experiments aimed at separating hybrid coding and observing the performance of single coding in sORFs prediction. Specifically, as shown in Table III and Fig. 6, the hybrid coding was divided into one-hot and gkm encoding, which were then fed into MLP and CNN models (e.g., one-hot + CNN and gkm + MLP). We uesd a training and evaluation strategy trained on the Pro-6318 dataset and tested on the Ara-2125 (Fig. 6A) and Pro-1228 (Fig. 6B) datasets to evaluate the performance of hybrid coding on both eukaryotic and prokaryotic datasets.

TABLE III. DESIGN AND RESULTS OF ABLATION EXPERIMENTS FOR HYBRID CODING

Dataset	Methods	SN	SP	ACC	MCC
Ara-2125	PsORFs	0.4706	0.8613	0.6974	0.443
	one-hot+CNN	0.4815	0.9562	0.7188	0.4973
	k-mer+MLP	0.6918	0.8216	0.7267	0.5078
	gkm+MLP	0.5153	0.9501	0.7327	0.5168
	one-hot+k-mer+CNN	0.5976	0.9082	0.7529	0.5322
	one-hot+gkm+CNN	0.5962	0.9205	0.7584	0.5462
	one-hot+gkm+UAsORFs	0.6316	0.9031	0.7682	0.5571
Pro-1228	PsORFs	0.8698	0.8997	0.8908	0.7814
	one-hot+CNN	0.6458	0.9525	0.8051	0.6333
	k-mer+MLP	0.798	0.8432	0.8215	0.6424
	gkm+MLP	0.8739	0.9546	0.9142	0.8311
	one-hot+k-mer+CNN	0.8278	0.8964	0.8705	0.7418
	one-hot+gkm+CNN	0.8772	0.9584	0.923	0.8482
	one-hot+gkm+UAsORFs	0.9137	0.9435	0.9292	0.8582

According to the data presented in Table III and Fig. 6D, one-hot+gmk+CNN and one-hot+k-mer+CNN outperform one-hot+CNN, gmk+MLP, and k-mer+MLP on both the prokaryotic and eukaryotic datasets, which indicates that the hybrid coding scheme has a better prediction than the single coding model. It is worth noting that gmk+MLP achieved better predictive performance than k-mer+MLP, especially on the prokaryotic dataset, where ACC and MCC are improved by 9.27% (0.9142-0.8215) and 18.87% (0.8311-0.6424) , demonstrating the effectiveness of gkm (l=5, k=3) features in distinguishing coding and non-coding fields regions. Meanwhile, one-hot+CNN has outperformed PsORFs in Ara-2125, which proves the effectiveness of global order information.

In conclusion, our results suggest that there is a complementary relationship between one-hot coding and gkm features, and their combination helps deep learning methods to capture coding features more comprehensively.

B. Multi Species Predictions Results

To evaluate the performance of the different models, we conducted two experiments: (a) Training on the prokaryotic dataset Pro-1282 and testing on the remaining five datasets (Hum-7111, Ara-2125, Mou-7385 and Pro-6318, Bac-150). (b) Training on the prokaryotic dataset Pro-6318 and testing on the remaining five datasets.

Since there is no overlap of sequences in the test and training datasets, the multi-species validation is considered as independent dataset testing. With these two multi-species experiments, we evaluate the generalization performance of the model in multi-species prediction. These experimental designs help to validate the model's ability to generalize across different biological species and provide an important reference for further model improvement.

In our study, we evaluated seven different computational algorithms, some of which have been tested in the original literature on sORF [33]. As shown in Table IV，tools such as codingCapacity, PsORFs, CPPred-sORF, and DeepCPP employ discrete encoding schemes based solely on biological features. In contrast, UAsORFs enhances this discrete encoding framework by incorporating one-hot encoding, thereby increasing the representational capacity of sORFs. Furthermore, U-Net can integrate one-hot encoded features with spatial features, improving the accuracy of identifying specific regions or categories within sORFs data and enhancing segmentation performance, thus capturing details that other tools might overlook. The CA mechanism can emphasize crucial channels within the one-hot encoded data, ensuring that key classification information is prioritized and enabling the capture of significant features that might be missed by other tools.

As can be seen in Fig. 7 and 8, UAsORFs showed improvements across various independent datasets, particularly in terms of ACC and MCC metrics. On eukaryotic datasets, UAsORF achieved the highest ACC and MCC, surpassing the best-performing tool codingCapacity. In the Mou-7385 dataset, our method exhibited increases in ACC and MCC by 2.38% (0.5935-0.5697) and 7.57% (0.2398-0.1641). Similar improvements were observed in the Hum-7111 and Ara-2125 datasets. In the Pro-6318 dataset, UAsORF outperformed PsORFs but slightly lagged behind codingCapacity. On the Bac-150 dataset, UAsORF saw improvements in ACC and MCC by 2.15% (0.79-0.7685) and 8.96% (0.4715-0.3819).

From Fig. 8, it is evident that experiment (b) achieved better predictive performance compared to the training evaluation strategy of experiment (a). On the Bac-150 dataset, compared to PsORFs, UAsORFs showed approximately 5.12% increase in ACC (0.8-0.7488) and approximately 27.09% increase in MCC (0.5764-0.3055). With an increase in the number of training samples, UAsORFs more effectively captured sORF sequence features, resulting in better predictive performance. This also underscores the importance of constructing high-quality training and evaluation data.

In summary, our study demonstrates that UAsORFs exhibit strong generalization capabilities, showcasing excellent cross-

species predictive ability, and demonstrating their capacity to distinguish coding sORFs from non-coding ones. Furthermore, by altering training and evaluation strategies, we further validate the outstanding performance of the UAsORFs model in cross-species prediction.

C. UAsORFs Ablation Experiments

To investigate the robustness and reliability of the UAsORFs model, we conducted a series of ablation experiments. As shown in Table V, we systematically remove various components from the UAsORFs model, including the U-Net, CA and other modules. We utilized the Pro-1228 dataset for training and compared their predictive performance on both the eukaryotic (Hum-7111) and the prokaryotic (Pro-6318) dataset.

Fig. 9A and Fig. 9B show a performance comparison of the ACC and MCC across three eukaryotic test datasets (Hum-7111, Ara-2125, and Mou-7385) and two prokaryotic test datasets (Pro-6318 and Bac-150) under different methods. Fig. 9C and Fig. 9D show a performance comparison of ACC and MCC between Base, CA+LS and joining U-Net block on multi-species test datasets. Performance comparison of ACC index and MCC index on multi-species test dataset by adding CA(Fig. 9E), LS(Fig. 9F) block with Base et al.

From Table V and Fig. 9, it is evident that compared with the Base version, after adding U-Net, CA, and LS modules, the UAsORFs model improves the ACC(Fig. 9A) and MCC (Fig. 9B)to 62.39% and 29.59% on the eukaryotic dataset, and 91.7% and 83.67% on the prokaryotic dataset respectively. Specifically, in Fig. 9C and Fig. 9D, compared Base and CA+LS, the inclusion of the U-Net module in UAsORFs lead to significant improvements in ACC and MCC, increasing by 2.19% and 3.36% on the Hum-7111 test dataset. Fig. 9E and Fig. 9F demonstrate the enhancement in predictive performance after incorporating the CA and LS modules. The experimental results demonstrate the effectiveness of using U-Net, CA, and LS modules for extracting sORFs features.

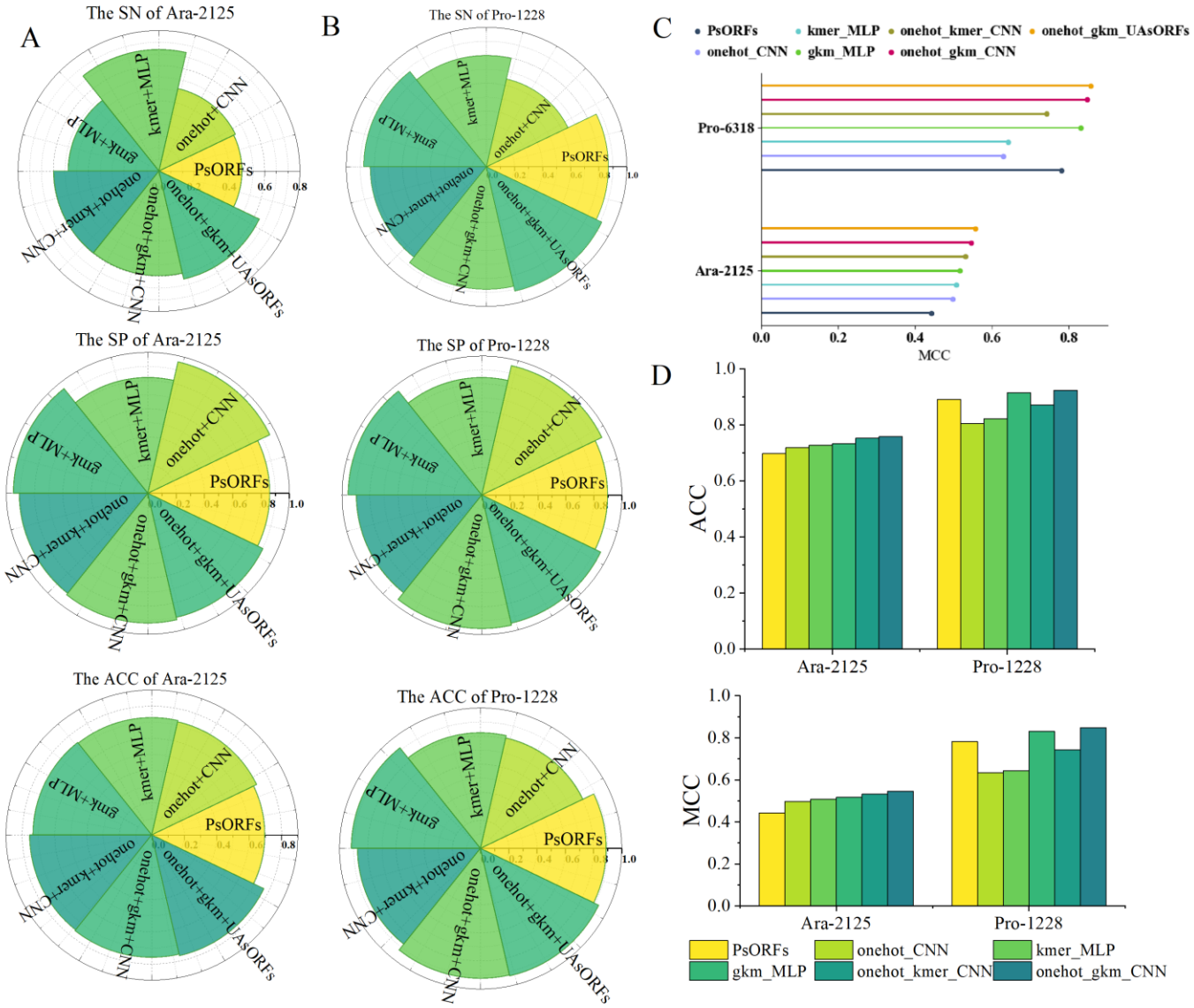

Fig. 6. Performance comparison of hybrid coding.

TABLE IV. COMPARISON OF UASORFS WITH SEVERAL OTHER PREDICTION TOOLS

Method	Year	Feature	Model
CPPred	2019	ORF length, ORF coverage, ORF integrity, Fickett score, Hexamer score, PI, Gravy, instability, CTD features	SVM classifier
MiPepid	2019	4-kmer	logistic regression
CPPred-sORF	2020	GCcount, mRNN-11codons and all features used by CPPred	SVM classifier
DeepCPP	2020	maximum ORF length, mean hexamer score, k-mer, ORF coverage, Fickett score, g-gap and nucleotide bias	CNN
PsORFs	2021	Codon frequency	Random forest
codingCapacity	2023	z-curve, codon frequency, k-mer and all features used by CPPred-sORF	Random forest
Our-method	2024	gmk	U-Net,CA

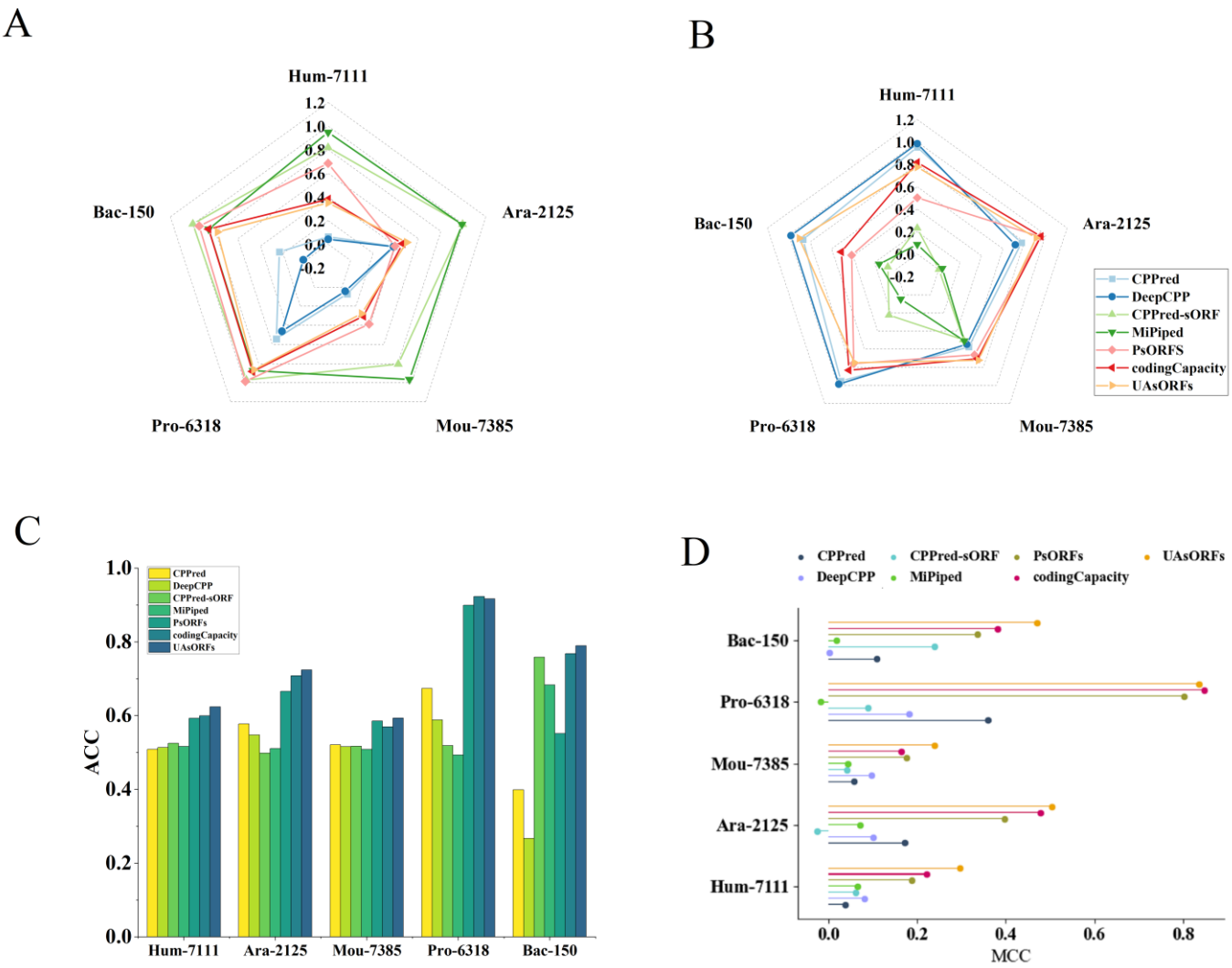

Fig. 7. The result of SN(A), SP(B), ACC(C)and MCC(D)for experiment (a).

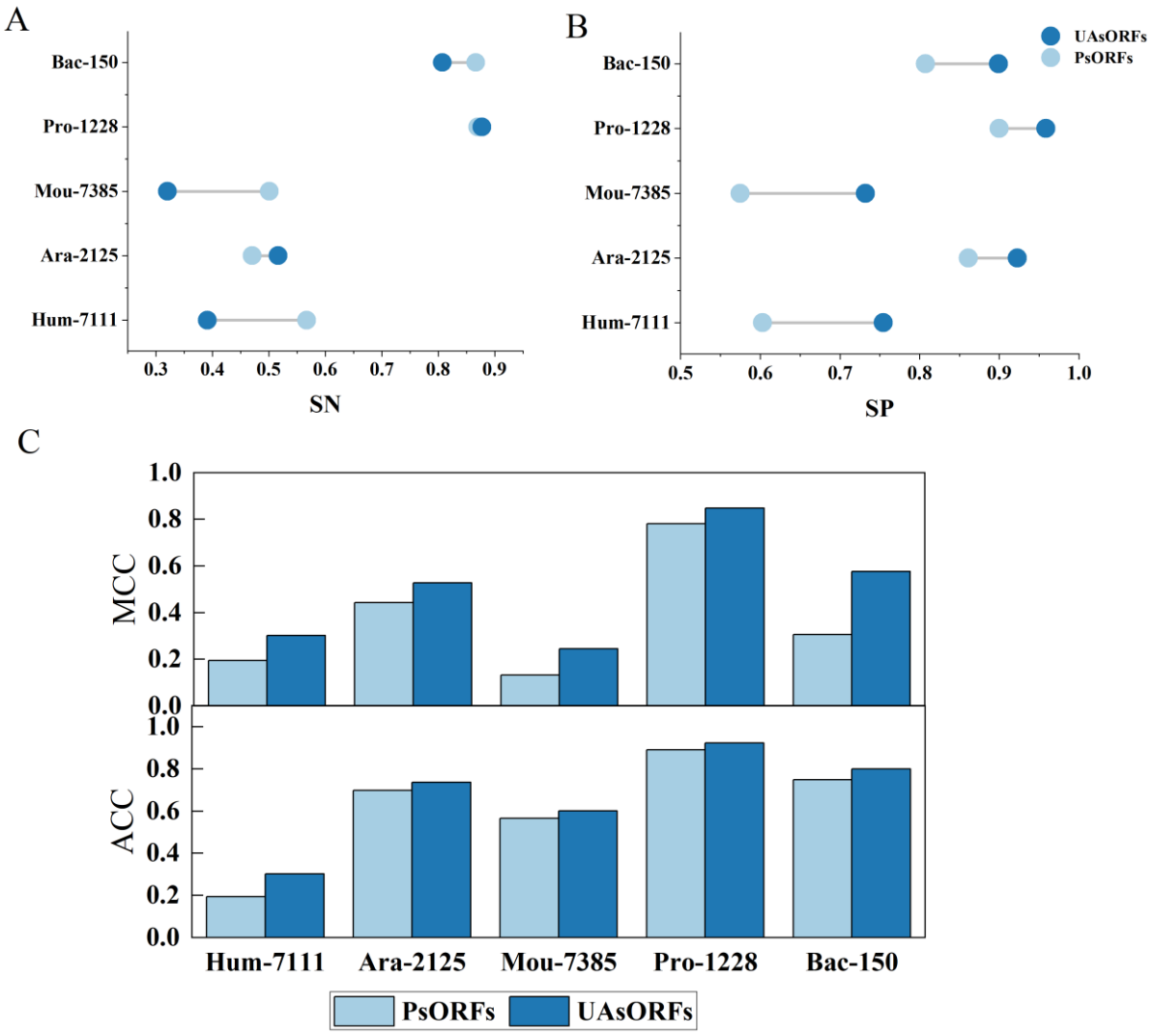

Fig. 8. The results of SN(A), SP(B), ACC and MCC(C) for PsORFs and UAsORFs using Pro-6318 train dataset.

TABLE V. DESIGN AND RESULTS OF ABLATION EXPERIMENTS FOR HYBRID CODING

Dataset	Method	U-Net	CA	LS	ACC	MCC
Hum-7111	Base	-	-	-	0.5719	0.1719
	CA+LS	-	√	√	0.6020	0.2623
	U-Net+LS	√	-	√	0.5679	0.1773
	U-Net+CA	√	√	-	0.5945	0.2301
	U-Net+CA+LS	√	√	√	0.6239	0.2959
Pro-6318	Base	-	-	-	0.8900	0.7864
	CA+LS	-	√	√	0.9094	0.8033
	U-Net+LS	√	-	√	0.8960	0.7983
	U-Net+CA	√	√	-	0.9156	0.8322
	U-Net+CA+LS	√	√	√	0.9170	0.8367

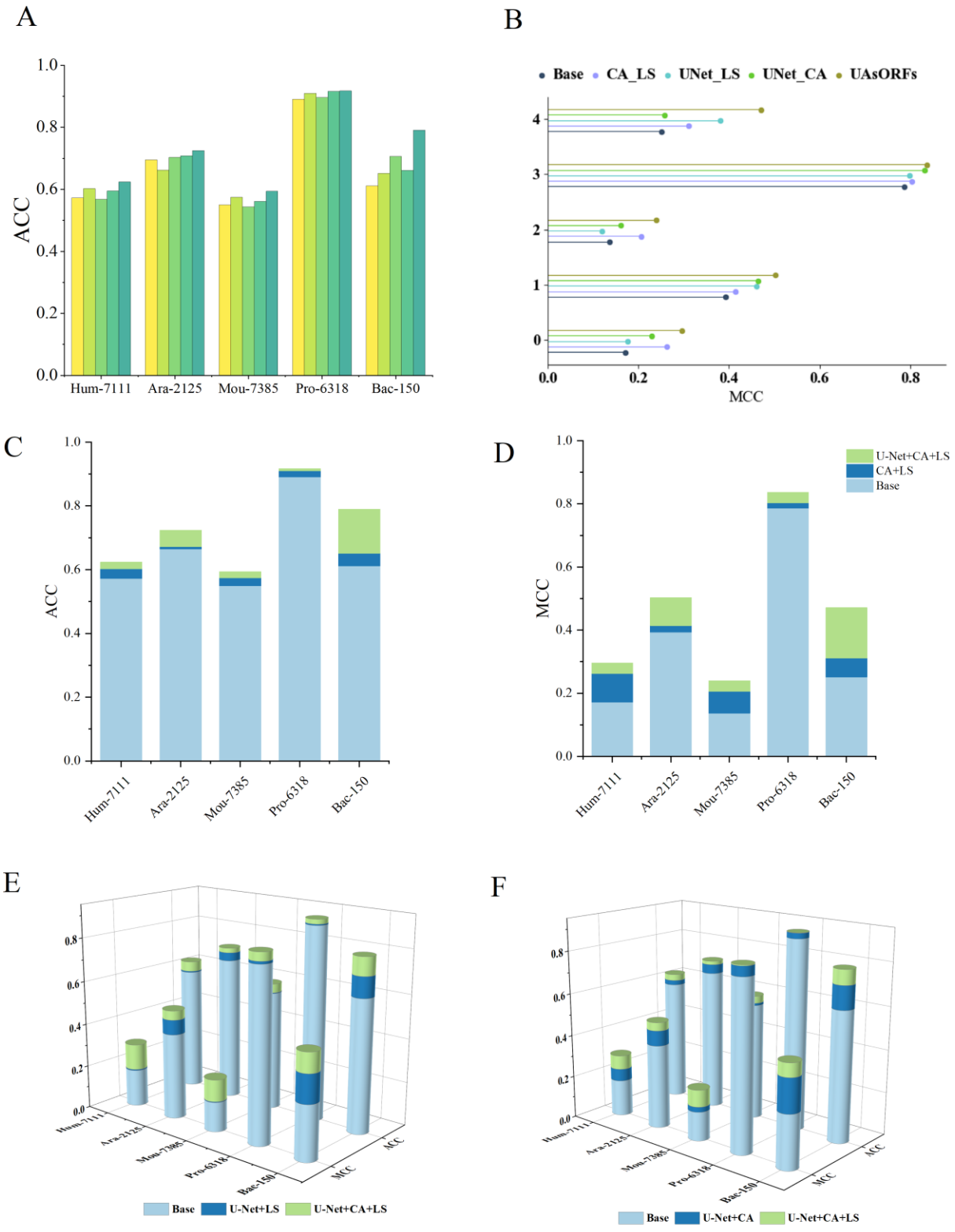

Fig. 9. Comparison of multi-species prediction performance of U-Net, CA and LS blocks of UAsORFs on independent test datasets.

D. Summarize

Through cross-species experiments and ablation studies on datasets from different species, we comprehensively evaluated the performance of the UAsORFs model, and verified the effectiveness of hybrid coding, U-Net, CA, and LS in the UAsORFs, which can effectively distinguish coding sORFs from non-coding ones. There are two main reasons for the outperformance of UAsORFs over the other methods. Firstly, the hybrid coding strategy can access and combine the global and local sequence information of sORFs to enhance the sORF representation. Secondly, deep learning effectively and autonomously learns to extract the features of sORFs, the feature fusion mechanism of up-sampling and intermediate variables in the U-Net module contribute to deep feature mining of sORFs. The CA attention mechanism is able to better capture the complex dependencies between nucleotides, thereby improving the understanding of interactions in sORF sequences. The channel attention mechanism can adaptively learn the importance of each channel, which enhances the model's representation of nucleotide pairing features.

VI. Conclusions

In this work, our study uses a hybrid encoding of one-hot and gkm coding, which retains both the global sequence order information and captures biological features. This approach fully utilizes the advantages of both methods, enhancing the encoding capability of the sequences and greatly avoiding the shortcomings such as insufficient sequence features and human intervention caused by a single encoding method. Additionally, we propose a deep learning architecture called UAsORFs, the deep learning framework distinguish between coding sORFs and non-coding sORFs through autonomous learning. The framework used in this study does not require extensive manual extraction of features, effectively learns essential sORFs features across multiple species and achieve remarkable predictive performance for multi-species sORFs. Additionally, the UAsORFs is a new, novel and efficient method for the prediction of protein-coding sORFs.

The study has seversal limitations. Firstly, the smaller sORFs dataset restricts the ability of the model to learn sORFs features. This is supported by the Multi-species predictions experiment (b), which demonstrates that increasing training samples allows UAsORFs to better capture sORFs sequence features, resulting in better predictive performance. Additionally, models trained on prokaryotic species exhibit suboptimal performance on eukaryotic protein-coding sORFs, suggesting that the current prokaryotic models may not capture certain features present in eukaryotic organisms. Future research should focus on expanding the sORF dataset and constructing a large-scale multi-species dataset to enhance capture features of eukaryotic sORFs and improve the applicability for sORF prediction. Furthermore, exploring species-specific characteristics and integrating other types of biological data (e.g., epigenetic marks, RNA-Seq data) could lead to the development of new biological sequence encoding schemes and further enhance prediction accuracy.

Acknowledgment

All the data mentioned in this article can be available at http://guolab.whu.edu.cn/codingCapacity/download.html.

We would like to thank the research team that provided the dataset.

Funding

This work was supported by the Outstanding Youth Innovation Teams in Higher Education of Shandong Province (2019KJN048).

Author Contributions

Ziling Wang: writing-original draft, methodology, analysis of results, data curation, validation. Wenxi Yang: analysis of the results, supervision, validation. Zhijian Qu: corresponding author, supervision, writing–review & editing, project administration, methodology, funding acquisition.

References

[1] WEI C, ZHANG J, YUAN X. Enhancing the prediction of protein coding regions in biological sequence via a deep learning framework with hybrid encoding. Digital Signal Processing, 2022, 123: 103430.

[2] SIEBER P, PLATZER M, SCHUSTER S. The Definition of Open Reading Frame Revisited. Trends Genet, 2018, 34(3): 167-70.

[3] WRIGHT B W, YI Z, WEISSMAN J S, et al. The dark proteome: translation from noncanonical open reading frames. Trends Cell Biol, 2022, 32(3): 243-58.

[4] FUCHS S, KUCKLICK M, LEHMANN E, et al. Towards the characterization of the hidden world of small proteins in Staphylococcus aureus, a proteogenomics approach. PLOS Genetics, 2021, 17(6): e1009585.

[5] ATAKAN M M, TüRKEL İ, ÖZERKLIĞ B, et al. Small peptides: could they have a big role in metabolism and the response to exercise? J Physiol, 2024, 602(4): 545-68.

[6] SANDMANN C-L, SCHULZ J F, RUIZ-ORERA J, et al. Evolutionary origins and interactomes of human, young microproteins and small peptides translated from short open reading frames. Molecular Cell, 2023, 83(6): 994-1011.e18.

[7] PRENSNER J R, ABELIN J G, KOK L W, et al. What Can Ribo-Seq, Immunopeptidomics, and Proteomics Tell Us About the Noncanonical Proteome? Molecular & Cellular Proteomics, 2023, 22(9): 100631.

[8] VAKIRLIS N, VANCE Z, DUGGAN K M, et al. De novo birth of functional microproteins in the human lineage. Cell Reports, 2022, 41(12): 111808.

[9] PAULI A, NORRIS M L, VALEN E, et al. Toddler: an embryonic signal that promotes cell movement via Apelin receptors. Science, 2014, 343(6172): 1248636.

[10] MATSUMOTO A, PASUT A, MATSUMOTO M, et al. mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. Nature, 2017, 541(7636): 228-32.

[11] NELSON B R, MAKAREWICH C A, ANDERSON D M, et al. A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. Science, 2016, 351(6270): 271-5.

[12] ANDERSON D M, MAKAREWICH C A, ANDERSON K M, et al. Widespread control of calcium signaling by a family of SERCA-inhibiting micropeptides. Sci Signal, 2016, 9(457): ra119.

[13] ZHENG X, WANG M, LIU S, et al. A lncRNA-encoded mitochondrial micropeptide exacerbates microglia-mediated neuroinflammation in retinal ischemia/reperfusion injury. Cell Death Dis, 2023, 14(2): 126.

[14] ZHENG C, WEI Y, ZHANG P, et al. CRISPR/Cas9 screen uncovers functional translation of cryptic lncRNA-encoded open reading frames in human cancer. J Clin Invest, 2023, 133(5).

[15] PRENSNER J R, ENACHE O M, LURIA V, et al. Noncanonical open reading frames encode functional proteins essential for cancer cell survival. Nat Biotechnol, 2021, 39(6): 697-704.

[16] LAURESSERGUES D, COUZIGOU J M, CLEMENTE H S, et al. Primary transcripts of microRNAs encode regulatory peptides. Nature, 2015, 520(7545): 90-3.

[17] LEE C, ZENG J, DREW B G, et al. The mitochondrial-derived peptide MOTS-c promotes metabolic homeostasis and reduces obesity and insulin resistance. Cell Metab, 2015, 21(3): 443-54.

[18] MARTINEZ T F, LYONS-ABBOTT S, BOOKOUT A L, et al. Profiling mouse brown and white adipocytes to identify metabolically relevant small ORFs and functional microproteins. Cell Metabolism, 2023, 35(1): 166-83.e11.

[19] LI J, SMITH L S, ZHU H-J. Data-independent acquisition (DIA): An emerging proteomics technology for analysis of drug-metabolizing enzymes and transporters. Drug Discovery Today: Technologies, 2021, 39: 49-56.

[20] PALAZZO A F, KOONIN E V. Functional Long Non-coding RNAs Evolve from Junk Transcripts. Cell, 2020, 183(5): 1151-61.

[21] VOSS R F. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. Phys Rev Lett, 1992, 68(25): 3805-8.

[22] STADEN R, MCLACHIAN A D. Codon preference and its use in identifying protein coding regions in long DNA sequences. Nucleic Acids Research, 1982, 10(1): 141-56.

[23] SHEPHERD J C. Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. Proc Natl Acad Sci U S A, 1981, 78(3): 1596-600.

[24] CLAVERIE J M, SAUVAGET I, BOUGUELERET L. K-tuple frequency analysis: from intron/exon discrimination to T-cell epitope mapping. Methods Enzymol, 1990, 183: 237-52.

[25] ZHANG C-T, ZHANG R. Analysis of distribution of bases in the coding sequences by a diagrammatic technique. Nucleic acids research, 1991, 19 22: 6313-7.

[26] CHEN W, FENG P M, DENG E Z, et al. iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. Anal Biochem, 2014, 462: 76-83.

[27] RAJAPAKSE J C, LOI SY H. Markov encoding for detecting signals in genomic sequences. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2005, 2(2): 131-42.

[28] CHOONG A C H, LEE N K. Evaluation of convolutionary neural networks modeling of DNA sequences using ordinal versus one-hot encoding method; proceedings of the 2017 International Conference on Computer and Drone Applications (IConDA), F 9-11 Nov. 2017, 2017 .

[29] FU L, PENG Q, CHAI L. Predicting DNA Methylation States with Hybrid Information Based Deep-Learning Model. IEEE/ACM Trans Comput Biol Bioinform, 2020, 17(5): 1721-8.

[30] ZHU M, GRIBSKOV M. MiPepid: MicroPeptide identification tool using machine learning. BMC Bioinformatics, 2019, 20(1): 559.

[31] TONG X, LIU S. CPPred: coding potential prediction based on the global description of RNA sequence. Nucleic Acids Res, 2019, 47(8): e43.

[32] TONG X, HONG X, XIE J, et al. CPPred-sORF: Coding Potential Prediction of sORF based on non-AUG. bioRxiv, 2020.

[33] YU J, GUO L, DOU X, et al. Comprehensive evaluation of protein-coding sORFs prediction based on a random sequence strategy. Front Biosci (Landmark Ed), 2021, 26(8): 272-8.

[34] YU J, JIANG W, S.-B. Z, et al. Prediction of protein-coding small ORFs in multi-species using integrated sequence-derived features and the random forest model. Methods: A Companion to Methods in Enzymology, 2023.

[35] ZHANG Y, JIA C, FULLWOOD M J, et al. DeepCPP: a deep neural network based on nucleotide bias information and minimum distribution similarity feature selection for RNA coding potential prediction. Brief Bioinform, 2021, 22(2): 2073-84.

[36] GHANDI M, LEE D, MOHAMMAD-NOORI M, et al. Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features. PLoS Computational Biology,10,7(2014-7-17), 2014, 10(7): e1003711.

[37] HAFT D H, DICUCCIO M, BADRETDIN A, et al. RefSeq: an update on prokaryotic genome annotation and curation. Nucleic Acids Res, 2018, 46(D1): D851-d60.

[38] OLEXIOUK V, MENSCHAERT G. Using the sORFs.Org Database. Curr Protoc Bioinformatics, 2019, 65(1): e68.

[39] BERARDINI T Z, REISER L, LI D, et al. The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. Genesis, 2015, 53(8): 474-85.

[40] HEMM M R, WEAVER J, STORZ G. Escherichia coli Small Proteome. EcoSal Plus, 2020, 9(1).

[41] ARNIKER S B, KWAN H K, LAW N F, et al. DNA numerical representation and neural network based human promoter prediction system; proceedings of the 2011 Annual IEEE India Conference, F 16-18 Dec. 2011, 2011.

[42] WEI C, YE Z, ZHANG J, et al. CPPVec: an accurate coding potential predictor based on a distributed representation of protein sequence. BMC Genomics, 2023, 24(1): 264.

[43] BERNARD G, GREENFIELD P, RAGAN M A, et al. k-mer Similarity, Networks of Microbial Genomes, and Taxonomic Rank . mSystems, 2018, 3(6).

[44] HATZIGEORGIOU A G. Translation initiation start prediction in human cDNAs with high accuracy. Bioinformatics, 2002, 18(2): 343-50.

[45] WEN J, LIU Y, SHI Y, et al. A classification model for lncRNA and mRNA based on k-mers and a convolutional neural network. BMC Bioinformatics, 2019, 20(1): 469.

[46] LAFFERTY J D, MCCALLUM A, PEREIRA F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data; proceedings of the International Conference on Machine Learning, F, 2001 .

[47] ZHOU H, HU B, YI N, et al. Balancing High-performance and Lightweight: HL-UNet for 3D Cardiac Medical Image Segmentation. Academic Radiology, 2024.

[48] WANG B, QIN J, LV L, et al. DSML-UNet: Depthwise separable convolution network with multiscale large kernel for medical image segmentation. Biomedical Signal Processing and Control, 2024, 97: 106731.

[49] LUO K, TU F, LIANG C, et al. RPA-UNet: A robust approach for arteriovenous fistula ultrasound image segmentation. Biomedical Signal Processing and Control, 2024, 95: 106453.

[50] HOU Q, ZHOU D, FENG J. Coordinate Attention for Efficient Mobile Network Design; proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), F 20-25 June 2021, 2021.

[51] HU J, SHEN L, SUN G. Squeeze-and-Excitation Networks; proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, F 18-23 June 2018, 2018 .