biology

MDPI

Article

D-sORF: Accurate Ab Initio Classification of Experimentally Detected Small Open Reading Frames (sORFs) Associated with Translational Machinery

Nikos Perdikopanis [1,2,3,*], Antonis Giannakakis [4,5], Ioannis Kavakiotis [3] and Artemis G. Hatzigeorgiou [3,6,*]

[1] Department of Electrical and Computer Engineering, University of Thessaly, 38221 Volos, Greece
[2] Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, 15784 Athens, Greece
[3] Department of Computer Science and Biomedical Informatics, University of Thessaly, 38221 Volos, Greece; ikavakiotis@gmail.com
[4] Department of Molecular Biology and Genetics, Democritus University of Thrace, 68100 Alexandroupolis, Greece; antgian@mbg.duth.gr
[5] University Research Institute of Maternal and Child Health and Precision Medicine, National and Kapodistrian University of Athens, 11527 Athens, Greece
[6] Hellenic Pasteur Institute, 11521 Athens, Greece
[*] Correspondence: nikosp@uoa.gr (N.P.); arhatzig@uth.gr (A.G.H.); Tel.: +30-21072-75178 (N.P.); +30-24210-74758 (A.G.H.)

Simple Summary: Small open reading frames (sORFs; fewer than 300 nucleotides or fewer than 100 amino acids) are short DNA sequences that can regulate cellular processes or produce functional peptides. Identifying these sORFs, especially in non-genic regions, remains challenging despite advances in sequencing technology. To address this, we developed D-sORF, a machine-learning framework that predicts coding sORFs using the nucleotide context and motifs around start codons. D-sORF achieves 94.74% precision and 92.37% accuracy, outperforming experimental methods such as ribosome sequencing (Ribo-Seq) in identifying peptide-producing transcripts and filtering out false positives. Unlike traditional conservation-based methods, D-sORF's significant advantage is recognizing sORFs with low sequence similarity. Its robust prediction capabilities make it a valuable tool for researchers. It can enhance our understanding of sORFs' roles, potentially leading to discoveries in terms of gene regulation and new therapeutic targets. By accurately distinguishing small coding sequences from non-coding ones, D-sORF significantly contributes to genomic research and its applications in medicine and biotechnology.

Abstract: Small open reading frames (sORFs; <300 nucleotides or <100 amino acids) are widespread across all genomes, and an increasing variety of them appear to be translating from non-genic regions. Over the past few decades, peptides produced from sORFs have been identified as functional in various organisms, from bacteria to humans. Despite recent advances in next-generation sequencing and proteomics, accurate annotation and classification of sORFs remain a rate-limiting step toward reliable and high-throughput detection of small proteins from non-genic regions. Additionally, the cost of computational methods utilizing machine learning is lower than that of biological experiments, and they can be employed to detect sORFs, laying the groundwork for biological experiments. We present D-sORF, a machine-learning framework that integrates the statistical nucleotide context and motif information around the start codon to predict coding sORFs. D-sORF scores directly for coding identity and requires only the underlying genomic sequence, without incorporating parameters such as the conservation, which, in the case of sORFs, may increase the dispersion of scores within the significantly less conserved non-genic regions. D-sORF achieves 94.74% precision and 92.37% accuracy for small ORFs (using the 99 nt medium length window). When D-sORF is applied to sORFs associated with ribosomes, the identification of transcripts producing peptides (annotated by the Ensembl IDs) is similar to or superior to experimental methodologies based on ribosome-sequencing (Ribo-Seq) profiling. In parallel, the recognition of putative negative data, such as the intron-containing transcripts that associate with ribosomes, remains remarkably low, indicating that

check for updates

Citation: Perdikopanis, N.; Giannakakis, A.; Kavakiotis, I.; Hatzigeorgiou, A.G. D-sORF: Accurate Ab Initio Classification of Experimentally Detected Small Open Reading Frames (sORFs) Associated with Translational Machinery. *Biology* **2024**, *13*, 563. https://doi.org/10.3390/biology13080563

Academic Editor: Youlian Pan

Received: 31 May 2024
Revised: 15 July 2024
Accepted: 24 July 2024
Published: 26 July 2024

D-sORF could be efficiently applied to filter out false-positive sORFs from Ribo-Seq data because of the non-productive ribosomal binding or noise inherent in these protocols.

Keywords: small open reading frames; sORF; machine learning; motif prediction; ribosome sequencing; genomic annotation

1. Introduction

Small open reading frames (sORFs; <100 amino acids) are widespread across all genomes. Their predominantly non-genic location also contributes to the low conservation scores, traditionally used to indicate functional significance in short sequences [1]. Advances in recent decades have revealed various peptides encoded by sORFs in various organisms, from bacteria to humans [2–19]. These studies have illuminated the roles of microproteins encoded by sORFs in regulating essential biological processes, including metabolism [7,13], DNA damage response [20], endocytosis [21], immune surveillance [22,23], development [10,14], and cell death [12]. For instance, specific microproteins have been implicated in fine-tuning metabolic pathways and the orchestration of cellular responses to DNA damage [20], highlighting their potential as therapeutic targets and biomarkers.

The classification of long intergenic, interleaving, and antisense non-coding transcripts (lncRNAs) has traditionally used an arbitrary cut-off ORF length of fewer than 100 amino acids. It also relies on the low scores from alignment-based algorithms for nucleotide and protein conservation. As a result, many lncRNAs may have been erroneously misclassified as non-coding [24]. An emerging understanding is that lncRNAs have risen relatively recently in evolution and may represent a de novo source of novel protein-coding genes [25]. These transcripts share all the features of coding genes (Cap, polyA, $5'$, and $3'$ UTRs). Still, they tend to harbor much smaller ORFs, contain fewer introns, exhibit weak sequence constraints, and show limited phylogenetic conservation [26–28]. To date, protein-coding and functional sORFs have been found in various transcripts. These include untranslated regions of mRNA ($5'$ and $3'$ UTRs), lncRNAs, and microRNA transcripts (pri-miRNAs) [19,29–31].

Advances in ribosome-sequencing profiling (Ribo-Seq) [32–34] and proteomics [35,36] provided the tools necessary to identify protein-coding sORFs. They revealed that the translation of non-coding regions could also be pervasive [37,38]. Various computational methods for ribosome-profiling data have been developed that rely upon specific distribution features of ribosome footprints, like the ORF and FLOSS scores [39–41], and allow the genome-wide identification of many long non-coding RNAs (lncRNAs), $5'$ and $3'$ UTRs [37,38] associated with ribosomes [39,42,43]. Sequencing of ribosome-protected mRNA fragments (RFPs) from samples treated with antibiotics that halt ribosomes at translation initiation sites (TIS) [44] allows the identification of sORFs with experimentally validated TIS positions. However, these datasets have only recently begun to be compiled compared to traditional antibiotics, such as cycloheximide [41]. During the last few years, corresponding public resources have been established, like the GWIPS-viz browser [45], sm-Prot [46], OpenProt [47], and www.sorfs.org (1 March 2023) [39–41] databases, to correctly classify sORFs and discover novel micropeptides.

Many lncRNAs, as well as $5'$ and $3'$ UTRs of mRNAs [37,38], have been reported to show ribosome-profiling patterns similar to canonical non-coding RNAs (e.g., rRNA) that are known not to be translated, implying that many lncRNAs are unlikely to produce functional peptides [48]. However, the ribosome footprints and the translational machinery heavily depend on the transcript abundance and cellular types and states. Genomes undergo widespread transcription during stress to produce a diverse range of non-coding RNAs [49–51]. Therefore, a large proportion of sORFs are expressed in response to stress and exported to the cytoplasm to associate with the translational machinery [49–51]. This fact may indicate that computational annotation pipelines based on experimental data may inherit a significant range of transcriptional and translational "noise", which may act as a rate-limiting step toward the identification of unknown small protein-coding genes or other types of regulatory sORFs that could associate with ribosomes to control the translation of other genes [52]. In addition, mass spectroscopy analyses have found many fewer peptides than predicted sORFs, indicating that accurate annotation of sORFs is critical for both their reliable identification and their unbiased nomenclature classification as protein-coding, functional sORFs or biological/technical "noise" [53].

The problem of identifying small coding ORFs was addressed for the first time nearly a decade ago through an algorithm based only on sequence information [54]. Since then, various sORF-specific ab initio prediction pipelines have been implemented. A group used a combination of cross-species conservation analysis and a phylogenetic assessment of the alignment-based mutations for their synonymous or nonsynonymous nature (Ka/Ks values) [18,54,55]. Others used a hexamer composition bias between coding and non-coding sequences at consecutive windows of 30 bp with a step size of 3 bp [56], coupled with a phylogenetic-based assessment of synonymous versus nonsynonymous substitutions rates, to select more functional sORFs [57]. However, alignment-based methods can be limited depending on the corresponding sORF's length, conservation, and amino acid composition [42], and algorithms comparing coding vs. non-coding sequences have been thus far performed in small windows, not using the sequence features of a translating ORF, like the presence of a translation initiation site (TIS) [44,58].

In this paper, we present D-sORF, a support vector machine-based framework utilizing prediction accuracy supported by a machine-learning model fed with features extracted from the nucleotide composition of the ORF and the sequence information around the start codon, i.e., the translation initiation site (TIS), in a large set of well-established proteins [44]. D-sORF can be applied in a genome-wide manner or just on sORFs associated with ribosomes to classify them based on a large window of sequence identity to the TIS and ORF nucleotide composition without the incorporation of parameters like conservation and phylogenicity. D-sORF calculates and classifies sORFs between 60 and 100 amino acids in length with 97.88% precision and 96.60% accuracy. For sORFs between 33 and 60 amino acids, the precision reaches 94.74%, while the classification accuracy is 92.37%. For sORFs shorter than 33 amino acids, 91.16% precision and 88.15% accuracy are achieved. Remarkably, when D-sORF is applied to a selection of sORFs from the sORFs.org database annotated by both Ribo-Seq data and Ensembl Gene ID (i.e., they represent true microproteins), 935 of the 1075 sORFs are identified as positive, achieving 77.96% accuracy and outperforming all the other computational approaches, including FLOSS.

2. Materials and Methods

In this section, we describe the D-sORF framework as well as the datasets used for the evaluation of our method.

2.1. Positive Datasets

The positive datasets exclusively originated from protein-coding regions (Figure 1A). Initially, we identified and retrieved 43,084 protein sequences corresponding to 19,031 unique UniProt IDs from the Ensembl Repository. It is important to mention that we also retrieved sequences with UniProt/Swiss-Prot IDs. Meanwhile, a small upstream region of 100 nucleotides was also retrieved for each sequence to extract features around the translation initiation site.

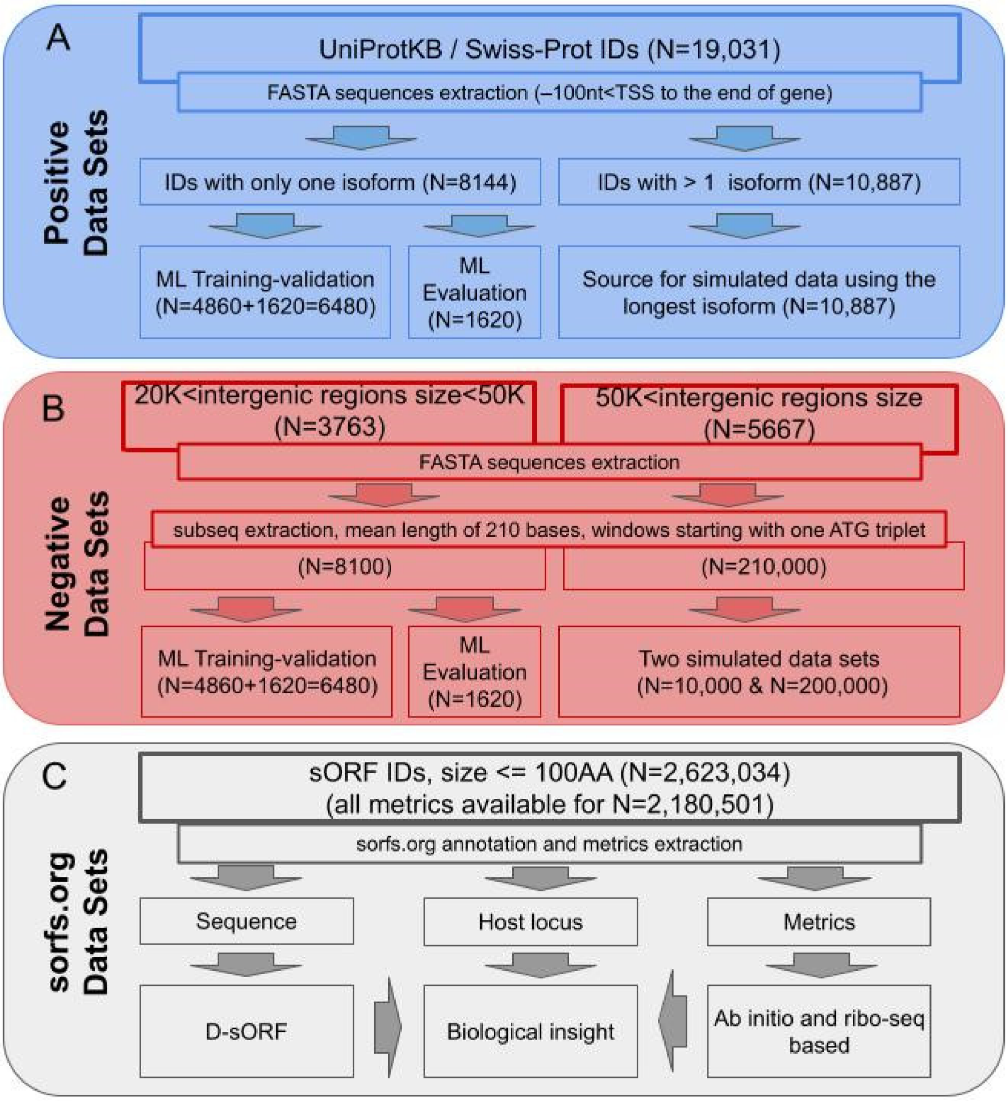

A
Positive Data Sets
UniProtKB / Swiss-Prot IDs (N=19,031)
FASTA sequences extraction (–100nt<TSS to the end of gene)
IDs with only one isoform (N=8144)
IDs with > 1 isoform (N=10,887)
ML Training-validation (N=4860+1620=6480)
ML Evaluation (N=1620)
Source for simulated data using the longest isoform (N=10,887)
B
Negative Data Sets
20K<intergenic regions size<50K (N=3763)
50K<intergenic regions size (N=5667)
FASTA sequences extraction
subseq extraction, mean length of 210 bases, windows starting with one ATG triplet
(N=8100)
(N=210,000)
ML Training-validation (N=4860+1620=6480)
ML Evaluation (N=1620)
Two simulated data sets (N=10,000 & N=200,000)
C
sorfs.org Data Sets
sORF IDs, size <= 100AA (N=2,623,034) (all metrics available for N=2,180,501)
sorfs.org annotation and metrics extraction
Sequence
Host locus
Metrics
D-sORF
Biological insight
Ab initio and ribo-seq based

Figure 1. Dataset creation workflow for the datasets used for training and validation of the D-sORF algorithm. (**A**) All the positive datasets originated from protein-coding sequences with an Ensembl ID. For the curated TIS positions, only Swiss-Prot labelled transcripts were selected. The source dataset was split randomly into training and simulation datasets. The training dataset was subsequently parted in ML model training and evaluation subsets. Only one isoform (the longest) per gene was selected for constructing the simulated positive dataset used for validation. (**B**) For the negative dataset construction, intergenic regions, non-overlapping with coding loci, were selected. Two different negative validation datasets were created. One was balanced (N = 10,000) and one imbalanced (N = 200,000) compared to the positive dataset (N = 10,000). (**C**) The sORFs dataset was created from the orfs.org repository by obtaining the corresponding nucleotide sequence for all the sORFs, together with the host region coordinates, annotations and all the metrics (PhyloP, PhastCon, ORFscore, FLOSS) available in the repository. For each sORF sequence, a 100 nts upstream region was also extracted and combined, then forwarded to the D-sORF algorithm. The resulting scores were compared with the algorithmic metrics available in the sorfs.org repository in order to gain biological insight.

This initial dataset was split into two parts. The first part consisted of 8144 coding sequences with only one isoform in Swiss-Prot to avoid unnecessary complexity deriving from alternative start sites. This first part was retained for the machine-learning component of the algorithm. It was divided into two non-overlapping sets for training and testing, respectively (see Section 2.6). The second part of the initial dataset (i.e., the remaining

34,940 sequences) was used to construct a positive simulated sORF dataset, which corresponded to 10,887 genes with alternative isoforms, from which the longest isoform for each gene was selected as the source sequence (for simplicity).

2.2. Negative Datasets

The negative dataset (Figure 1B) was constructed from the RefSeq database (hg38) [59], where the gene-mediated regions' coordinates and sizes were annotated. To avoid the inclusion of sequences rich in functional information (i.e., promoters), the first 2000 bases upstream of each gene were discarded. Next, the intergenic space was used to extract the negative datasets for training and evaluation. The negative dataset for training was extracted from the 20 K to 50 K size intergenic regions. To ensure that the negative training dataset was balanced with the positive dataset (as described above), 8100 non-annotated/non-functional regions starting with ATG and a mean length of 210 bases were extracted. Regions larger than 50 K in length were utilized (not overlapping with the training datasets) to extract 210,000 sequences split into two datasets of 10,000 and 200,000 (balanced and unbalanced relative to the positive simulated dataset, respectively) to be used as simulated negative datasets. Since the identification of functional sORFs is known to be a heavily imbalanced task, the second negative simulated sORF dataset was constructed to be over 20 times larger compared to the positive simulated sORF dataset (i.e., 10,000 sequences).

2.3. Small ORF Datasets

Data for the Ribo-Seq-derived sORFs (Figure 1C) were obtained from sorf.org, a public repository for sORFs identified by ribosome footprint sequencing. The analysis encompassed human sORFs with lengths of fewer than 100 AAs. Among the various information provided by the repository, the corresponding annotation regarding the sORF's host genomic region was used to group the sORFs as exonic (located in the coding sequence of a gene), intronic (located in the intronic part of a gene), intergenic (sORFs located between genes), ncRNA (sORFs located on non-coding RNA), 3′-UTR (sORFs located in the 3′-UTR), 5′-UTR (sORFs located in the 5′-UTR), NMD (nonsense-mediated decay; if the coding sequence, following the appropriate reference of a transcript, finishes >50 bp from a downstream splice site, then it is tagged as NMD), TEC (to be experimentally confirmed; used for non-spliced EST clusters that have polyA features), NSD (non-stop decay; transcripts that have polyA features, including a signal, without a prior stop codon in the CDS), and sORFs (sORFs corresponding to Ensembl-annotated protein-coding ORFs of fewer than or equal to 100 AAs). The latter subgroup was a biological positive dataset representing true coding sORFs. From the 2,623,034 downloaded records, sorfs.org provides (FLOSS score [37], ORFscore [39], PhyloP [55], and PhastCon [60]) metrics for 2,180,501 sORFs. We selected the last mentioned for a detailed investigation.

2.4. D-sORF Machine-Learning Framework

D-sORF is a modular machine-learning framework consisting of three sequential components, namely data preprocessing and transformation, machine-learning algorithm selection and evaluation, and sORF predictor (Figure 2).

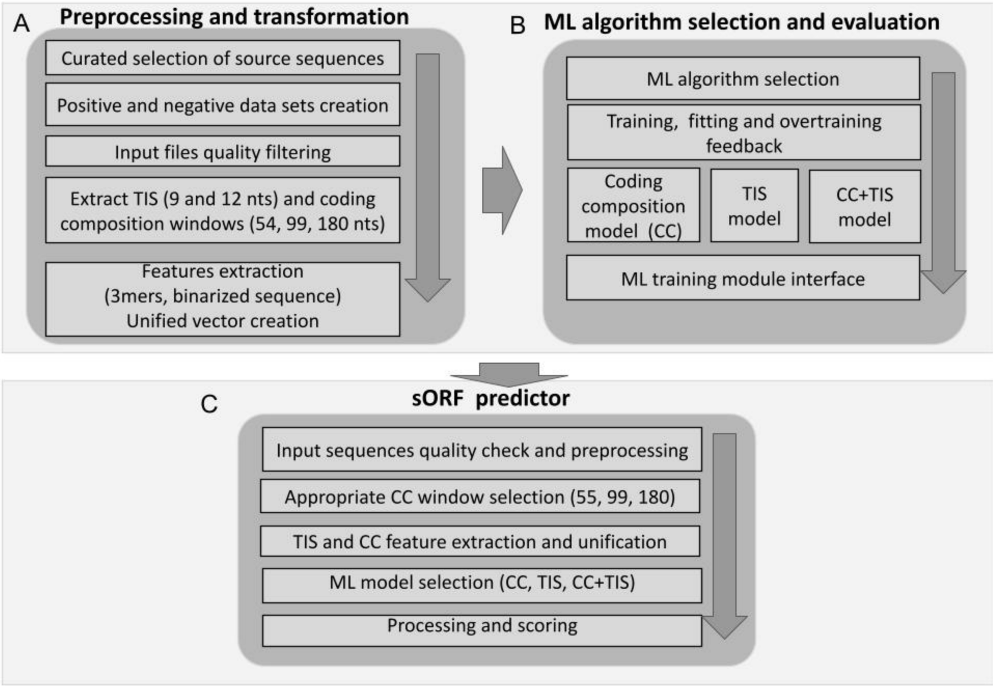

A
Preprocessing and transformation
Curated selection of source sequences
Positive and negative data sets creation
Input files quality filtering
Extract TIS (9 and 12 nts) and coding composition windows (54, 99, 180 nts)
Features extraction (3mers, binarized sequence) Unified vector creation
B
ML algorithm selection and evaluation
ML algorithm selection
Training, fitting and overtraining feedback
Coding composition model (CC)
TIS model
CC+TIS model
ML training module interface
C
sORF predictor
Input sequences quality check and preprocessing
Appropriate CC window selection (55, 99, 180)
TIS and CC feature extraction and unification
ML model selection (CC, TIS, CC+TIS)
Processing and scoring

Figure 2. Workflow of the D-sORF Framework. (**A**) In the "preprocessing and transformation" module, the input sequence datasets were cleared from ambiguous and low-quality sequences and sequence windows for the TIS and CC (coding composition) were separated for feature extraction. A vector was created for all the CC and TIS features individually as well as from the unification of the corresponding TIS and CC features. These vectors were the drivers of the training process. (**B**) In the "ML algorithm selection and evaluation" module, comparison and selection between well-known ML algorithms (e.g., SVM, RF) took place. For each feature vector (CC, TIS and combined CC with TIS), a separate output model was trained to be used from the D-sORF main module (predictor). (**C**) The D-sORF predictor was fed with sORFs for processing. The best fitting window among 55, 99, and 180 was selected based on the sORF length. The user could select which model to apply (CC, TIS or combinatory CC+TIS mode) The Diana-sORF returns a score in the range [0 to 1]. Higher scoring values provide the sORF with a greater chance of being coding.

2.5. Data Preprocessing and Transformation

The data preprocessing and transformation module takes as input the positive and negative datasets in standard FASTA format (Figure 2A). This module analyzes the raw FASTA data, transforming them into the appropriate internal structures while performing the essential quality control to exclude unqualified sequences, e.g., sequences with ambiguities (only A, C, T, and G are allowed). Following this essential initial step, the module continues the feature extraction process, extracting two separate sequence windows for the TIS and coding composition (CC) sampling regions to score two separate features for the coding potential (Supplementary Figure S1). The TIS window includes nucleotides at positions −7 to +5 centered at the given start codon of the sORF sequence. It represents the translation initiation site (TIS) pattern, excluding the start codon and thus avoiding the potential bias of start codon usage in sORFs. Subsequently, each nucleotide window is transformed into a new binarized representation based on the code A = 0001, C = 0010, T = 0100, and G = 1000 (binary sequences with the greatest Hamming distance), resulting in a 36-element vector.

To capture the coding composition (CC) of a given sORF, three different sequence length CC windows of 54 nts, 99 nts, and 180 nts were extracted downstream of the first +90 nts until the corresponding stop codon. The first 90 nts were omitted from training and testing to avoid plausible signal peptides known to precede the protein-specific nucleotide context [61]. Signal peptides have a very specific amino acid composition [2] and can produce a negative bias during the training procedure. All the windows were extracted on a 3-step basis to maintain the in-frame information. Subsequently, a 64-element vector was created containing the frequency of each of the 64 different 3-mers. Finally, the TIS and CC-derived vectors were unified.

2.6. ML Algorithm Selection and Evaluation

The initial dataset, consisting of 16,200 examples (8100 positive and 8100 negative), was randomly split to create the training (9720—60%), validation (3240—20%), and test (3240—20%) subsets. The positive–negative ratio was maintained in the datasets produced above. Three well-known machine-learning algorithms were investigated, namely Support Vector Machine (SVM) [62], k-Nearest Neighbor (k-NN) [63], and Random Forest (RF) [64].

To ensure unbiased model selection based on performance, the hyper-parameter space of the algorithms was tested for the best combination toward the best cross-validation score. The hyper-parameters optimized for SVM were (a) the kernel type used in the algorithm, (b) the kernel coefficient gamma, and (c) the penalty parameter C. The hyper-parameters optimized for k-NN were (a) the number of neighbors, (b) the weight function used in the prediction, and (c) the algorithm used to compute the nearest neighbor. Finally, the hyper-parameters tuned for RF were (a) the number of trees in the forest, (b) the function to measure the quality of a split, and (c) the number of features to consider when looking for the best split.

The search in the hyper-parameter space was performed exhaustively, considering all the parameter combinations. The tuning of the hyper-parameters was performed on the training data using 10-fold cross-validation. After obtaining the best hyper-parameter combination for each algorithm, the three algorithms (SVM, k-NN, RF) were re-trained on the training set using the selected hyper-parameter combination. The final accuracy of each model was measured using the initially reserved test set. Trained models compose the D-sORF predictor core. SVM (RBF kernel) achieved the best performance (ACC = 98.1%), followed by RF (ACC = 97.5%), and finally, k-NN (ACC = 94%). Hence, the analysis proceeded using the SVM classifier. To obtain the probabilities for each respective label, the parametric approach based on Platt's sigmoid model [3] was used, based on fitting a logistic regression model (sigmoid) to a classifier's scores. The validation set was used for the probability calibration, which differs from the model-fitting dataset.

2.7. sORF Predictor

The final standalone algorithm of the D-sORF framework is the sORF predictor (Figure 2C), which is based on the trained ML model and can be applied for the classification of a given sORF. Similar to the training and evaluation module, two types of windows were selected for the TIS and CC feature extraction (Figure 3). The CC window can be 54, 99, or 180 nts according to the sORF length. In the default operation mode, the algorithm selects the longest windows for feature extraction but allows this option to be parameterized by the user. After that, the TIS and CC features are concatenated inline, resulting in a 100-dimension feature vector, which is classified by the appropriate ML model of 54 nts, 99 nts, or 180 nts, respectively, returning a score ranging from 0 to 1. The user can configure the D-sORF predictor with any trained models, depending on the sORF length range under study.

The downstream analysis proceeds by incorporating big data originating from sorfs.org. It produces a series of reports that categorize the outputs according to the annotation, host region, D-sORF classification, and all the metrics for encoding the ability or classification parameters provided by the repository.

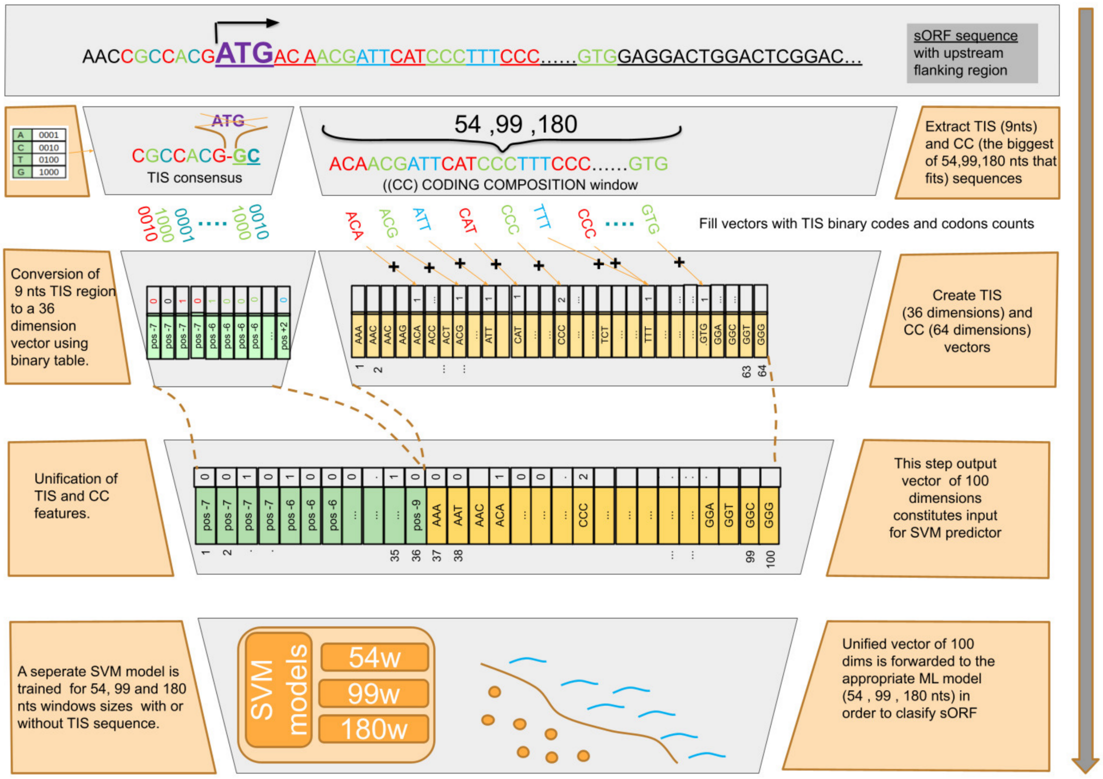

AACCGCCACGATGACAACGATTCATCCCTTTCCC......GTGGAGGACTGGACTCGGAC...
sORF sequence with upstream flanking region
ATG
CGCCACG-GC
TIS consensus
54 ,99 ,180
ACAACGATTCATCCCTTTCCC......GTG
((CC) CODING COMPOSITION window
Extract TIS (9nts) and CC (the biggest of 54,99,180 nts that fits) sequences
Fill vectors with TIS binary codes and codons counts
Conversion of 9 nts TIS region to a 36 dimension vector using binary table.
Create TIS (36 dimensions) and CC (64 dimensions) vectors
Unification of TIS and CC features.
This step output vector of 100 dimensions constitutes input for SVM predictor
A seperate SVM model is trained for 54, 99 and 180 nts windows sizes with or without TIS sequence.
SVM models
54w
99w
180w
Unified vector of 100 dims is forwarded to the appropriate ML model (54 , 99 , 180 nts) in order to clasify sORF

Figure 3. Workflow of the D-sORF predictor. The signal peptide on the sORF sequence is not excluded for the CC sequence (user can enable this feature). The start codon is removed in order to extract the TIS pattern sequence. Both the TIS and CC features are used for the classification and automatic selection of the best fitting window sequence (54, 99, 180) and ML model, according to the sORF length. Processing starts using as an input the sORF sequence combined with the upstream flanking region. After that, the TIS (9 nts) and CC (the biggest among 54, 99, 180 nts) sequences are extracted. The TIS sequences are then converted to a 36-dimension vector using the same binary table as in the training module. This constitutes the first part of the final 100-dimension vector, while the coding composition 64-dimension vector constitutes the second part. The derived vector is then forwarded to the appropriate ML model (54, 99, and 180 nts) for the sORF to be classified.

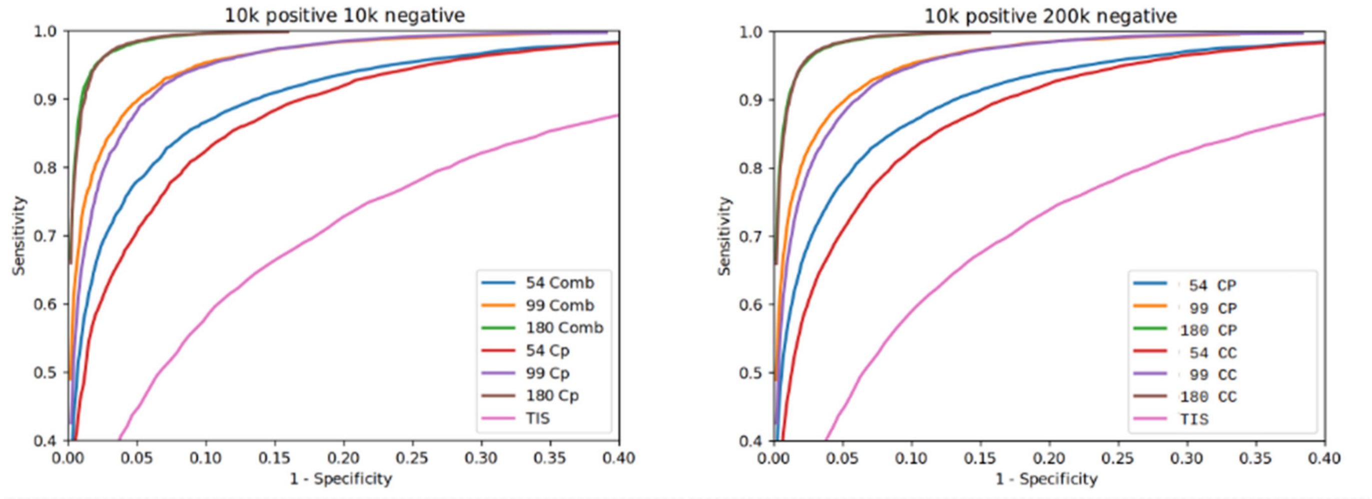
3. Results

Two different approaches were employed for the evaluation of D-sORF. Specifically, different types of simulated datasets were used for the initial algorithmic evaluation of D-sORF. Furthermore, the biological significance of D-sORF was evaluated by large-scale analysis of experimentally validated sORFs and by comparing D-sORF's performance to the available metrics for coding potential (FLOSS score, ORFscore, PhyloP, PhastCon) and to the different sORF annotation classes. The Ensembl ID-labeled class of sORFs annotated as small proteins was used as the most representative biological positive dataset, while the sORFs annotated as introns were selected as a biological negative dataset.

3.1. Algorithm Evaluation

To evaluate the predictive strength of the classification models created from the features, the TIS and CC alone or in combination, the algorithm was tested using the following seven sets of features: TIS on its own, CC on its own with three different length windows (54, 99, and 180 nts), and combinations of TIS with each of the different length CC windows. When applying the problem of recognizing coding sORFs on real data (genome, transcripts, and ribosome-associated transcripts), the ratio of positive to negative data is heavily unbalanced, with negative data usually being over-represented. To partly model this imbalance in our evaluation procedure, two different pairs of validation datasets were created: the first pair represented a balanced combination of 10,000 positive with 10,000 negative samples, whereas the second contained an over-represented negative dataset (200,000), i.e., twenty times larger than the positive set.

We performed two sets of evaluation procedures, one for each balanced and one for each unbalanced group of datasets. In each experiment, we compared the seven previously mentioned models (TIS, CC54, CC99, CC180, CC54+TIS, CC99+TIS, CC180+TIS). In all cases, the classifiers with the combined features performed better compared to the CC and TIS alone. For the balanced dataset, the classification of an sORF using only the smallest size, CC54, reached 0.87 precision, 0.86 sensitivity, 0.87 accuracy, and 0.87 specificity (Supplementary Table S1A). When the algorithm encompassed the TIS features (i.e., CC54+TIS), the precision, sensitivity, accuracy, and specificity of D-sORF changed to 0.91, 0.85, 0.88, and 0.92, respectively (Supplementary Table S1B). The other two models, CC99 and CC180, presented similar improvements (Supplementary Table S1C–F). The combined model was termed CP to represent the unified coding potential feature. Since it showed the best performance, we may conclude that both features (TIS and CC) have added value when combined, thus highlighting the potential role of the TIS in sORFs' encoding ability (Supplementary Figure S2). The trade-off between sensitivity and specificity and the differences between CC and CP were demonstrated using ROC curve diagrams (Figure 4).



(**A**) (**B**)

Figure 4. D-sORF ROC curves. Sensitivity vs. specificity diagrams following integration of the TIS and CC features into a unified coding potential (CP) feature. The diagrams depict the framework performance after the integration of the TIS in 54, 99, and 180 nts CC window mode (**A**) for balance data, and (**B**) for imbalanced data.

Similarly, in the unbalanced dataset, the classifiers exhibited the same behavior as in the balanced set. The unified model CP, combining the two different features (i.e., TIS and CC180), performed better and exhibited the best performance compared to the TIS and CC models alone. CC180+TIS achieves 0.69, 0.95, 0.97, and 0.97 (0.63, 0.96, 0.97, and 0.97 for CC180), while CC99+TIS shows 0.47, 0.90, 0.95, and 0.95 (0.40, 0.91, 0.93, and 0.93 for CC99) for precision, sensitivity, accuracy, and specificity (Supplementary Table S1I,J,K,L). The CC54+TIS model shows 0.34, 0.84, 0.91, and 0.92 for the same metrics (0.26, 0.85, 0.88, and 0.88 for CC54) (Supplementary Table S1G,H). Based on these results, we can conclude that although the larger the CC window, the better the performance, D-sORF retains good scores for sensitivity and specificity for all the window sizes (Supplementary Figure S2).

It is worth noting that D-sORF is a threshold-enabled algorithm, and users can adjust the threshold for fine-tuning. All the above metrics were obtained using a threshold value of 0.5 (scores above 0.5 count as coding). The sensitivity and specificity are balanced for a threshold of 0.9. Using this value significantly increases the algorithm's precision. For instance, the precision in the unbalanced dataset reaches 0.68 from 0.34 for CC54+TIS, 0.77 from 0.47 for CC99+TIS, and 0.87 from 0.69 for CC180+TIS (Supplementary Table S1H,L).

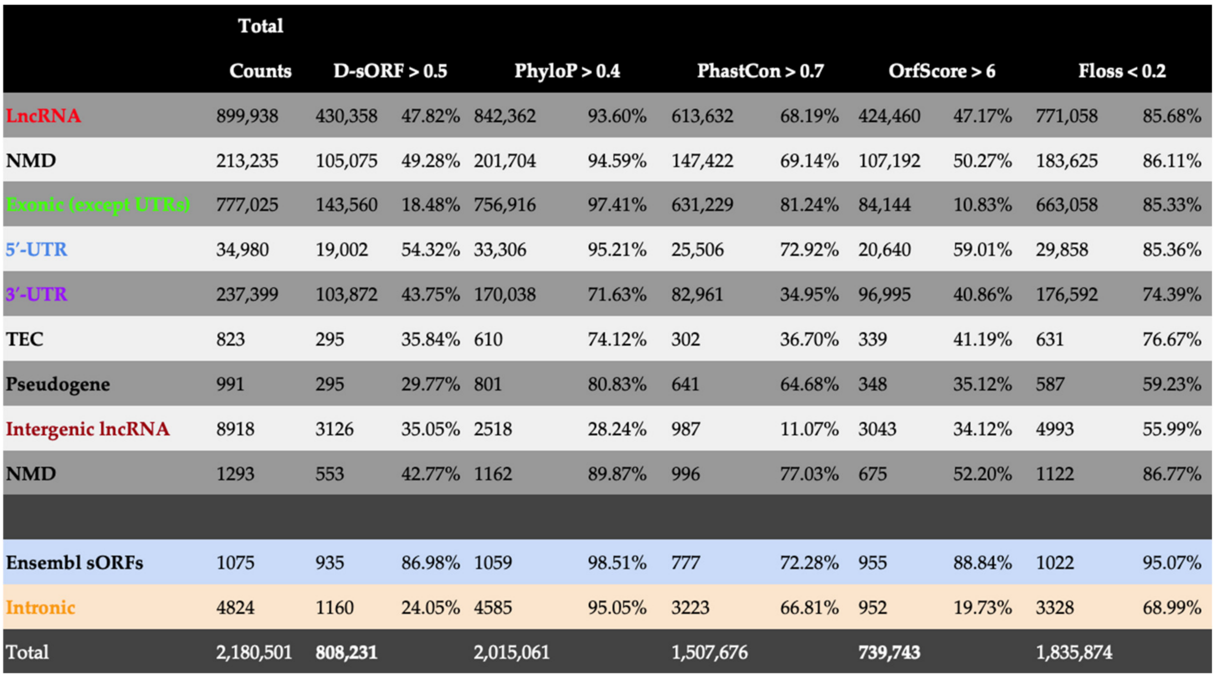
3.2. Application of D-sORF in Annotating Experimentally Detected sORFs in Genomic Regions Enriched with Coding and Non-Coding Transcripts

D-sORF's performance on real-life sORFs was tested using the sorfs.org repository of experimentally validated sORFs regarding their start codon and association with the ribosome footprinting coding patterns. In addition, the sORFs used were already annotated by sorfs.org regarding their location around or within the main mRNA coding frame, as annotated by Ensembl. Exonic sORFs are located in the exons of mRNAs but out of frame, while intronic sORFs are in the introns of mRNA genes. Moreover, 5′ and 3′ UTRs represent sORFs located in mRNAs' corresponding exonic but non-genic regions. LncRNAs, TEC, pseudogene, and NMD groups are sORFs located in transcripts annotated by Ensembl as Long Non-Coding RNAs, as To-be-Experimentally Confirmed protein-coding, as Pseudogenes, and as transcripts with premature stop codons likely to be degraded through the Non-sense Mediated Decay pathway, correspondingly. Finally, sORFs detected in lncRNAs but located at distal genomic sites (>5000 bp ± of any mRNA) were studied separately from the rest of the lncRNAs, as Intergenic (Supplementary Figure S3). The above biotype groups contain transcripts harboring a spectrum of expected or known coding/non-coding capacity. Based on current knowledge, the intronic regions show the most biological relevance as a negative control (highly non-coding group).

At the same time, the 5′ UTR sORFs represent an established genomic location of a known functional subclass of sORFs, named upstream ORFs (uORFs), with an established role in mostly repressing but also, for a subset of them, regulating the translation of the mRNA ORFs [65]. However, we considered 5′ UTR sORFs as a test group, not as a biological positive reference group of genomic regions with mainly and highly coding sORFs. Instead, we selected a subset of transcripts from the sorfs.org database corresponding to the Ensembl IDs of validated microproteins (i.e., <100 amino acids).

The resulting D-sORF scores across all the above annotation groups were compared to all the available classification metrics used in the sorfs.org database to classify the detected sORFs' coding potential, such as the FLOSS score, ORFscore, PhyloP, and PhastCon (Figure 5). The PhyloP and PhastCon scores are based on phylogenetic information, while the ORFscore and FLOSS scores algorithmically analyze different distribution parameters from ribosome-footprinting experiments. For each type of score, the cut-off values suggested in the corresponding reference papers on sorfs.org were used [40,41]. The PhyloP and PhastCon algorithms showed a relative constitutive detection rate across all the annotation groups, including the positive (Ensembl sORFs) and negative (Intronic) groups. The PhyloP system scored for the coding sORFs with an increased frequency, while PhastCon had a decreased frequency rate in both of the above groups compared to the corresponding detection rates of the experimentally-based algorithms of ORFscore and FLOSS.

In contrast, D-sORF detected coding sORFs across all the groups (including positive and negative groups) with a frequency rate almost similar to the experimentally based algorithm ORFscore. In addition, D-sORF's classification of coding sORFs showed a comparable detection rate to the experimentally based FLOSS score with some important discrepancies, which could be attributed to the increased sensitivity of the FLOSS score, as indicated by its increased detection rate (68.99%) within the intronic negative group. Regarding the positive group of annotated micropeptides (Ensembl sORFs) contained in the sorfs.org database, D-sORF scored as positive 935 out of 1075 (86.98%), showing similar discriminatory ability to ORFscore (955 out of 1075, 88.84%) and proving its biological significance in identifying true-positive peptide-encoding sORFs. A representative subset is presented in Appendix A.



	Total Counts	D-sORF > 0.5		PhyloP > 0.4		PhastCon > 0.7		OrfScore > 6		Floss < 0.2	
LncRNA	899,938	430,358	47.82%	842,362	93.60%	613,632	68.19%	424,460	47.17%	771,058	85.68%
NMD	213,235	105,075	49.28%	201,704	94.59%	147,422	69.14%	107,192	50.27%	183,625	86.11%
Exonic (except UTRs)	777,025	143,560	18.48%	756,916	97.41%	631,229	81.24%	84,144	10.83%	663,058	85.33%
5′-UTR	34,980	19,002	54.32%	33,306	95.21%	25,506	72.92%	20,640	59.01%	29,858	85.36%
3′-UTR	237,399	103,872	43.75%	170,038	71.63%	82,961	34.95%	96,995	40.86%	176,592	74.39%
TEC	823	295	35.84%	610	74.12%	302	36.70%	339	41.19%	631	76.67%
Pseudogene	991	295	29.77%	801	80.83%	641	64.68%	348	35.12%	587	59.23%
Intergenic lncRNA	8918	3126	35.05%	2518	28.24%	987	11.07%	3043	34.12%	4993	55.99%
NMD	1293	553	42.77%	1162	89.87%	996	77.03%	675	52.20%	1122	86.77%
Ensembl sORFs	1075	935	86.98%	1059	98.51%	777	72.28%	955	88.84%	1022	95.07%
Intronic	4824	1160	24.05%	4585	95.05%	3223	66.81%	952	19.73%	3328	68.99%
Total	2,180,501	808,231		2,015,061		1,507,676		739,743		1,835,874	

Figure 5. Genome-wide application of D-sORF to classify experimentally validated sORFs as coding. The table contains all the sORFs from the sorfs.org database and shows the absolute count and percentage of sORFs classified as positive by D-sORF, PhyloP, PhastCon, ORFscore, and FLOSS at their corresponding cut-off values. The sORFs have been annotated based on sorfs.org classification parameters as exonic (sORFs located in the exonic part of a gene), intronic (sORFs located partially or completely overlapping introns), intergenic (sORFs in lncRNAs at distal sites > 5000 bp), lncRNA (sORFs overlapping long non-coding RNAs), 3′ UTR (sORFs in 3′-UTR regions of genes), 5′ UTR (sORFs in 5′-UTR regions of genes), Ensembl sORF (sORFs corresponding to Ensembl ID protein-coding ORFs of fewer or equal to 100 AA), NMD (classified as nonsense-mediated decay if sORF overlaps the coding sequence of a transcript that finishes >50 bp from a downstream splice site) and TEC (to be experimentally confirmed as protein-coding, overlapping EST clusters that have polyA features). The colors associated with each sORF class indicate the genomic region from which they originate, as illustrated in Supplementary Figure S3.

Reassuringly, D-sORF, unlike PhyloP, PhastCon, and FLOSS, scored low (~24%) in the mainly non-coding class of intronic sORFs, again similarly to the experimentally based ORFscore (~20%), establishing its biological relevance in understanding coding potentiality in rare or newly evolved sORFs. It is important to note that the 5′ UTR sORFs showed the second to highest positivity detection after the Ensembl sORFs by both D-sORF (~54%) and ORFscore (~59%) (Figure 5). In summary, D-sORF performed significantly better (as ORFscore) than the FLOSS score. For example, in the lncRNA class of sORFs, which is expected to be enriched for peptide-encoding sORFs, D-sORF showed a comparable prediction score (47.82%) to ORFscore (47.17%). In comparison, FLOSS and PhyloP produced significantly greater positive counts (85.68% and 93.60%) for lncRNAs.

To gain more insight, 1075 Ensembl sORFs and 4848 intronic sORFs with prediction scores by all the metrics were plotted in Venn diagrams. In the Ensembl sORF positive group, significant overlapping of common positive sORFs was observed (573/1075 ~54%) for all the metrics and (804 ~75%) for D-sORF, ORFscore, and FLOSS combined (Figure 6A). On the contrary, in the intronic category, slight overlapping of common positive sORFs (137/4824 ~3%) was observed (Figure 6B). It remains to be investigated whether the above small percentage of intronic sORFs that were detected as positive by all the metrics and by D-sORF, ORFscore, and FLOSS score combined could represent a subset of circular RNA transcripts that translate microproteins with significant biological roles in human diseases [66].

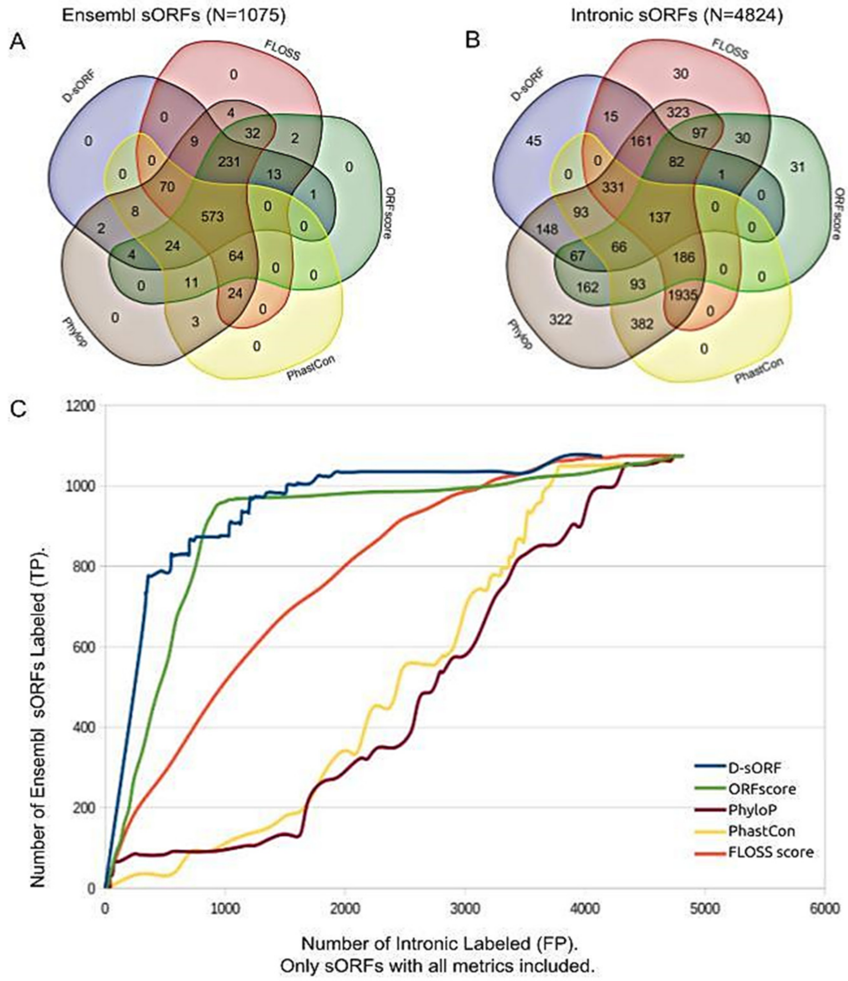


Figure 6. Evaluation of D-sORF compared to all the metrics available in the sorfs.org repository. Venn diagram of (**A**) Ensembl-validated sORFs and (**B**) intronic sORFs. Presenting the results of all the different algorithms and metrics. Counts and percentages of sORFs that were classified as positive and are in common between D-sORF, PhyloP, ORFscore, and FLOSS are shown. Only sORFs with values from all the scores were used. (**C**) The relation of the prediction from the Ensembl-annotated sORFs and intronic sORFs at different thresholds for D-sORF, PhyloP, ORFscore, and FLOSS.

Importantly, using the Ensembl-labeled sORFs as a positive set and intronic sORFs as a negative one, D-sORF's biological specificity surpassed PhyloP, PhastCon, and FLOSS score both in accuracy (77.96% vs. 22.00%, 40.31%, and 42.69%) and precision (44.63% vs. 18.76%, 19.43%, and 23.49%). The metrics are slightly lower than ORFscore. Statistical measures for all the metrics are presented in Supplementary Table S2.

Finally, all the metric scores for the Ensembl and intronic sORFs were plotted against each other. Especially for PhyloP and ORFscore, we normalized and shifted to the 0–1 interval to have comparable values. Figure 6C shows that D-sORF performed significantly better than FLOSS score (the main computational classification tool used in processing ribosome profiling information) and surpassed all the other algorithms/metrics. The above results indicate that even though the D-sORF algorithm is purely trained on in silico datasets with strictly ab initio computational processes, it can accurately predict coding sORFs better than experimentally based Ribo-Seq algorithms and alignment-based frameworks. In essence, D-sORF can be applied as an unbiased sORF classifier that can be used alone or with other metrics to functionally classify transcripts and describe various biological aspects of sORF classes.

4. Discussion

Our knowledge of the translational "vocabulary" of the human genome has expanded thanks to ribosome profiling, or "Ribo-seq", which has identified hundreds of open reading frames (ORFs) within long non-coding RNAs (lncRNAs) and putative untranslated regions (UTRs) of protein-coding genes. Reference gene annotation programs, however, have been cautious in incorporating these ORFs due to unknowns regarding their physiological roles and experimental repeatability. Nonetheless, it is evident that many "Ribo-seq ORFs" have physiological significance, whereas others control gene regulation or produce stable proteins. In the end, the lack of a standardized ORF annotation pipeline has produced a vicious cycle: studies looking into the roles of Ribo-Seq ORFs will be impeded by their lack of recognition as long as reference annotation databases continue to ignore them [67].

Here, we present D-sORF, which could be used alone or in combination with experimental-based metrics to filter down candidate encoding sORFs from different annotation groups and, therefore, increase the rate of their experimental validation. In addition, D-sORF can be used to discriminate ORFs that associate with ribosome footprint patterns but show true-negative coding potential and thus may have a non-coding regulatory function instead. D-sORF's significant advantage is that it can predict positive and negative sORFs equally well. Thus it can be applied to functionally classify lncRNAs and other non-coding transcripts that either encode sORFs and thus represent a recently evolved class of protein-coding genes or regulate the translational machinery in a non-coding manner.

D-sORF is an ab initio scoring system that accurately distinguishes coding from non-coding sORFs based on the nucleotide context of a 180 nts coding ORF composition window and a 9 nts window around the TIS of experimentally validated protein-coding ORFs, compared to windows of similar length in carefully curated intergenic regions used as negative datasets. D-sORF represents the first-of-its-kind classification algorithm powered by trained machine-learning models to predict putative translating sORFs in the human genome accurately. Compared to phylogenetic and sequence conservation approaches, like PhyloP and PhastCon, D-sORF exhibited superior performance in accurately depicting, through prediction, the sORF encoding resemblance to protein ORFs. This indicates that, in general, ab initio nucleotide similarity approaches are more effective than cross-species comparison pipelines in utilizing the sequence information of recently evolved small ORFs with little similarity to known proteins and low conservation in the amino acid sequences.

Specifically, PhastCon and PhyloP showed a general trend to predict low or high scores for all the sORFs regardless of their annotation class, including the negative group. For instance, although PhyloP scores high for Ensembl sORFs, it also scores high for intronic sORFs (~94%), showing ~5% specificity and 22% accuracy (see Supplementary Table S2). However, D-sORF, FLOSS, and ORFscore performed with the highest accuracy rate and represented the top three algorithms with the highest prediction accuracy, which were able to discriminate true positives from the Ensembl sORF class. Significantly, these three algorithms also scored the lowest for the best-known negative set of the intronic sORF class.

Furthermore, although D-sORF is based only on sequence training, it showed better discriminatory ability than the Fragment Length Organization Similarity Score (FLOSS), representing the most recent Ribo-Seq algorithmic scoring described by [37]. This algorithm identifies true ribosome footprints by measuring the magnitude of disagreement between the RPF-length distribution of Ensembl-annotated protein-coding sequences and the RPF-length distribution of an sORF. D-sORF's performance is similar to ORFscore, another Ribo-Seq sORF annotation algorithm that, in contrast to the FLOSS system, is known to have stringent parameters for detecting true ribosome footprints outside the annotated CDS of protein-coding genes (i.e., in lncRNAs, $5'$ and $3'$ UTR regions) [39]. This indicates that Ribo-Seq-based algorithms might suffer from a translational bias, depending on the pattern of translational efficiency they score.

Our results show that D-sORF showed remarkable prediction sensitivity and was comparable only to Ribo-Seq-based algorithms. Nevertheless, D-sORF's ab initio approach and carefully curated positive and negative training datasets, coupled with deep-learning models, may represent what is needed to facilitate an alignment-free and data-independent functional classification of experimentally validated sORFs.

D-sORF has significant potential in real-world biological research, particularly in studying low-conservation and/or low-similarity sORFs. These small but crucial genomic elements can encode functional microproteins in various regulatory processes. By accurately predicting coding sORFs, D-sORF can help toward achieving robust annotation, identification, and experimental validation of novel microproteins. This capability, which does not rely on experimental data, is especially relevant in shedding light on previously unexplored aspects of gene regulation and cellular function and understanding diseases where microproteins might play a role in the metabolism, DNA repair, and immune response pathways.

5. Conclusions

Our study introduces D-sORF, a machine-learning framework designed to predict coding small open reading frames (sORFs) using the nucleotide context and motif information around the start codons. D-sORF demonstrates high precision (94.74%) and accuracy (92.37%) in identifying sORFs, surpassing traditional and experimental methods such as ribosome sequencing (Ribo-Seq).

D-sORF enhances the prediction accuracy by outperforming conservation-based approaches, making it particularly effective in recognizing recently evolved sORFs with low sequence similarity. Additionally, the tool effectively filters out false positives, such as intronic sORFs, often mistakenly identified by other methods. Its broad applicability allows D-sORF to be applied to various genomic datasets. It provides a robust method for distinguishing coding sORFs from non-coding sequences without relying on parameters like the sequence conservation. Furthermore, D-sORF can functionally characterize non-coding transcripts, potentially identifying novel protein-coding genes or regulatory elements.

Overall, D-sORF represents a significant advancement in genomic research, offering a reliable method for sORF detection that can facilitate discoveries in gene regulation and protein function. This tool holds promise for enhancing our understanding of the genomic landscape and its implications for biology and medicine.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/biology13080563/s1, Figure S1: Sequence feature extraction windows; Figure S2: Balanced/unbalanced dataset performance; Figure S3: sORF annotation; Table S1: Model performance; Table S2: Algorithm comparison metrics.

Author Contributions: Conceptualization, N.P., A.G.H. and A.G.; methodology, N.P., A.G.H. and A.G.; software, N.P.; validation, N.P., A.G. and I.K.; formal analysis, N.P., A.G., I.K. and A.G.H.; investigation, N.P. and A.G.; resources, N.P. and A.G.H.; data curation, N.P. and A.G.; writing—original draft preparation, N.P., A.G. and I.K.; writing—review and editing, N.P., A.G.H., A.G. and I.K.; visualization, N.P.; supervision, N.P. and A.G.H.; project administration, A.G.H. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the Hellenic Foundation for Research and Innovation (HFRI) under the 1st Call for HFRI Research Projects to Support Faculty Members and Researchers and Procure High- Value Research Equipment grant (Project Number 2563). Additionally, it was funded by the European Union - NextGenerationEU through Greece 2.0—National Recovery and Resilience Plan, under the call “Flagship actions in interdisciplinary scientific fields with a special focus on the productive fabric” (ID 16618), project name “Bridging big omic, genetic and medical data for Precision Medicine implementation in Greece” (project code TAEDR-0539180).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Dataset available on request from the authors. (https://github.com/nperdi/D-sORF; Contact: arhatzi@uth.gr).

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

The dataset comprises fifty predicted small open reading frames (sORFs) identified by the D-sORF tool as potentially functional. Each entry in the table includes the unique sORF identifier assigned by sORFs.org, the length of the sORF in nucleotides, and the translated amino acid sequence. The dataset includes sORFs from various studies, providing a comprehensive resource for understanding the functional potential of these small but significant genomic elements. Analysis of the sequences can reveal common motifs, structural properties, and potential roles in cellular processes. All the sORFs in the list are labeled by Ensembl as sORFs, as declared by sORFs.org.

sORF ID	Length	Sequence
SorfId=andreev_2015:538832	300	MVLRRLLAALLHSPQLVERLSESRPIRRAAQLTAFALLQAQLRGQDAARRLQDLAAGPVGSLCRRAERFRDAFTQELRRGLRGRSGPPPGSQRGPGANI
SorfId=elkon_2015:50819	210	MLRWLRDFVLPTAACQDAEQPTRYETLFQALDRNGDGVVDIGELQEGLRNLGIPLGQDAEEVGRRRGAA
SorfId=fritsch_2012:225652	180	MRQLKGKPKKETSKDKKERKQAMQEARQQITTVVLPTLAVVVLLIVVFVYVATRPTITE
SorfId=andreev_2015:496184	213	MAPWSREAVLSLYRALLRQGRQLRYTDRDFYFASIRREFRKNQKLEDAEARERQLEKGLVFLNGKLGRII
SorfId=rubio_2014:98637	141	MKESSIDDLMKSLDKNSDQEIDFKEYSVFLTMLCMAYNDFFLEDNK
SorfId=andreev_2015:629427	261	MTDRYTIHSQLEHLQSKYIGTGHADTTKWEWLVNQHRDSYCSYMGHFDLLNYFAIAENESKARVRFNLMEKMLQPCGPPADKPEEN
SorfId=crappe_2014:52683	249	MSATWTLSPEPLPPSTGPPVGAGLDAEQRTVFAFVLCLLVVLVLLMVRCVRILLDPYSRMPASSWTDHKEALERGQFDYALV
SorfId=calviello_2016:870882	252	MEALGSGHYVGGSIRSMAAAALSGLAVRLSRPQGTRGSYGAFCKTLTRTLLTFFDLAWRLRKNFFYFYILASVILNVHLQVYI
SorfId=andreev_2015:491102	189	MAAATLTSKLYSLLFRRTSTFALTIIVGVMFFERAFDQGADAIYDHINEGVRACAIPDLGPA
SorfId=calviello_2016:375783	189	MTKAGSKGGNLRDKLDGNELDLSLSDLNEVPVKELVSIVPRGPGAHLRRALAPTWATYFPLL
SorfId=liu_HEK_2013:259226	204	MLSRLQELRKEEETLLRLKAALHDQLNRLKVTPRPSPAPPPFLPSDHFPPTWASVVTTCCGQFSATH
SorfId=gonzalez_2014:380995	267	MSIMDHSPTTGVVTVIVILIAIAALGALILGCWCYLRLQRISQSEDEESIVGDGETKEPFLLVQYSAKGPCVERKAKLMTPNGPEVHG
SorfId=calviello_2016:540138	264	MYCLQWLLPVLLIPKPLNPALWFSHSMFMGFYLLSFLLERKPCTICALVFLAALFLICYSCWGNCFLYHCSDSPLPESAHDPGVVGT
SorfId=calviello_2016:185382	207	MAPAAAPSSLAVRASSPAATPTSYGVFCKGLSRTLLAFFELAWQLRMNFPYFYVAGSVILNIRLQVHI
SorfId=andreev_2015:177588	294	MEEISLANLDTNKLEAIAQEIYVDLIEDSCLGFCFEVHRAVKCGYFYLEFAETGSVKDFGIQPVEDKGACRLPLCSLPGEPGNGPDQQLQRSPPEFQ
SorfId=andreev_2015:409703	114	MITDVQLAIFANMLGVSLFLLVVLYHYVAVNNPKKQE
SorfId=andreev_2015:37830	279	MTKLAQWLWGLAILGSTWVALTTGALGLELPLSCQEVLWPLPAYLLVSAGCYALGTVGYRVATFHDCEDAARELQSQIQEARADLARRGLRF
SorfId=andreev_2015:13871	279	MWPVFWTVVRTYAPYVTFPVAFVVGAVGYHLEWFIRGKDPQPVEEEKSISERREDRKLDELLGKDHTQVVSLKDKLEFAPKAVLNRNRPEKN
SorfId=park_2016:1581934	237	MREENKGMPSGGGSDEGLASAAARGLVEKVRQLLEAGADPNGVNRFGRRAIQVAGAPGPRRQGARERGARPRRIGAGT
SorfId=gonzalez_2014:153430	150	VRDKKLLNDLNGAVEDAKTARLFNITSSALAASCIILVFIFLRYPLTDY

sORF ID	Length	Sequence
SorfId=fritsch_2012:974998	183	MDAVSQVPMEVVLPKHILDIWVIVLIILATIVIMTSLLLCPATAVIIYRMRTHPILSGAV
SorfId=andreev_2015:137699	264	MSTSVPQGHTWTQRVKKDDEEEDPLDQLISRSGCAASHFAVQECMAQHQDWRQCQPQVQAFKDCMSEQQARRQEELQRRQEQAGAHH
SorfId=andreev_2015:321927	207	VGVGKNKCLYALEEGIVRYTKEVYVPHPRNTEAVDLITRLPKGAVLYKTFVHVVPAKPEGTFKLVAML
SorfId=calviello_2016:206180	141	MPAGVPMSTYLKMFAASLLAMCAGAEVVHRYYRPDLVSAGGLLFYF
SorfId=gonzalez_2014:52803	162	VNIDHFTKDITMKNLVEPSLSSFDMAQKRIHALMEKDSLPRFVRSEFYQELIK
SorfId=elkon_2015:525198	114	MVVDRLFLWTFIIFTSVGTLVIFLDATYHLPPPDPFP
SorfId=andreev_2015:272691	150	MDINRRRYPAHLARSSSRKYTELPHGAISEDQAVGPADIPCDSTGQTST
SorfId=calviello_2016:187121	222	LLYLGRDYPKGADYFKKRLKNIFLKNKDVKNPEKIKELIAQGEFVMKELEALYFLRKYRAMKQRYYSDTNKTN
SorfId=loayza_puch_2016:211324	231	MDLRRVKEYFSWLYYQYQIISCCAVLEPWERSMFNTILLTIIAMVVYTAYVFIPIHIRLAWEFFSKICGYHSTISN
SorfId=gonzalez_2014:438290	78	MLDIFILMFFAIIGLVILSYIIYLL
SorfId=cenik_2015:853557	171	MADVSERTLQLSVLVAFASGVLLGWQANRLRRRYLDWRKRRLQDKLAATQKKLDLA
SorfId=andreev_2015:169859	174	MPTGKQLADIGYKTFSTSMMLLTVYGGYLCSVRVYHYFQWRRAQRQAAEEQKTSGIM
SorfId=crappe_2014:5055	144	DVGSLDEKMKSLDVNQDSELKFNEYWRLIGELAKEIRKKKDLKIRKK
SorfId=calviello_2016:673102	225	MFDIKAWAEYVVEWAAKDPYGFLTTVILALTPLFLASAVLSWKLAKMIEAREKEQKKKQKRQENIAKAKRLKKD
SorfId=ingolia_2012:368307	69	LGMEEEDVIEVYQEQTGGHSTV
SorfId=andreev_2015:611100	273	MPKRKAKGDAKGDKAKVKDEPQRRSARLSAKPAPPKPEPRPKKASAKKGEKLPKGRKGKADAGKDGNNPAKNRDASTLQSQKAEGTGDAK
SorfId=elkon_2015:4281	108	MPQTFSGGRGFELDSGSDDMDPGRPRPPRDPDELH
SorfId=calviello_2016:40163	273	MRDPVSSQYSSFLFWRMPIPELDLSELEGLGLSDTATYKVKDSSVGKMIGQATAADQEKNPEGDGLLEYSTFNFWRAPIASIHSFELDLL
SorfId=elkon_2015:564093	213	MPARRLLLLLTLLLPGLGVSDRGAWGGGQLATAGSGPGQRRGAGAGVRAGSATAAARCPVSPAVGGSGRA
SorfId=andreev_2015:217825	87	VERIKERVEEKEGIPPQQQRLIYSGKQM
SorfId=calviello_2016:764879	126	VSKCSEEIKNYIEERSGEDPLVKGIPEDKNPFKEKGSCVIS
SorfId=calviello_2016:642790	204	VKAAFTRAYNKEAHLTPYSLQAIKASRHSTSPSLDSEYNEELNEDDSQSDEKDQDAIETDAMIKVVF
SorfId=andreev_2015:308992	210	MTTLIPILSTFLFEDFSKASGFKGQRPETLHERLTLVSVYAPYLLIPFILLIFMLRSPYYKYEEKRKKK
SorfId=gonzalez_2014:172668	129	VSKAAADLMTYCDAHACEDPLITPVPTSENPFREKKFFCALL
SorfId=andreev_2015:216730	162	VIPKRTNRPGISTTDRGFPRARYRARTTNYNSSRSRFYSGFNSRPRGRVYRSG
SorfId=gonzalez_2014:645009	93	VFRYSLQKLAYTVSRTGRQVLGERRQRAPN
SorfId=werner_2015:456667	147	MVITSENDEDRGGQEKESKEESVLAMLGIIGTILNLIVIIFVYIYTTL
SorfId=iwasaki_2016:288070	114	MAEGNTLISVDYEIFGKVQGVFFRKHTQVCGLQALGW
SorfId=andreev_2015:619791	195	MMNNGLLQQPSALMLLPCRPVLTSVALNANFVSWKSRTKYTITPVKMRKSGGRDHTGGNKDRGI
SorfId=werner_2015:555128	159	MGCHSSKSTTVAAESQKLEEEREGREPGLETGTQAADCKDAPLKDGTPEPKS

References

1. Ladoukakis, E.; Pereira, V.; Magny, E.G.; Eyre-Walker, A.; Couso, J.P. Hundreds of putatively functional small open reading frames in Drosophila. *Genome Biol.* **2011**, *12*, R118. [CrossRef]
2. Bliss, M. Banting's, Best's, and Collip's accounts of the discovery of insulin. *Bull. Hist. Med.* **1982**, *56*, 554–568.
3. Wadler, C.S.; Vanderpool, C.K. A dual function for a bacterial small RNA: SgrS performs base pairing-dependent regulation and encodes a functional polypeptide. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 20454–20459. [CrossRef] [PubMed]
4. Casson, S.A.; Chilley, P.M.; Topping, J.F.; Evans, I.M.; Souter, M.A.; Lindsey, K. The POLARIS gene of Arabidopsis encodes a predicted peptide required for correct root growth and leaf vascular patterning. *Plant Cell* **2002**, *14*, 1705–1721. [CrossRef]
5. Rohrig, H.; Schmidt, J.; Miklashevichs, E.; Schell, J.; John, M. Soybean ENOD40 encodes two peptides that bind to sucrose synthase. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 1915–1920. [CrossRef]
6. Lee, D.H.; Acharya, S.S.; Kwon, M.; Drane, P.; Guan, Y.; Adelmant, G.; Kalev, P.; Shah, J.; Pellman, D.; Marto, J.A.; et al. Dephosphorylation enables the recruitment of 53BP1 to double-strand DNA breaks. *Mol. Cell* **2014**, *54*, 512–525. [CrossRef] [PubMed]
7. Dong, X.; Wang, D.; Liu, P.; Li, C.; Zhao, Q.; Zhu, D.; Yu, J. Zm908p11, encoded by a short open reading frame (sORF) gene, functions in pollen tube growth as a profilin ligand in maize. *J. Exp. Bot.* **2013**, *64*, 2359–2372. [CrossRef]
8. Kastenmayer, J.P.; Ni, L.; Chu, A.; Kitchen, L.E.; Au, W.C.; Yang, H.; Carter, C.D.; Wheeler, D.; Davis, R.W.; Boeke, J.D.; et al. Functional genomics of genes with small open reading frames (sORFs) in *S. cerevisiae*. *Genome Res.* **2006**, *16*, 365–373. [CrossRef]
9. Gleason, C.A.; Liu, Q.L.; Williamson, V.M. Silencing a candidate nematode effector gene corresponding to the tomato resistance gene Mi-1 leads to acquisition of virulence. *Mol. Plant-Microbe Interact.* **2008**, *21*, 576–585. [CrossRef]
10. Kondo, T.; Hashimoto, Y.; Kato, K.; Inagaki, S.; Hayashi, S.; Kageyama, Y. Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nat. Cell Biol.* **2007**, *9*, 660–665. [CrossRef]
11. Galindo, M.I.; Pueyo, J.I.; Fouix, S.; Bishop, S.A.; Couso, J.P. Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biol.* **2007**, *5*, e106. [CrossRef] [PubMed]
12. Hashimoto, Y.; Niikura, T.; Tajima, H.; Yasukawa, T.; Sudo, H.; Ito, Y.; Kita, Y.; Kawasumi, M.; Kouyama, K.; Doyu, M.; et al. A rescue factor abolishing neuronal cell death by a wide spectrum of familial Alzheimer's disease genes and Abeta. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 6336–6341. [CrossRef] [PubMed]
13. Lee, C.; Zeng, J.; Drew, B.G.; Sallam, T.; Martin-Montalvo, A.; Wan, J.; Kim, S.J.; Mehta, H.; Hevener, A.L.; de Cabo, R.; et al. The mitochondrial-derived peptide MOTS-c promotes metabolic homeostasis and reduces obesity and insulin resistance. *Cell Metab.* **2015**, *21*, 443–454. [CrossRef] [PubMed]
14. Pauli, A.; Norris, M.L.; Valen, E.; Chew, G.L.; Gagnon, J.A.; Zimmerman, S.; Mitchell, A.; Ma, J.; Dubrulle, J.; Reyon, D.; et al. Toddler: An embryonic signal that promotes cell movement via Apelin receptors. *Science* **2014**, *343*, 1248636. [CrossRef]
15. Chanut-Delalande, H.; Hashimoto, Y.; Pelissier-Monier, A.; Spokony, R.; Dib, A.; Kondo, T.; Bohère, J.; Niimi, K.; Latapie, Y.; Inagaki, S.; et al. Pri peptides are mediators of ecdysone for the temporal control of development. *Nat. Cell Biol.* **2014**, *16*, 1035–1044. [CrossRef] [PubMed]
16. Magny, E.G.; Pueyo, J.I.; Pearl, F.M.; Cespedes, M.A.; Niven, J.E.; Bishop, S.A.; Couso, J.P. Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. *Science* **2013**, *341*, 1116–1120. [CrossRef]
17. Tonkin, J.; Rosenthal, N. One small step for muscle: A new micropeptide regulates performance. *Cell Metab.* **2015**, *21*, 515–516. [CrossRef] [PubMed]
18. Crappé, J.; Van Criekinge, W.; Trooskens, G.; Hayakawa, E.; Luyten, W.; Baggerman, G.; Menschaert, G. Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs. *BMC Genom.* **2013**, *14*, 648. [CrossRef]
19. Andrews, S.J.; Rothnagel, J.A. Emerging evidence for functional peptides encoded by short open reading frames. *Nat. Rev. Genet.* **2014**, *15*, 193–204. [CrossRef]
20. Slavoff, S.A.; Heo, J.; Budnik, B.A.; Hanakahi, L.A.; Saghatelian, A. A human short open reading frame (sORF)-encoded polypeptide that stimulates DNA end joining. *J. Biol. Chem.* **2014**, *289*, 10950–10957. [CrossRef]
21. Yosten, G.L.C.; Liu, J.; Ji, H.; Sandberg, K.; Speth, R.; Samson, W.K. A 5′-upstream short open reading frame encoded peptide regulates angiotensin type 1a receptor production and signalling via the β-arrestin pathway. *J. Physiol.* **2016**, *594*, 1601–1605. [CrossRef] [PubMed]
22. Schwab, S.R.; Li, K.C.; Kang, C.; Shastri, N. Constitutive display of cryptic translation products by MHC class I molecules. *Science* **2003**, *301*, 1367–1371. [CrossRef] [PubMed]
23. Wang, R.F.; Parkhurst, M.R.; Kawakami, Y.; Robbins, P.F.; Rosenberg, S.A. Utilization of an alternative open reading frame of a normal gene in generating a novel human cancer antigen. *J. Exp. Med.* **1996**, *183*, 1131–1140. [CrossRef] [PubMed]
24. Ruiz-Orera, J.; Messeguer, X.; Subirana, J.A.; Alba, M.M. Long non-coding RNAs as a source of new peptides. *eLife* **2014**, *3*, e03523. [CrossRef] [PubMed]
25. McLysaght, A.; Guerzoni, D. New genes from non-coding sequence: The role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Philos. Trans. R. Soc. B Biol. Sci.* **2015**, *370*, 20140332. [CrossRef] [PubMed]
26. Strullu-Derrien, C.; Selosse, M.-A.; Kenrick, P.; Martin, F.M. The origin and evolution of mycorrhizal symbioses: From palaeomycology to phylogenomics. *New Phytol.* **2018**, *220*, 1012–1030. [CrossRef] [PubMed]

27. Cabili, M.N.; Trapnell, C.; Goff, L.; Koziol, M.; Tazon-Vega, B.; Regev, A.; Rinn, J.L. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **2011**, *25*, 1915–1927. [CrossRef]
28. Kutter, C.; Watt, S.; Stefflova, K.; Wilson, M.D.; Goncalves, A.; Ponting, C.P.; Odom, D.T.; Marques, A.C. Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genet.* **2012**, *8*, e1002841. [CrossRef] [PubMed]
29. Laing, W.A.; Martínez-Sánchez, M.; Wright, M.A.; Bulley, S.M.; Brewster, D.; Dare, A.P.; Rassam, M.; Wang, D.; Storey, R.; Macknight, R.C.; et al. An upstream open reading frame is essential for feedback regulation of ascorbate biosynthesis in Arabidopsis. *Plant Cell* **2015**, *27*, 772–786. [CrossRef]
30. Lauressergues, D.; Couzigou, J.M.; Clemente, H.S.; Martinez, Y.; Dunand, C.; Bécard, G.; Combier, J.P. Primary transcripts of microRNAs encode regulatory peptides. *Nature* **2015**, *520*, 90–93. [CrossRef]
31. Couso, J.-P.; Patraquim, P. Classification and function of small open reading frames. *Nat. Rev. Mol. Cell Biol.* **2017**, *18*, 575–589. [CrossRef] [PubMed]
32. Ingolia, N.T.; Ghaemmaghami, S.; Newman, J.R.S.; Weissman, J.S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **2009**, *324*, 218–223. [CrossRef] [PubMed]
33. Ingolia, N.T.; Lareau, L.F.; Weissman, J.S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **2011**, *147*, 789–802. [CrossRef] [PubMed]
34. Ingolia, N.T.; Brar, G.A.; Rouskin, S.; McGeachy, A.M.; Weissman, J.S. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protoc.* **2012**, *7*, 1534–1550. [CrossRef] [PubMed]
35. Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* **2003**, *422*, 198–207. [CrossRef] [PubMed]
36. Mann, M.; Jensen, O.N. Proteomic analysis of post-translational modifications. *Nat. Biotechnol.* **2003**, *21*, 255–261. [CrossRef] [PubMed]
37. Ingolia, N.T.; Brar, G.A.; Stern-Ginossar, N.; Harris, M.S.; Talhouarne, G.J.; Jackson, S.E.; Wills, M.R.; Weissman, J.S. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep.* **2014**, *8*, 1365–1379. [CrossRef] [PubMed]
38. Andreev, D.E.; O'Connor, P.B.; Fahey, C.; Kenny, E.M.; Terenin, I.M.; Dmitriev, S.E.; Cormican, P.; Morris, D.W.; Shatsky, I.N.; Baranov, P.V. Translation of 5′ leaders is pervasive in genes resistant to eIF2 repression. *eLife* **2015**, *4*, e03971. [CrossRef] [PubMed]
39. Bazzini, A.A.; Johnstone, T.G.; Christiano, R.; Mackowiak, S.D.; Obermayer, B.; Fleming, E.S.; Vejnar, C.E.; Lee, M.T.; Rajewsky, N.; Walther, T.C.; et al. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.* **2014**, *33*, 981–993. [CrossRef]
40. Olexiouk, V.; Van Criekinge, W.; Menschaert, G. An update on sORFs.org: A repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res.* **2018**, *46*, D497–D502. [CrossRef]
41. Olexiouk, V.; Crappé, J.; Verbruggen, S.; Verhegen, K.; Martens, L.; Menschaert, G. sORFs.org: A repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res.* **2016**, *44*, D324–D329. [CrossRef] [PubMed]
42. Cohen, S.M. Everything old is new again: (linc)RNAs make proteins! *EMBO J.* **2014**, *33*, 937–938. [CrossRef]
43. Raj, A.; Wang, S.H.; Shim, H.; Harpak, A.; Li, Y.I.; Engelmann, B.; Stephens, M.; Gilad, Y.; Pritchard, J.K. Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *eLife* **2016**, *5*, e13328. [CrossRef]
44. Kozak, M. An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.* **1987**, *15*, 8125–8148. [CrossRef]
45. Michel, A.M.; Kiniry, S.J.; O'Connor, P.B.F.; Mullan, J.P.; Baranov, P.V. GWIPS-viz: 2018 update. *Nucleic Acids Res.* **2018**, *46*, D823–D830. [CrossRef]
46. Li, Y.; Zhou, H.; Chen, X.; Zheng, Y.; Kang, Q.; Hao, D.; Zhang, L.; Song, T.; Luo, H.; Hao, Y.; et al. SmProt: A Reliable Repository with Comprehensive Annotation of Small Proteins Identified from Ribosome Profiling. *Genom. Proteom. Bioinform.* **2021**, *19*, 602–610. [CrossRef] [PubMed]
47. Brunet, M.A.; Brunelle, M.; Lucier, J.F.; Delcourt, V.; Levesque, M.; Grenier, F.; Samandi, S.; Leblanc, S.; Aguilar, J.D.; Dufour, P.; et al. OpenProt: A more comprehensive guide to explore eukaryotic coding potential and proteomes. *Nucleic Acids Res.* **2019**, *47*, D403–D410. [CrossRef]
48. Chen, Y.; Long, W.; Yang, L.; Zhao, Y.; Wu, X.; Li, M.; Du, F.; Chen, Y.; Yang, Z.; Wen, Q.; et al. Functional Peptides Encoded by Long Non-Coding RNAs in Gastrointestinal Cancer. *Front. Oncol.* **2021**, *11*, 777374. [CrossRef] [PubMed]
49. Leong, H.S.; Dawson, K.; Wirth, C.; Li, Y.; Connolly, Y.; Smith, D.L.; Wilkinson, C.R.; Miller, C.J. A global non-coding RNA system modulates fission yeast protein levels in response to stress. *Nat. Commun.* **2014**, *5*, 3947. [CrossRef]
50. Mazin, P.V.; Fisunov, G.Y.; Gorbachev, A.Y.; Kapitskaya, K.Y.; Altukhov, I.A.; Semashko, T.A.; Alexeev, D.G.; Govorun, V.M. Transcriptome analysis reveals novel regulatory mechanisms in a genome-reduced bacterium. *Nucleic Acids Res.* **2014**, *42*, 13254–13268. [CrossRef]
51. Giannakakis, A.; Zhang, J.; Jenjaroenpun, P.; Nama, S.; Zainolabidin, N.; Aau, M.Y.; Yarmishyn, A.A.; Vaz, C.; Ivshina, A.V.; Grinchuk, O.V.; et al. Contrasting expression patterns of coding and noncoding parts of the human genome upon oxidative stress. *Sci. Rep.* **2015**, *5*, 9737. [CrossRef]
52. Pircher, A.; Gebetsberger, J.; Polacek, N. Ribosome-associated ncRNAs: An emerging class of translation regulators. *RNA Biol.* **2014**, *11*, 1335–1339. [CrossRef]
53. Slavoff, S.A.; Mitchell, A.J.; Schwaid, A.G.; Cabili, M.N.; Ma, J.; Levin, J.Z.; Karger, A.D.; Budnik, B.A.; Rinn, J.L.; Saghatelian, A. Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.* **2013**, *9*, 59–64. [CrossRef] [PubMed]

54. Hurst, L.D. The Ka/Ks ratio: Diagnosing the form of sequence evolution. *Trends Genet.* **2002**, *18*, 486. [CrossRef]
55. Pollard, K.S.; Hubisz, M.J.; Rosenbloom, K.R.; Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **2010**, *20*, 110–121. [CrossRef]
56. Hanada, K.; Zhang, X.; Borevitz, J.O.; Li, W.-H.; Shiu, S.-H. A large number of novel coding small open reading frames in the intergenic regions of the Arabidopsis thaliana genome are transcribed and/or under purifying selection. *Genome Res.* **2007**, *17*, 632–640. [CrossRef]
57. Hanada, K.; Akiyama, K.; Sakurai, T.; Toyoda, T.; Shinozaki, K.; Shiu, S.-H. sORF finder: A program package to identify small open reading frames with high coding potential. *Bioinformatics* **2010**, *26*, 399–400. [CrossRef]
58. Chugunova, A.; Navalayeu, T.; Dontsova, O.; Sergiev, P. Mining for Small Translated ORFs. *J. Proteome Res.* **2018**, *17*, 1–11. [CrossRef] [PubMed]
59. O'Leary, N.A.; Wright, M.W.; Brister, J.R.; Ciufo, S.; Haddad, D.; McVeigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D.; et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **2016**, *44*, D733–D745. [CrossRef]
60. Siepel, A.; Bejerano, G.; Pedersen, J.S.; Hinrichs, A.S.; Hou, M.; Rosenbloom, K.; Clawson, H.; Spieth, J.; Hillier, L.W.; Richards, S.; et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **2005**, *15*, 1034–1050. [CrossRef]
61. McGeoch, D.J. On the predictive recognition of signal peptide sequences. *Virus Res.* **1985**, *3*, 271–286. [CrossRef] [PubMed]
62. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
63. Fix, E.; Hodges, J.L. *Discriminatory Analysis: Nonparametric Discrimination: Consistency Properties*; USAF School of Aviation Medicine: Randolf Fiend, TX, USA, 1951.
64. Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J.C.; Sheridan, R.P.; Feuston, B.P. Random forest: A classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958. [CrossRef] [PubMed]
65. Dever, T.E.; Ivanov, I.P.; Hinnebusch, A.G. Translational regulation by uORFs and start codon selection stringency. *Genes Dev.* **2023**, *37*, 474–489. [CrossRef] [PubMed]
66. Liu, H.; Hao, W.; Yang, J.; Zhang, Y.; Wang, X.; Zhang, C. Emerging roles and potential clinical applications of translatable circular RNAs in cancer and other human diseases. *Genes Dis.* **2023**, *10*, 1994–2012. [CrossRef]
67. Mudge, J.M.; Ruiz-Orera, J.; Prensner, J.R.; Brunet, M.A.; Calvet, F.; Jungreis, I.; Gonzalez, J.M.; Magrane, M.; Martinez, T.F.; Schulz, J.F.; et al. Standardized annotation of translated open reading frames. *Nat. Biotechnol.* **2022**, *40*, 994–999. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.