Molecular Therapy
Nucleic Acids
Review

A review from biological mapping to computation-based subcellular localization

Jing Li,[1,2] Quan Zou,[1] and Lei Yuan[3]

[1]Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, 1 Chengdian Road, Quzhou, Zhejiang 324000, China; [2]School of Biomedical Sciences, University of Hong Kong, Hong Kong, China; [3]Department of Hepatobiliary Surgery, Quzhou People's Hospital, 100 Minjiang Main Road, Quzhou, Zhejiang 324000, China

Subcellular localization is crucial to the study of virus and diseases. Specifically, research on protein subcellular localization can help identify clues between virus and host cells that can aid in the design of targeted drugs. Research on RNA subcellular localization is significant for human diseases (such as Alzheimer's disease, colon cancer, etc.). To date, only reviews addressing subcellular localization of proteins have been published, which are outdated for reference, and reviews of RNA subcellular localization are not comprehensive. Therefore, we collated (the most up-to-date) literature on protein and RNA subcellular localization to help researchers understand changes in the field of protein and RNA subcellular localization. Extensive and complete methods for constructing subcellular localization models have also been summarized, which can help readers understand the changes in application of biotechnology and computer science in subcellular localization research and explore how to use biological data to construct improved subcellular localization models. This paper is the first review to cover both protein subcellular localization and RNA subcellular localization. We urge researchers from biology and computational biology to jointly pay attention to transformation patterns, interrelationships, differences, and causality of protein subcellular localization and RNA subcellular localization.

INTRODUCTION

Proteins synthesized on ribosomes must be transported to their corresponding subcellular structures to perform normal biological functions. If the proteins are not in the correct locations, serious disease may result (Figure 1). For instance, protein retention in the endoplasmic reticulum can lead to diabetes and eyelid albinism, protein accumulation in the endoplasmic reticulum can lead to abnormal signal transduction (Alzheimer's disease), and abnormal transporters can lead to delayed spinal epiphyseal dysplasia.[1] In addition, research on protein subcellular localization (SCL) is beneficial for designing targeted drugs.

SCL of proteins is helpful to predict protein function, reveal molecular interaction mechanisms, and understand complex physiological processes, which make the study of protein SCL greatly significant to cell biology, proteomics, and drug design. Proteins are usually transcribed by RNA molecules, which are distributed in different parts of the cell. Therefore, the corresponding RNA molecules can be used to determine the locations of certain proteins.[2] Protein SCL and RNA SCL are closely related to human diseases. For example, microtubule-associated protein tau (MAPT) is a related protein that encodes tau. Exon 10 encodes the fourth of four microtubule-binding domains (R), and disruption of the balance between 4R-tau and 3R-tau isoforms leads to the phosphorylation of tau protein, which is the authoritative hallmark of Alzheimer's disease.[3] In addition, the translocation of RNA can lead to the generation of fragile X syndrome and fragile X-associated tremor ataxia syndrome (FXTAS) (Figure 1).[4]

SCL is significant for the study of antiviral clues and the discovery of targeted drugs. In the following, SARS-CoV-2 is used as an example to explain the relationship between virus and SCL. The outbreak of the COVID-19 pandemic in 2019 severely affected human health and life. The novel coronavirus threatens people mainly by affecting their respiratory system.[5] In addition, the novel coronavirus may also cause sepsis, acute heart damage, and multiple organ dysfunction in people with poor resistance. The length of coronavirus RNA genomes range from 26 to 32 kb.[6] SARS-CoV-2 virus binds to host cell surface receptors via the spike (S) protein during entry into host cells.[7] SARS-CoV-2 enters host cells mainly by using S protein to infect human cells. The S protein binds to the host receptor via the receptor-binding domain (RBD) in the S1 subunit, and subsequently, the S2 subunit fuses with the cell membrane. The angiotensin-converting enzyme 2 (ACE2) protein of human cells binds to the S protein of SARS-CoV-2 to complete the infection process. In other words, research on the localization of human ACE2 protein can aid in understanding infection of the human body by the novel coronavirus.[8] Conserved enzymes such as main protease or 3C-like protease (Mpro or 3CLpro), papain-like protease (PLpro), nonstructural protein 12 (nsp12), and RNA-dependent RNA polymerase (RdRP) can be used as drug targets to help researchers search for anti-coronavirus drugs.[9] In other words, research on RNA SCL (genomics research)

https://doi.org/10.1016/j.omtn.2023.04.015.

Correspondence: Quan Zou, Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, 1 Chengdian Road, Quzhou, Zhejiang 324000, China.
E-mail: zouquan@nclab.net
Correspondence: Lei Yuan, Department of Hepatobiliary Surgery, Quzhou People's Hospital, 100 Minjiang Main Road, Quzhou, Zhejiang 324000, China.
E-mail: icbbsuse@sina.com

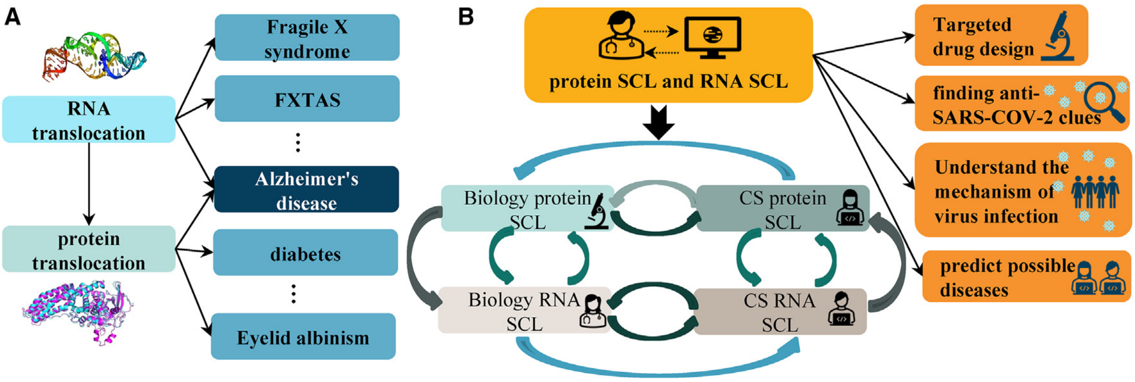

Figure 1. Research significance of subcellular localization problem
(A) Protein and RNA dislocation and disease relationship. (B) Significance of the subcellular localization of B proteins and RNA.

can help researchers identify routes of coronavirus infection into the human body, which can lead to the development of safer and more effective strategies (Figure 1).[10]

Subcellular location errors of RNA transcripts not only limit the time and space of translation but also lead to protein toxicity and cellular error responses.[11] In addition, the spatial distribution of RNA affects the place of translation, leading to the differences of protein concentration and location. In turn, these proteins limit cell function and response to the environment (Figure 1).

In SCL problems, biological sequences are the main form of data. To date, there are multiple publicly available RNA SCL databases. These databases were filtered and integrated several times to generate a relative centralized and complete database. In contrast, there is almost no separate database of protein SCL. Protein SCL data are included in a comprehensive database, which can hinder data collection for studying protein SCL. See supplemental information for changes in RNA SCL database and protein SCL database (Table 1; Figure 2).

To our knowledge, there have been reviews only on protein SCL,[28–33] and no literature has integrated both RNA SCL and protein SCL. Furthermore, these reviews did not delve into the intrinsic links between proteins in biology and computational biology. The advantages of our review are as follows:

(1) This is the first review that covers both protein SCL and RNA SCL, including changes in protein and RNA SCL databases and changes in computer algorithms for protein and RNA SCL problems.
(2) A general approach to protein and RNA SCL on the basis of traditional machine learning is summarized, which provides a reference for biological researchers to help them find more effective strategies for SCL from a biological perspective.
(3) Some suggestions on the construction, verification, and maintenance of SCL models are given.
(4) Future research directions are suggested (i.e., discussing protein SCL and RNA in the mode transformation and mutual relationship between traditional biology and computational biology), which is very important for the thorough study of SCL.

RELATED DATABASE

Protein SCL database

Proteins are an indispensable structural component of the human body and participate in various life activities. Understanding the positions of proteins in cells is helpful for understanding the mechanism of protein activity, which is important for the design of target protein drugs and the study of biology.

UniProt[12] (Table 1) is a comprehensive protein database containing more than 500,000 pieces of protein information (protein SCL, protein structure, and interactions). The database incorporates 5 million virus-related proteins. LOCATE[13] is a mouse protein subcellular location database published in 2006. LOCATE′[14] is a database of SCLs of proteins for mouse and human species published in 2008. LOCATE′ absorbed data from LIFEdb,[34] Mouse Genome Informatics,[35] UniProt,[36] and Ensembl.[37] LOCATE has the advantage of containing automatic classification calculations, experimental localization images and identification of protein sorting signals. eSLDB[15] is a protein SCL database for species including *Homo sapiens*, *Mus musculus*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, and *Arabidopsis thaliana*. The main source of data is SwissProt. The biggest advantage of eSLDB is that it uses a more detailed positioning method. eSLDB uses more detailed localization methods, including experimentally determined, homology-based, and predicted. PSORTb 3.0[16] is a protein SCL database for bacteria. PSORTdb[16–18] is a SCL database of bacterial and archaea proteins that provides certain clues for drug research and development. PSORTdb is the latest version of Gram-negative bacteria protein SCL that supports subcellular location prediction simultaneously. Notably, the database that was no longer accessible is abandoned.

RNA SCL database

With the development of technology, an increasing number of databases have been constructed, which provide the data basis for computational protein SCL and RNA SCL methods (Table 1 and Figure 2).

Table 1. Protein database and RNA database

Database	Link	Time	Reference	Size	Advantage	Variety
Protein database						
UniProt	https://www.uniprot.org	2019	UniProt Consortium[12]	5,000,000	focus on the improving the number of viral reference proteomes	–
LOCATE	http://locate.imb.uq.edu.au	2006, 2008	Fink et al. and Sprenger et al.[13,14]	122,765	containing automatic calculation of classification, experimental positioning of image data	human and membrane.
eSLDB	http://gpcr.biocomp.unibo.it/esldb/	2007	Pierleoni et al.[15]	143,222	more detailed positioning methods	Homo sapiens, Mus musculus, Caenorhabditis elegans, Saccharomyces cerevisiae, and Arabidopsis thaliana
PSORTdb	http://db.psort.org	2016	Rey et al., Yu et al., Peabody et al. and Gardy et al.[16–19]	1,443	up-to-date database of Gram-negative bacteria	Grame-negative bacteria
RNA database						
lncRNAdb	http://www.lncrnadb.org	2011	Amaral et al.[20]	150	containing a comprehensive list of lncRNA	–
lncRNAdb version 2.0	http://lncrnadb.org	2011, 2015	Heinzelmann et al. and Quek et al.[21,22]	283	data from 287 papers and the database are managed manually	–
lncATLAS	https://www.encodeproject.org	2017	Mas-Ponte et al.[23]	6,768	specifically for human cells	–
LncSLdb	http://bioinformatics.xidian.edu.cn/lncSLdb	2018	Wen et al.[24]	11,000	it was most complete at that time the data can be searched by gene symbol, genomic coordinates, and sequence similarity	–
EVmiRNA	http://bioinfo.life.hust.edu.cn/EVmiRNA	2018	Liu et al.[25]	1,000	the first database focusing on miRNA expression profiles of extracellular vesicles	–
RNALocate version 2.0	http://www.rnalocate.org	2017, 2022	Zhang et al. and Cui et al.[26,27]	213,000	with prediction tools	–

lncRNAdb version 2.0[21,22] and lncATLAS[23] belong to the long coding RNA database. Among these RNAs, each RNA sequence in lncRNAdb contains structural information and SCL information. lncATLAS contains 6,768 long noncoding RNA (lncRNA) sequences distributed in 15 cell sites. RNA sequences were obtained using RNA sequencing techniques. lncSLdb[24] is a lncRNA SCL database. Specifically, the database is based on data from three species and covers more than 11,000 transcripts. Among these transcripts, subcellular sites include chromosomes, ribosomes, and cytoplasm. lncSLdb includes entries from lncRNAdb,[20] RNALocate,[26] and lncATLAS, and it requires glycine lengths of more than 200 nt. In addition, the data can be searched by gene symbol, genomic coordinates, and sequence similarity. In biology, micrononcoding RNAs (miRNAs) are involved in posttranscriptional regulation. Research has shown that the miRNAs in normal and diseased cells were different.[21,38,39] Further understanding of miRNAs can help scholars improve their understanding of the pathological mechanism. EVmiRNA[25] is an miRNA database that includes the relationship between miRNA and disease, which is the first database focusing on miRNA expression profiles of extracellular vesicles.

RNALocate version 2.0[27] is an RNA SCL database developed through manual management that includes more than 213 thousand samples and covers 104 species and 171 SCLs. In addition, RNALocate also includes lncSLdb, lncATLAS, and EVmiRNA samples, which allows RNALocate to cover a wider range of RNA types, species, and SCLs. RNALocate is the primary version of RNALocate version 2.0. To date, RNALocate version 2.0 is the latest and most comprehensive RNA SCL database.

As we can see from Figure 2, there are references between the mainstream RNA databases. lncRNA localization data are an important part of RNA SCL. lncRNAdb, lncRNAdb version 2.0, lncATLAS, and lncSLdb are all SCL databases of lncRNAs.

SCL MODELS

SCL models based on traditional machine learning

Protein SCL models based on traditional machine learning

In the past half century, breakthrough progress has been made in the SCL of proteins on the basis of traditional machine learning (Table 2; Figure 3). In 1994, Nakashima and Nishikawa[58] first tried to distinguish intracellular and extracellular proteins by amino acid

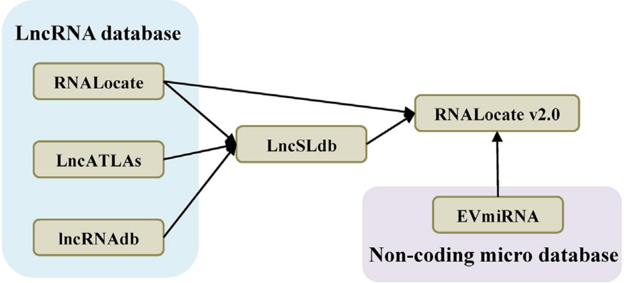

Figure 2. Cross-referencing of databases

composition (AAC) and achieved good results. Hua and Sun[59] established the first protein SCL prediction system on the basis of a support vector machine (SVM) in 2001 and 2002 on the basis of previous studies. Although the SCL models based on traditional machine learning before 2017 are of great significance for research on the SCL problem, only those after 2017 are presented because of limited space (pre-2017 models are shown in supplemental information section 3 and Table S1).

Models for SCL of eukaryotic proteins. In view of the special biological functions of multilocus proteins, the gposc-ecc-mplc model and gnegc-ecc-mploc model were constructed to generate multiple protein SCL models of Gram-negative bacteria. The pLoc_bal_mEuk model is designed to identify the eukaryotic protein SCL, which decreases the prediction bias to a certain extent through data balance processing.[60]

Models for SCL of bacterial proteins. PSORT, which predicts protein SCL on the basis of amino acid sequence and source information, is one of the early methods to use computer technology for protein SCL. An apolar algorithm was used to build the PSORT model.[61] Subsequently, the decision algorithm was used to improve PSORT to generate Psort-II.[62] Shen et al.[29] used the multikernel SVM model to construct a protein SCL model, which is a multilabel classification model. The PsePSSM, DWT, and AvBlock were used to extract features. The integrated kernel was applied to train the SVM. Experimental results showed that combining features can promote the ability of the model. The pLoc_bal-mGops[43] model used hypothesis training samples IHTS to solve the data imbalance problem by adding some hypothesis or theory samples to a smaller subset. PseAAC theory was used to extract features and combined with multilabel Gaussian regression ML-GKR to construct the classifier.

Models for SCL of animals and plants. Although the Ploco-Manimal model has advantages in protein prediction, the model is trained on an extremely imbalanced dataset. Therefore, the model will inevitably have bias.[40] Ding et al.[41] constructed a model for the human protein SCL. Protein SCL is a multilabel classification problem that is transformed into a binary classification for analysis. The PsePSSM, PSSM-DWT, PSSM-AB, PsePP, PP-DWT, and PP-AB were applied for feature extraction. Subsequently, the features were used as kernels and combined with a multikernel learning on the basis of kernel target alignment (KTA-MKL). The KNB algorithm was used to train the model. Pan et al. constructed a protein SCL model using an embedding method. The model is based on protein-protein information and uses minimum redundancy maximum relevance (MRMR) for feature selection to further analyze important embedded features. Plant-msubp is a protein SCL model based on ensemble learning. The dataset was derived from the UniProt database and searched using the keywords "SUBCELLULAR LOCATION AND Review: Yes." Those protein sequences marked "PROBABLE," "POSSIBLE," AND "BY SIMILARITY" were removed. Three methods were used to extract features, including AAC, DIPEP, PseAAC, and NCC AAC. Then, these feature matrices were input into the SVM for training to generate the classification model.[48] Considering that other protein SCL models only analyze the features of the protein, Liu et al.[46] proposed the embedding algorithm, which can take nodes as coding information. The embedding algorithm can link the network with the traditional classification algorithm, which provides a new idea for the problem of protein SCL. On the basis of this link, Mashup was trained as a protein SCL model. Only the optimal feature subset and classifier were applied to train the model. The synthetic minority oversampling technique (SMOTE) was used to balance the training dataset. The decision tree, k-nearest neighbor (KNN), random forest (RF), and SVM were used for classification. The experimental result found that the trained model with features after selection was better. Therefore, feature selection is a significant step for protein SCL.[47] Given the inability of neural networks to accurately interpret features within the model, IPSORT was developed to achieve the accuracy of TargetP using the amino acid index and alphabet indexing to concatenate the approximate pattern.[63] The pLoc_bal-mAnimal[42] and Ploc_bal-mHum[44] models were built using the same methodology as pLoc_bal-mGops.[43] In addition, there are separate SCL models for viruses and yeasts. The pLoc_bal-mVirus[45] constructed the model in the same way that pLoc_bal-mGops[43] does. Bayes was used in the study of protein SCL. Drawid et al.[64] extracted the interval probability of proteins on the basis of 30 different features to generate a Bayesian system, which was tested on yeast genes.

RNA SCL models based on traditional machine learning

RNA SCL models contain mainly lncRNA and mRNA models.

Models for SCL of lncRNA. lncRNAs are found in the nucleus, chromatin, and cytoplasm of cells. With the development of biotechnology and computer science, an increasing number of lncRNA sequences have been discovered. Because of the low conservative nature of lncRNAs, functional annotation for lncRNAs is difficult. Therefore, it is urgent to annotate lncRNAs with computational techniques (Table 2; Figure 3).

The training and testing data for Lnlocate[49] are sourced from the RNALocate database. The dataset includes 301, 152, and 25 RNA sequences in the cytoplasm, nucleus, and exosome, respectively. Lnlocate uses the SMOTE technique to integrate unbalanced data into balanced data. Lnlocate first uses the k-mer to learn elementary

Table 2. Models of protein SCL and RNA SCL based on traditional machine learning

Model/author	Species	Algorithm	Time	Reference
Models of protein SCL				
pLoc-mAnimal	Animals	ML-GKR	2017	Cheng et al.[40]
Ding	human	KTA-MKL	2017	Ding et al.[41]
pLoc_bal-mAnimal	animal	ML-GKR	2019	Cheng et al.[42]
pLoc_bal_mGops	Gram-positive bacteria	ML-GKR	2019	Xiao et al.[43]
Ploc_bal-mHum	human	ML-GKR	2019	Chou et al.[44]
pLoc_bal-mVirus	virus	ML-GKR	2019	Xiao et al.[45]
shen	Gram-negative bacteria	SVM	2020	Shen et al.[29]
Liu	animal, plant, fungi, GN bacteria, and Gram-positive bacteria	SVM/RF	2021	Liu et al.[46]
Pan	human	DT, KNN, SVM	2021	Pan et al.[47]
Plant-msubp	plant	SVM	2021	Sahu et al.[48]
Models of RNA SCL				
lncLocator	–	RF, SVM	2018	Cao et al.[49]
iLoc-lncRNA	–	SVM	2018	Su et al.[50]
Locate-R	–	LD-SVM	2020	Ahmad et al.[51]
mRNALocater	–	SVM	2021	Tang et al.[52]
SubLocEP	–	LGBM	2021	Li et al.[53]
zhang	–	LGBM, XGBoost, CatBoost	2022	Zhang et al.[54]
TACOS	–	tree stacking	2022	Jeon et al.[55]
RNAlight	–	LGBM	2023	Yuan et al.[56]
Shubham	–	XGBoost	2023	Choudhury et al.[57]

features and then uses an unsupervised stack encoder AE to extract more advanced features. Then, RF and SVM are used for the basic classifier. Finally, the four basic classifiers are presented (including k-mer-based RF, K-mer-based SVM, AE-based RF and AE-based SVM). The output is fed into a three-layer neural network model. Lnlocate is a model that combines traditional machine learning and deep learning. ILoc-lncRNA[50] integrated nucleotide characteristics through binomial distribution and used the features for lncRNA prediction. The ILoc-lncRNA dataset is from RNALocate. When processing the data, the CD-Hit was required to be greater than 80%, resulting in 655 RNA sequences (the number of cytoplasmic, ribosome, and exosome is 426, 43, and 30, respectively). In addition, ILoc-lncRNA performed feature analysis using principal-component analysis (PCA), ANOVA, MRMR, and diffusion graphs and combined the retained features with SVM to construct the classifier. Locate-R[51] is a SCL model targeting long noncoding regions. The dataset is from RNALocate. CD-Hit is required to be greater than 80%, and the final dataset is exactly consistent with ILoc-lncRNA. Locate-R uses SMOTE to balance the class. Locate-R extracts the features of l-mers and n-gap l-mers and inputs the selected features into the local depth SVM (LD-SVM) to train the model. TACOS is a novel approach for cell-specific lncRNA SCL, which is the first application based on tree stacking and involves SCL of 10 different cell types. TACOS fuses 10 different feature descriptors and uses an appropriate tree to train the model.[55] RNAlight is an RNA SCL model based on light gradient boosting machine (LGBM) for identifying nucleotide k-aggregates that contribute to mRNA and lncRNA SCL. In addition, RNAlight was extended to other types of RNA SCL.[56]

Models for SCL of mRNA. MRNALocator is an mRNA SCL model. The electron-ion interaction pseudopotential (PseEIIP) values of trinucleotides and the pseudo-k-tuple nucleotide composition (PseKNC) were applied to extract features from RNA sequences. Subsequently, the nonfeature selection scheme is used to select features. The selected features are combined with LGBM, XGBoost, and CatBoost to train the mRNALocator.[52] SubLocEP[53] is a two-layer integrated model on the basis of machine learning, and the base classifiers are trained by LGBM. SubCLocEP is a SCL classification model for predicting prokaryotic mRNAs that uses both opportunistic sequence characteristics and physicochemical properties. To address human mRNA SCL, Zhang et al.[54] constructed a model based on traditional machine learning. MRNA subcellular location data was obtained from the RNALocate database. To achieve low data redundancy, CD-HIT-EST was applied to cluster the data with an 80% cutoff value. K-tuple (k-mer) nucleotide composition, pseudo-k-tuple nucleotide composition, position correlation scoring function and binary coding were utilized for feature extraction. To reduce the influence of noise, incremental feature selection was used to determine the optimal feature sets. The selected features are input into the SVM to generate the predictor. Choudhury et al.[57] evaluated both models on the basis of machine learning and deep learning, and finally determined on a hybrid technique that combines the XGBoost model and subject search. Short time is one of the main characteristics of this model.

Construction of SCL model based on traditional machine learning

After further insight into the protein SCL model and the RNA SCL model, we summarized the traditional machine learning method-based modeling for SCL. We expect to provide a clear research direction for biomedical researchers (Figure 4).

Data pretreatment

To date, the data types for SCL problems have been sequences (including protein sequences and RNA sequences). Raw data generally contain missing values, NAN values, and unreasonable values. Therefore, data cleaning is needed to improve data quality and the possibility of data mining. SCL is generally a multiclassification

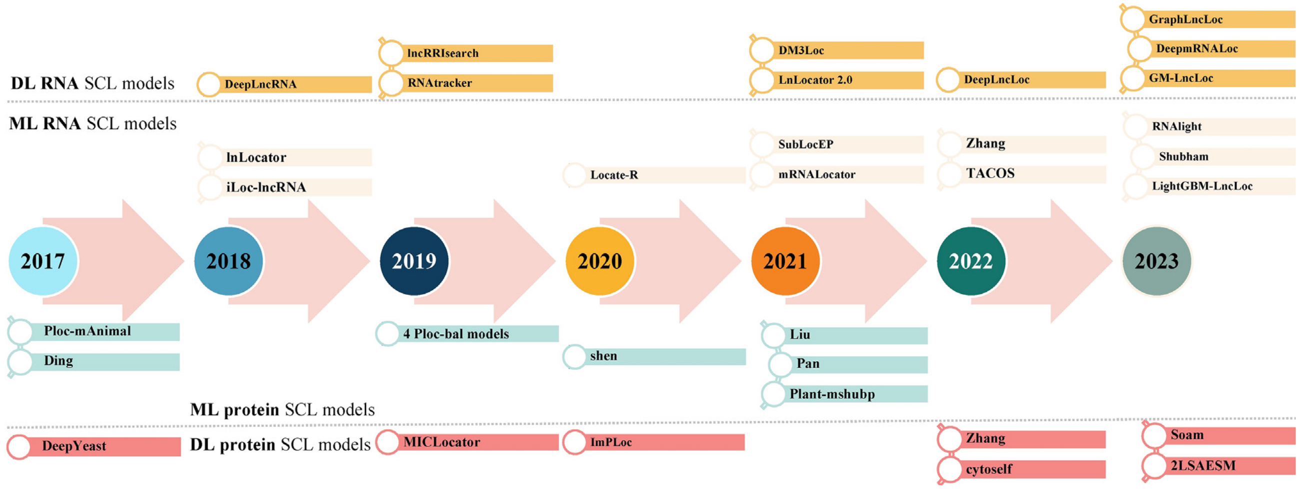

Figure 3. Published timeline of protein and RNA subcellular localization models

(DL RNA models (deep learning RNA models), ML RNA models (machine learning RNA models), ML protein models (machine learning protein models), DL protein models (deep learning protein models)).

problem, where the data imbalance problem will lead to serious bias of the model. In the training set, the data imbalance will cause the computer to focus on the data of rich classes. At this point, the model will be biased. Therefore, reducing the impact of data imbalance becomes critical.

Undersampling and oversampling are common approaches. Undersampling achieves data balance by reducing the number of samples of rich classes. The balanced data are constructed by saving all the rare samples and randomly selecting the same number of rare samples in the rich category. This approach is more reasonable when the amount of data is large enough. Obviously, the disadvantage of undersampling is that large amounts of data will be lost, even if the model learns only part of the overall pattern.[65] The principle of Easy-Ensemble is to undersample the original data many times, generate multiple different balanced training sets, and then train multiple different classification sets. The result can be obtained by combining the output of multiple classification sets.[66] Near-Miss uses KNN to select the most representative sample. In essence, Near-Miss is a prototype selection method, which means that the most representative samples are selected from rich classes for training, and it requires considerable computation.[67] In addition, clustering is also one of the directions. The majority class is clustered (the number of targets in the cluster is the same as the number of samples in the rare class). The center point of each cluster is taken as the new sample of rich classes, and all the samples of rare classes are combined to generate a balanced training set.[68]

Accordingly, oversampling is a more appropriate method when the amount of data is insufficient. The principle of oversampling is to balance the dataset by adding the rare samples. Rare samples are generated by using repetition, bootstrapping, or synthetic minority oversampling. In essence, neither undersampling nor oversampling has absolute advantages. The application of oversampling and undersampling depends on the target and the dataset. The combination of oversampling and undersampling is also one of the feasible directions. The Monte Carlo sample expansion method is a computational method whose principle is to identify the system through many random samples and then obtain the value to be calculated. The advantages of the Monte Carlo sample extension method are that it is simple, easy to understand and flexible.[69] SMOTE[70] is an improved scheme based on a random oversampling algorithm, which is usual to solve imbalanced data problems and has been widely accepted. The principle of the SMOTE algorithm is to analyze and simulate rare class samples and add artificial simulated new samples to the dataset to solve the imbalance problem in the original data. KNN technology is used in the simulation process of the SMOTE algorithm. The steps of generating new samples are as follows: (1) select n rare samples randomly, (2) find the rare samples of the initial expansion, (3) find the m rare samples closest to the initial sample above, and (4) select any point in the nearest m rare samples. The point is the newly added sample data.

The trait of adaptive synthetic sampling (ADASYN)[71] is to automatically conform the number of synthetic samples on the basis of the distribution of data. Instead of SMOTE, which roughly synthesizes the same number of rare class samples. The steps are as follows: (1) calculate the total number of samples G to be synthesized; (2) for each minority sample xi, determine its k-neighbor points and calculate the distribution proportion Γ; (3) and calculate the number of samples gi to be synthesized for each rare sample xi. Then, the SMOTE algorithm is used to synthesize new samples.

Feature extraction

Feature extraction is the process of converting raw data that cannot be recognized by traditional machine learning algorithms into numerical features that can be recognized by the algorithm. In traditional

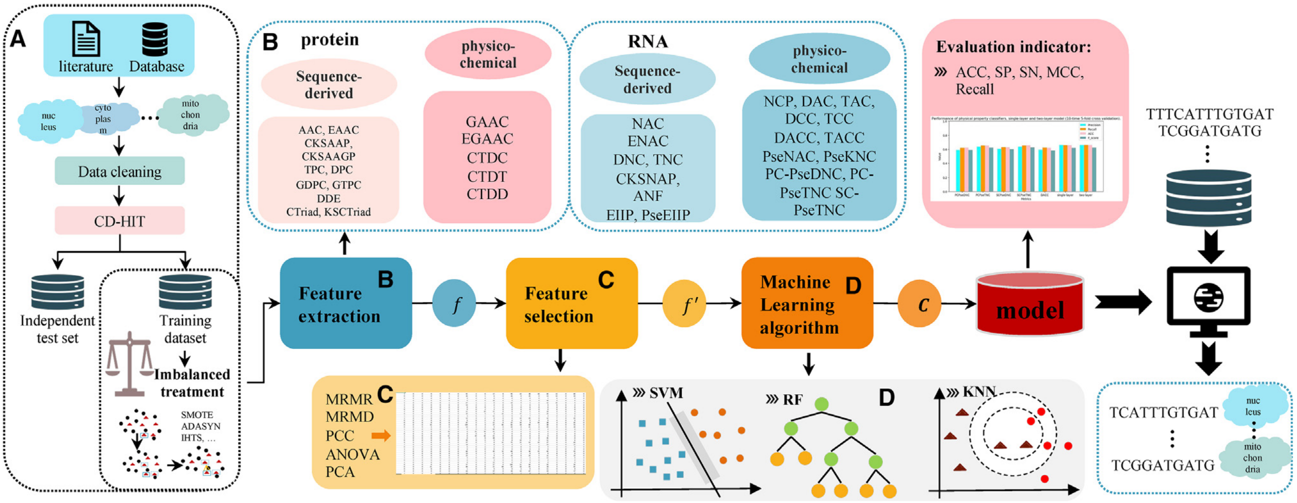

Figure 4. Flowchart of the protein and RNA subcellular localization model

(A) Data progress; (B) feature extraction; (C) feature selection; (D) machine learning algorithm (amino acid composition (AAC), Enhanced amino acid composition (EAAC), The composition of the k-spaced amino acid pair (CKSAAP), Composition of k-Spaced Amino Acid Group Pairs (CKSAAGP), tripeptide composition (TPC), dipeptide composition (DPC), Grouped dipeptide composition (GDPC), Grouped Tripeptide Composition (GTPC), conjoint triad (CTriad), Grouped amino acid composition (GAAC), Enhanced GAAC (EGAAC), The principle of nucleic acid composition (NAC), Dinucleotide Composition (DNC), Trinucleotide Composition (TNC), Enhanced nucleic acid composition (ENAC), The composition of k-spaced nucleic acid pairs (CKSNAP) accumulated nucleotide frequency (ANF), Electron-ion interaction pseudopotentials of trinucleotide (EIIP), Electron-ion interaction pseudopotentials of trinucleotide (PseEIIP), Nucleotide chemical property (NCP), Dinucleotide-based autocovariance (DAC), trinucleotide-based auto covariance (TAC), Dinucleotide-based Cross Covariance (DCC), Trinucleotide-based Cross Covariance (TCC), dinucleotide-based autocross covariance (DACC), Trinucleotide-based AutoCross Covariance (TACC). Pseudonucleic acid composition (PseNAC), pseudo-k-tuple nucleotide composition (PseKNC), parallel correlation pseudodinucleotide composition (PC-PseDNC), parallel correlation pseudotrinucleotide composition (PC-PseTNC), series correlation pseudodinucleotide composition (SC-PseDNC), and series correlation pseudotrinucleotide composition (SC-PseTNC). For more details, please check the section of feature extraction based on protein sequence and feature extraction methods for RNA.

machine learning, effective features are significant and are tightly connected with the performance of models.

Protein feature extraction methods

Feature extraction based on protein sequence

The protein sequence is mainly composed of 20 common amino acids, and the differences in different types of proteins can be directly reflected in the sequence.

The main purpose of AAC is to calculate the frequency of amino acid types in peptide sequences or proteins, and the feature dimension is 20 (because there are 20 amino acids). Enhanced AAC (EAAC) was developed on the basis of AAC, whose principle is to calculate AAC on the basis of a sequence window of fixed length (sequence sliding continuously from the N terminus to the C terminus). The composition of the k-spaced amino acid pair (CKSAAP) coding principle calculates the frequency residual of any split pair of residues (the maximum spacing is 5). Composition of k-spaced amino acid group pairs (CKSAAGP) is an advanced version of CKSAAP (maximum disability separation increased to 25). The purpose of tripeptide composition (TPC) and dipeptide composition (DPC) is to calculate tripeptide and dipeptide components, which have dimensions of 8,000 and 400, respectively. Grouped DPC (GDPC) is an advanced version of DPC that consists of 25 descriptors and calculates the ratio of amino acid type groups to the number of dipeptides. Accordingly, grouped TPC (GTPC) is an advanced version of TPC that outputs 125 descriptors that calculate the ratio of amino acid type groups to the number of tripeptides. Dipeptide deviation from expected mean (DDE) is based on the dipeptide layer, theoretical mean, and theoretical variance to extract features. BINARY uses a binary code to represent each amino acid, which is typically used to encode peptides of equal length. The coding principle of the conjoint triad (CTriad) is carried out by treating three consecutive amino acids as a single unit, considering the properties of amino acids and their neighbors. k-spaced conjoint triad (KSCTriad), a variant of CTriad, counts not only three consecutive amino acids but also consecutive amino acids separated by any number of residues.[72,73]

Feature extraction based on the physical and chemical properties of proteins

Grouped AAC (GAAC) is a feature extraction method based on physical and chemical properties, whose purpose is to calculate the frequency of each amino acid group. Enhanced GAAC (EGAAC) is an enhanced version of GAAC that adds fixed-length windows. In other words, EGAAC calculates the GAAC based on windows of fixed length. CTDC calculates the proportion of polar, neutral, and hydrophobic residues in a protein. CTDT calculates the percentage frequency of the neutral residue after the polar residue or in reverse. CTDD calculates the frequency over the entire sequence from the location of the first residue of a given group to the location of the residue occurring at a given frequency.

Feature extraction methods for RNA

Feature extraction based on RNA sequence

The principle of nucleic acid composition (NAC) is to calculate the frequency of each type of nucleic acid in the sequence. Dinucleotide composition (DNC) and trinucleotide composition (TNC) count dinucleotides and trinucleotides, whose dimensions are 16 and 64, respectively.

Enhanced NAC (ENAC), a variant of NAC, calculates NAC on the basis of a sequence window of fixed length (sequence slides continuously from the 5′ end to the 3′ end). Binary is a binary code for each nucleotide (A = 1000, C = 0100, G = 0010, T[U] = 0001), which is usually used to encode nucleotide sequences of equal length. The composition of k-spaced nucleic acid pairs (CKSNAP) calculates the frequency of nucleic acid pairs isolated from any nucleic acid. The principle of accumulated nucleotide frequency (ANF) is to calculate the nucleotide frequency information and the density of any nucleotide in the RNA sequence. Electron-ion interaction pseudopotentials of trinucleotide (EIIP) computes the electron-ion interaction pseudopotentials of trinucleotides in RNA sequences. Electron-ion interaction pseudopotentials of trinucleotide (PseEIIP) use the EIIP average of trinucleotides in each sample to construct the feature vector.

Feature extraction method based on the physical and chemical properties of RNA

Nucleotide chemical property (NCP) is determined according to the chemical structure and chemical properties of the nucleotides encoding (A = [1,1,1], C = [0, 0], G = [0, 1], U = [0, 1]). Dinucleotide-based autocovariance (DAC) and trinucleotide-based autocovariance (TAC) are calculated as the correlation of the same physicochemical indices between dinucleotides or trinucleotides separated by subsequent nucleic acids along the sequence. Dinucleotide-based cross-covariance (DCC), and trinucleotide-based cross-covariance (TCC) calculate the correlation of the same physicochemical index between two dinucleotides (trinucleotides) separated by a lag distance along the sequence. DAC and DCC are combined to generate the dinucleotide-based autocross covariance (DACC) code. Accordingly, TAC and TCC combine to form trinucleotide-based auto-Cross-covariance (TACC). Pseudo-NAC (PseNAC) is obtained on the basis of local sequence order information and remote sequence effects, whose derived feature extraction methods include pseudo-k-tuple nucleotide composition (PseKNC), parallel correlation pseudo-DNC (PC-PseDNC), parallel correlation pseudo-TNC (PC-PseTNC), series correlation pseudo-DNC (SC-PseDNC), and series correlation pseudo-TNC (SC-PseTNC). PseDNC is obtained by combining continuous local sequence order information and global sequence order information. PC-PseDNC is a variant of PseDNC that uses 38 default physicochemical indices, and PseDNC uses 6 default physicochemical indices.[72,73]

Feature selection

Feature selection is the process of selecting a subset of available features that are useful to the model (the useless features are eliminated). Feature selection can improve the performance of the model and help researchers understand the characteristics and underlying structure of the data, which plays an important role in further improving the model and algorithm.

The principle of the filtering method is to calculate the information S(i) of each feature Xi relative to the label y and obtain n results. Then, sort the n S(i) in descending order and output the top k features. Generally, the Pearson correlation coefficient, chi-square test, mutual information, and maximum information coefficient, distance correlation coefficient and variance selection method are used to measure S(i). Minimum redundancy maximum correlation is a filtering feature selection method whose goal is to maximize the correlation between features and categorical variables. MRMR considers not only the correlation between features and labels but also the correlation between features. The wrapping method is based on the hold-out method. Specifically, for each feature subset to be selected, the model is trained on the training set, and then the feature subset is selected on the test set according to the error. Greedy search, which is locally optimal, needs to be combined. Greedy algorithms can reduce the computational complexity.[74]

The embedding method is a feature selection method based on the penalty term, which selects features through the L1 regularization term. The L1 regularization method has the characteristic of a sparse solution, so it can be used for feature selection. The absence of features in L1 does not mean the features are unimportant. Because only one of two highly correlated features may be retained, the L2 regularization method can be used to determine which feature is crucial. MRMD2.0 (Max-Relevance-Max-Distance) achieves a balance between feature ordering, prediction accuracy, and stability by sorting high-dimensional features to get rid of important information. Compared with other feature selection algorithms, the biggest advantage of MRMD2.0 is stability, which can ensure that the feature vector with reduced dimensions can obtain excellent performance.[75]

The low variance feature extraction method drops the features with very low variance. In the multicollinearity feature extraction method, when there is correlation between any two features, multicollinearity will occur. Traditional machine learning expects that each feature should be independent of the others, which means low collinearity. The objective is to select features with low collinearity.

Feature selection on the basis of feature importance is also an important strategy. Decision tree, RF, LGBM, and XGBoost split the data using a feature that minimizes impurities. Finding the optimal features is a pivotal part of classification. The goal of PCA is to decrease the dimensionality of the high-dimensional feature space. The original features are reprojected into new dimensions (principal components). The goal is to find the number of features that explain the variance in the data.[76]

Traditional machine learning algorithm

Traditional machine learning is the study of a particular algorithm (rather than a specific algorithm) that allows a computer to learn in a training set to predict the new samples. Supervised learning algorithms are dominant in protein SCL and RNA SCL. On the basis of the understanding and summary of the protein SCL model and RNA SCL model, SVM is recognized as the commonly used traditional machine learning algorithm.

SVM is a classification algorithm. SVM presents sample features as points in space and uses functions to segment sample points. SVM is only suitable for binary classification problems. In the multiclassification problem, the multiclassification problem needs to be transformed into multiple binary classification problems to use SVM.[77] In addition, in the protein SCL and RNA SCL problems, the tree model is one of the methods. The tree model includes RF,[78] boosting,[79] XGBoost,[80] and LGBM.[81] The tree model is the internal node of tree mode signal generated by the learning of the dataset. The internal node represents the judgment of an attribute. Information entropy is used to measure uncertainty. The stronger the uncertainty, the greater the information entropy. The weaker the uncertainty, the lower the information entropy. In the feature selection of the tree model, the feature with maximum information gain is selected. After building the first layer of the tree for the feature with the greatest gain, recursion is continued to find the feature with the greatest gain in addition to the previous feature with the greatest gain. By that analog, until all the leaf nodes are output, the sample can be traversed.

In traditional machine learning, appropriate algorithms need to be selected according to the training set. In addition, ensemble learning is often used to construct models. Ensemble learning generally consists of two stages. The first stage uses a separate method whose aim is to obtain a base classifier by training samples. In the second stage, the voting method is used for ensemble learning (combining different base classifiers). When the difference between base classifiers is large, the effect of the integration model is more obvious.[82]

Model evaluation

The evaluation index is the standard to measure the performance of the model. Reasonable use of evaluation indicators is the basis of training excellent performance model. ACC (Equation 3) is a common indicator of evaluation, showing the accuracy of model prediction. However, ACC cannot evaluate the performance of models on unbalanced test sets. At this time, more comprehensive evaluation indicators were introduced, including F-score[83] and Matthews' correlation coefficient (MCC) (Equation 5). F-score (Equation 4) is a comprehensive evaluation indicator of precision (Equation 1) and recall (Equation 2). MCC[84] is designed to measure the correlation between actual and predicted labels, and it is a widely accepted comprehensive performance evaluation indicator. FP means that the actual sample is negative, but the predicted sample is positive. FN represents that the actual sample is negative, and the actual sample is also negative. TN means that the actual sample is positive, but the predicted sample is negative.[85]

$$precision = \frac{TP}{TP + FN} \qquad \text{(Equation 1)}$$

$$recall = \frac{TN}{TN + FP} \qquad \text{(Equation 2)}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \qquad \text{(Equation 3)}$$

$$F - score = \frac{2TP}{2TP + FP + FN} \qquad \text{(Equation 4)}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \qquad \text{(Equation 5)}$$

RNA SCL modeling based on deep learning

Protein SCL models based on deep learning

To study the SCL of proteins from multiple aspects, SCL models based on deep learning have been widely discussed (Table 3; Figure 3).

The image-based method can be used to analyze the spatial distribution and location changes of proteins in normal and cancerous tissues. Newberg and Murphy proposed a framework for analyzing protein spatial distribution, in which SCL features were used to identify protein subcellular patterns in images. NNPSL is a protein SCL model based on an artificial neural network (ANN). NNPSL constructs two kinds of neural networks for the SCL of prokaryotic and eukaryotic protein sequences: one network with a simple full-link layer and the other with a hidden layer. Among these networks, the subcellular sites of prokaryotes include not only the cytoplasm, extracellular space, mitochondria, and nucleus of eukaryotes but also the cytoplasm, extracellular space, and periplasmic space of prokaryotes.[106] The dataset proposed by Hubbard uses a self-organizing model to construct the SCL model (the self-organizing model is a typical neural network).[86] TargetP developed a neural network-based model using N-terminal amino acid sequences, whose target was to distinguish between mitochondria and chloroplasts.[87] Given the inability of neural networks to accurately interpret features within the model, IPSORT was developed to achieve the accuracy of TargetP using the amino acid index, and alphabet indexing concatenated the approximate pattern.[63] Bodén and Hawkins[88] used the targetP dataset to train the biased cyclic network model based on sequence. PredSL uses neural networks, Markov chains, contour-hidden Markov models and scoring matrices to construct protein SCL models based on N-terminal sequences, with subcellular locations including chloroplasts, thylakoids, mitochondria, secretory pathways, and others.[88] DeepYeast is trained to study protein SCL problems, considering the limitations of protein sequences. A convolutional neural network (CNN) is used for model training.[89] Kumar et al.[107] suggested that cancer markers could be identified

Table 3. Models of protein SCL based on deep learning

Model/author	Algorithm	Time	Reference
Models of protein SCL			
Hubbard	NN	2000	Cai et al.[86]
TargetP	NN	2002	Emanuelsson et al.[87]
PredSL	cyclic network	2005	Bodén et al.[88]
DeepYeast	CNN	2017	Pärnamaa and Parts[89]
MIC_Locator	CNN	2019	Yang et al.[90]
ImPLoc	CNN	2020	Long et al.[91]
zhang	CNN	2022	Zhang et al.[92]
Cytoself	self-supervised deep learning encodes	2022	Kobayashi et al.[93]
Aggarwal et al.	artificial intelligence stack ensemble approach	2023	Aggarwal et al.[94]
PScL-2LSAESM	two-level stacked autoencoder network	2023	Ullah et al.[95]
Ding et al.	multi-scale and multi-model deep neural network through an ensemble strategy	2023	Ding et al.[96]
Models of RNA SCL			
DeepLncRNA	NN	2018	Gudenas and Wang[97]
RNATracker	CNN	2019	Yan et al.[98]
LncRRIsearch	NN	2019	Fukunaga et al.[99]
LncLocator2.0	NLP	2021	Lin et al.[100]
DM3Loc	CNN	2021	Wang et al.[101]
DeepLncLoc	text CNN	2022	Zeng et al.[102]
GraphLncLoc	GCN	2023	Li et al.[103]
DeepmRNALoc	DNN	2023	Wang et al.[104]
GM-LncLoc	GCN	2023	Cai et al.[105]

by measuring changes in protein expression levels and SCL between normal tissues and cancerous tissues.

MIC_Locator is a protein SCL model that converts immunohistochemical images into the frequency domain to capture local features and achieves better performance on a multilabel classification model.[90] Zhang et al.[92] trained a protein SCL model on the basis of microscopic images. IF images were obtained from the HPA database, and different sizes were contained. The selective line search algorithm was used to crop the original image and obtain an image with a uniform size of 512 × 512. DenseNet was applied as the backbone network, which is an extension of ResNet. Meanwhile, multitask learning strategies and generating image masks were applied to improve the performance. In addition, the experiment found that the feature fusion method based on a CNN has better performance than the limit-based model. ImPLoc is a deep learning model for protein SCL based on immunohistochemical (IHC) images. The tissue map of the Human Protein Atlas (HPA) contains the IHC image, which allows the distribution of proteins at both the tissue and cellular levels to be visualized. The feature extraction method of deep CNN extracts the features of the image. The autoattention encoder was used to fuse multiple features to construct the model.[91] Cytoself is an image-based protein SCL model, which use self-supervised deep learning encodings. Cytoself is a rare model that applies self-supervised learning in SCL.[93] Aggarwal et al.[94] developed a model for targeting protein SCL in microscopy images on the basis of an artificial intelligence stack ensemble approach (by using 231,072 samples). Each sample consists of four images. Aggarwal et al.[94] use three of the four images to train the model, and the model involves 28 subcellular locations. PScL-2LSAESM is a model for protein SCL based on biological images, which is a novel two-level stacked autoencoder network (2L-SAE-SM) that aims to improve performance by integrating heterogeneous features. In the 2L-SAE-SM first level, each optimal heterogeneous feature set is incorporated to train the model. The outputs of the first level are defined as intermediate decision sets, which are averagely integrated to generate the second-level SAE-SM.[95] Ding et al.[96] proposed a multi-scale and multi-model deep neural network through an ensemble strategy. The model uses protein SCL on single-cell high-throughput images to train the model, which provides a foundation for the study of protein and gene functions.

From the perspective of protein SCL models based on deep learning, the CNN model is widely accepted. CNN is an effective approach for graph classification, which mainly includes convolution, pooling layer, and full connection layer. The main objective of the convolution layer is to retain the features of the image. The main function of the pooling layer is to limit reduction of the data, which can effectively avoid overfitting. Fully connected layer is used to output the classification results. The convolution layer can be used separately for feature extraction, which can effectively combine deep learning and traditional machine learning. In addition, self-supervised learning has also been introduced to the SCL, which may present new opportunities for SCL.

RNA SCL models based on deep learning

Considering the size of the RNA SCL training dataset, there are relatively few RNA SCL models based on deep learning so far (Table 3; Figure 3). Coincidentally, the current mainstream RNA SCL models based on deep learning focus on lncRNAs rather than mRNAs. The reason for this coincidence may be the limitation of the data.

DeepLncRNA[97] is a feedforward multilayer neural network model for SCL based on lncRNAs. The size of the training set was 93 SCL samples of human RNA, 48 belonging to the nucleus and 45 belonging to the cytoplasm. K-mer is used for feature extraction. k is set in the range of 2–5 to calculate the nucleic acid frequency. DeepLncRNA uses three rectified linear unit activation functions as the three hiding layers and softmax as the output layer. For overfitting problems, DeepLncRNA reduces the impact by randomly partitioning the connecting half of each layer. In addition, DeepLncRNA makes the dropout randomly block some hidden units in each layer, which improves the generalization ability of the model. The backpropagation algorithm is used to adjust the weights of the DeepLncRNA networks. A new deep neural network, known as RNATracker,[98] was used to predict the distribution of RNA transcripts in the subcellular compartment. The model uses CNN technology to construct the network, utilizing sequence information and secondary structure to train the model. Human and mouse lncRNA-lncRNA and lncRNA-RNA interactions are applied to train LncRRIsearch,[99] which can be used to predict synthetic interaction, tissue specific, and subcellular interaction patterns.

LncLocator2.0[100] is a CNN model targeting lncRNAs, whose data type is RNA sequence. LncLocator2.0 uses CNRCI values (the logarithmic ratio of the concentration of two samples) and sequence length (the sequence length of the samples is required to be less than 1,000) to filter the data. First, natural language processing (NLP) was used to analyze RNA sequences, and the results were connected to a CNN to predict location. The classification result contains the CNRCI value for each location, which can be considered to confirm the sublocation. DM3Loc is an mRNA SCL model based on deep learning, which is encoded by one-hot, where T and U share the same coding. Sequences are filled or truncated to fix the sequence length at 8,000 nt. A CNN is applied to train the model. At the same time, different attention weights were set to optimize the model.[101]

Like the protein SCL models based on deep learning, the effective algorithm of deep learning for RNA SCL model is also a CNN. Notably, NLP has also been used for SCL. Different from CNN, the train data type of NLP is sequence, not image. Notably, given the limitations of k-mer, DeepLncLoc combines a new subsequence embedding method to extract features and uses text CNN for model construction, which takes the advantage of opportunities for SCL.[102] GraphLncLoc is a model based on graph CNN, which uses sequence to graph transformation to train the model. The research object is lncRNA SCL data. Considering that k-mer can lose sequence information and ignore sequence patterns and motifs without length, Li et al.[103] transformed the sequence classification into a graph classification to extract high-level features and used graph CNNs to train the model. DeepmRNALoc is a deep neural network model based on eukaryotic mRNA, which includes five subcellular locations (cytoplasm, endoplasmic reticulum, extracellular region, mitochondria, and nucleus).[104] Given that previous models are based on low-level information and learn from a small data size, GM-lncLnc extracted the advanced features of lncRNA on the basis of the sequence information and combined with the graph structure. The introduction of meta-learning accelerates the progress of learning, which to some extent solves the problem of data size.[105]

To date, deep learning studies on RNA SCL have been limited by the size of RNA datasets. Over time, more RNA SCL entries will accumulate, and RNA SCL models based on deep learning with better performance can be expected. More data and more reasonable classification algorithms are expected to train better protein SCL and RNA SCL models by using deep learning. Traditional machine learning is more interpretable than deep learning. Traditional machine learning can help discover and track biological pathways, but deep learning may be more expressive. The combination of deep learning and traditional machine learning could also be an interesting direction. In addition, the effectiveness and ineffectiveness of the model require consideration. Researchers in computational biology may only focus on the performance of the model, whose goal is to discover the better performing models. Biologists may focus on other aspects, who may be interested in the inside of the model. The contrast between model effectiveness and ineffectiveness is also important to biologists. In other words, biologists prefer to know why models work in some cases but in other cases do not. For traditional machine learning, this process is visible and explainable. But the deep learning model will lack explanatory power.

In the protein SCL model, Shen et al. combined multiple features (including PsePSSM, DWT, AvBlock) to extract comprehensive information. It can comprehensively consider various features/information to increase the possibility of accurate predicting unknown samples. The model trained by Liu et al. based on multiple species, which is recommended for readers whose data covers multiple species. TACOS fuses 10 feature descriptors for lncRNA SCL, which is the first model based on tree stacking. SubLocEP extracts mRNA sequence information from multiple levels, including sequence information and physicochemical properties features. Integrated information helps TACOS and SubLocEP perform stably when making prediction on unknown samples. RNAlight has been extended to other types of RNA SCL, which is unique advantage. The machine learning models are usually trained using traditional algorithm, including SVM/RF/KNN/GKR. The models performance of SVM/RF are relatively stable on the small size training set. Therefore, Liu model (for protein), LncRNA (for lncRNA), and SubLocEP (mRNA) are recommended. RNAlight is recommended for those who need to study other types of RNA SCL.

Zhang et al. used HPA data to train the model, multitask learning, generate image masks, and feature fusion and CNN were combined to improve the performance of the model. Combining sequence features and graph structure help GM-lncLnc capture more information to improve the performance. DM3Loc is a CNN model for mRNA. One-hot encoding and weight setting help stabilize DM3Loc. Among SCL models based on deep learning, CNNs are widely considered to be expressive. This is confirmed by the performance of both Zhang model and DM3Loc. Aggarwal is competitive for readers who cover multiple subcellular location (28 SCLs). Compared with analyzing sequences or images, GM-lncLnc combining sequence information and graph structure is expected to have excellent performance.

GENERAL PROBLEMS AND FUTURE DIRECTIONS

Data description

The data in biology are complex and heavy. There are a lot of public data in different forms and species. In papers on SCL models, the description of datasets is often inadequate. In fact, data are the most important part of traditional machine learning. A clear description of the data is essential for the reader. When the amount of data is small, biocomputing researchers tend to use traditional machine learning because traditional machine learning methods are more likely to perform reliably. As more samples are discovered, researchers prefer to use deep learning.

The selection of learning methods

While reading SCL articles, we found that the authors rarely mentioned why the approach was chosen for traditional machine learning or deep learning. At the end of the articles, there is no explanation about why the model works well (or better than other mainstream tools). The researchers only use indicators to show the performance of model without in-depth analysis, which will hinder understanding of the paper. Biocomputing researchers should select the reasonable algorithm to train model, not the popular algorithm (or for any other reason).

The construction of models

When proving the high performance of models, researchers often use their own test sets. There is no baseline data to help them prove the performance of the model. Even if there is a public website with a benchmark, researchers will adjust the model to fit the benchmark. So far, there is no accurate way for researchers to verify the ultimate performance of the models. In addition, it is very difficult to reproduce a model that has been published. Some models fail to get the same results, which is a fatal blow to the validity of the study. In addition, data leakage during the training of the model is also very deadly. When computer researchers write code, some researchers are not comfortable with the complete separation of the training set and the test set. On the surface, the code looks reasonable, but in fact there is a data leak problem. For computer researchers, this is a fatal mistake.

The sharing of code and data

In the process of sorting out the SCL databases and models, we found that the online access tools of the databases and models in many published papers had been stopped, which affected the progress of other researchers and hindered the development of the SCL. It is important to share resources on more authoritative and stable servers. For instance, researchers can upload code to GitHub (https://github.com) and data resources to the more spacious Google Drive (https://accounts.google.com).

Future directions

Currently algorithms from other fields such as computer vision and NLP have been applied directly to biomedical problems, but there is a lack of biomedical-specific algorithms. In other words, researchers could develop algorithms on the basis of biology in the future.

To explore the relationship between protein SCL and RNA SCL, interdisciplinary collaboration between traditional and computational biology is vital. Through this collaboration, researchers can develop more effective algorithms and models to gain new insights into complex biological processes. Such a partnership is essential for advancing our understanding of these fundamental molecular components and ultimately improving human health.

CONCLUSION

Compared with traditional electron microscopy, fluorescence microscopy, and other traditional methods for SCL, the computational method can shorten the research time, save manpower and material resources, and reduce the possibility of experimental error. This paper reviews and discusses the current mainstream protein SCL and RNA SCL databases and prediction models, hoping to help readers understand the development process of SCL problems and the development of computer technology in biology. It is also expected to help biological researchers understand computer technology to further promote the development of SCL problems. The transcriptional dependence between RNA and protein may be an important research direction on the relationship between RNA SCL and protein SCL. The previous papers have treated protein SCL and RNA SCL as separate problems; this review is the first to cover both protein SCL and RNA SCL. In computational biology, different models use different benchmark datasets to evaluate the performance of the models. Strictly speaking, it is very difficult to compare different models without authoritative benchmark data. So far, the analysis of model strengths has been based on computer technology. In fact, authoritative benchmark data should be established, and the performance of the mentioned models on specific benchmark dataset should be presented. The difficulty collecting authoritative datasets and conducting rigorous tests on the proposed models are the limitations of this review.

In the future, we will collect the latest published protein and RNA SCL data and evaluate the models we mentioned. This will help the reader intuitively understand the advantages of the models. To the best of our knowledge, the research literature does not involve research on the relationship between protein SCL and RNA SCL. Relevant papers only discuss protein SCL and RNA SCL as independent issues. Given the predictive value to human disease, importance of finding target drugs, understanding viral mechanisms, and the significance of discovering antiviral drug clues, we call for multidisciplinary efforts between biology and computational biology fields to work together to solve the problem of protein and RNA SCL.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.omtn.2023.04.015.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (grants 62102063 and 62131004) and the Municipal Government of Quzhou (grant 2022D040).

AUTHOR CONTRIBUTIONS

Q.Z. and L.Y. proposed the idea of a SCL review and pointed out the limitations of the existing model. L.J. sorted the literature, analyzed the state-of-the-art models, and summarized the general machine learning methods for SCL. Q.Z., L.Y., and L.J. jointly determined future research directions for SCL.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Aridor, M., and Hannan, L.A. (2000). Traffic jam: a compendium of human diseases that affect intracellular transport processes. Traffic *1*, 836–851.
2. Shen, H.-B., Yang, J., and Chou, K.-C. (2007). Methodology development for predicting subcellular localization and other attributes of proteins. Expert Rev. Proteomics *4*, 453–463.
3. Liu, F., and Gong, C.-X. (2008). Tau exon 10 alternative splicing and tauopathies. Mol. Neurodegener. *3*, 8.
4. Cooper, T.A., Wan, L., and Dreyfuss, G. (2009). RNA and disease. Cell *136*, 777–793.
5. Wu, Y.-C., Chen, C.-S., and Chan, Y.-J. (2020). The outbreak of COVID-19: an overview. J. Chin. Med. Assoc. *83*, 217–220.
6. Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., et al. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet *395*, 565–574.
7. Lau, Y.L., and Peiris, J.S.M. (2005). Pathogenesis of severe acute respiratory syndrome. Curr. Opin. Immunol. *17*, 404–410.
8. Zhang, H., Penninger, J.M., Li, Y., Zhong, N., and Slutsky, A.S. (2020). Angiotensin-converting enzyme 2 (ACE2) as a SARS-CoV-2 receptor: molecular mechanisms and potential therapeutic target. Intensive Care Med. *46*, 586–590.
9. Zumla, A., Chan, J.F.W., Azhar, E.I., Hui, D.S.C., and Yuen, K.-Y. (2016). Coronaviruses—drug discovery and therapeutic options. Nat. Rev. Drug Discov. *15*, 327–347.
10. Naqvi, A.A.T., Fatima, K., Mohammad, T., Fatima, U., Singh, I.K., Singh, A., Atif, S.M., Hariprasad, G., Hasan, G.M., and Hassan, M.I. (2020). Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: structural genomics approach. Biochim. Biophys. Acta, Mol. Basis Dis. *1866*, 165878.
11. Bashirullah, A., Cooperstock, R.L., and Lipshitz, H.D. (1998). RNA localization in development. Annu. Rev. Biochem. *67*, 335–394.
12. UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. *47*, D506–D515.
13. Fink, J.L., Aturaliya, R.N., Davis, M.J., Zhang, F., Hanson, K., Teasdale, M.S., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., and Teasdale, R.D. (2006). LOCATE: a mouse protein subcellular localization database. Nucleic Acids Res. *34*, D213–D217.
14. Sprenger, J., Lynn Fink, J., Karunaratne, S., Hanson, K., Hamilton, N.A., and Teasdale, R.D. (2008). LOCATE: a mammalian protein subcellular localization database. Nucleic Acids Res. *36*, D230–D233.
15. Pierleoni, A., Martelli, P.L., Fariselli, P., and Casadio, R. (2007). eSLDB: eukaryotic subcellular localization database. Nucleic Acids Res. *35*, D208–D212.
16. Rey, S., Acab, M., Gardy, J.L., Laird, M.R., DeFays, K., Lambert, C., and Brinkman, F.S.L. (2005). PSORTdb: a protein subcellular localization database for bacteria. Nucleic Acids Res. *33*, D164–D168.
17. Yu, N.Y., Laird, M.R., Spencer, C., and Brinkman, F.S.L. (2011). PSORTdb—an expanded, auto-updated, user-friendly protein subcellular localization database for Bacteria and Archaea. Nucleic Acids Res. *39*, D241–D244.
18. Peabody, M.A., Laird, M.R., Vlasschaert, C., Lo, R., and Brinkman, F.S.L. (2016). PSORTdb: expanding the bacteria and archaea protein subcellular localization database to better reflect diversity in cell envelope structures. Nucleic Acids Res. *44*, D663–D668.
19. Gardy, J.L., Spencer, C., Wang, K., Ester, M., Tusnády, G.E., Simon, I., Hua, S., DeFays, K., Lambert, C., and Nakai, K. (2003). PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria. Nucleic Acids Res. *31*, 3613–3617.
20. Amaral, P.P., Clark, M.B., Gascoigne, D.K., Dinger, M.E., and Mattick, J.S. (2011). lncRNAdb: a reference database for long noncoding RNAs. Nucleic Acids Res. *39*, D146–D151.
21. Heinzelmann, J., Henning, B., Sanjmyatav, J., Posorski, N., Steiner, T., Wunderlich, H., Gajda, M.R., and Junker, K. (2011). Specific miRNA signatures are associated with metastasis and poor prognosis in clear cell renal cell carcinoma. World J. Urol. *29*, 367–373.
22. Quek, X.C., Thomson, D.W., Maag, J.L.V., Bartonicek, N., Signal, B., Clark, M.B., Gloss, B.S., and Dinger, M.E. (2015). lncRNAdb v2. 0: expanding the reference database for functional long noncoding RNAs. Nucleic Acids Res. *43*, D168–D173.
23. Mas-Ponte, D., Carlevaro-Fita, J., Palumbo, E., Hermoso Pulido, T., Guigo, R., and Johnson, R. (2017). LncATLAS database for subcellular localization of long noncoding RNAs. RNA *23*, 1080–1087.
24. Wen, X., Gao, L., Guo, X., Li, X., Huang, X., Wang, Y., Xu, H., He, R., Jia, C., and Liang, F. (2018). lncSLdb: a resource for long non-coding RNA subcellular localization. Database, 2018.
25. Liu, T., Zhang, Q., Zhang, J., Li, C., Miao, Y.-R., Lei, Q., Li, Q., and Guo, A.-Y. (2019). EVmiRNA: a database of miRNA profiling in extracellular vesicles. Nucleic Acids Res. *47*, D89–D93.
26. Zhang, T., Tan, P., Wang, L., Jin, N., Li, Y., Zhang, L., Yang, H., Hu, Z., Zhang, L., Hu, C., et al. (2017). RNALocate: a resource for RNA subcellular localizations. Nucleic Acids Res. *45*, D135–D138.
27. Cui, T., Dou, Y., Tan, P., Ni, Z., Liu, T., Wang, D., Huang, Y., Cai, K., Zhao, X., Xu, D., et al. (2022). RNALocate v2. 0: an updated resource for RNA subcellular localization with increased coverage and annotation. Nucleic Acids Res. *50*, D333–D339.
28. Wang, Z., Zou, Q., Jiang, Y., Ju, Y., and Zeng, X. (2014). Review of protein subcellular localization prediction. Curr. Bioinform. *9*, 331–342.
29. Shen, Y., Ding, Y., Tang, J., Zou, Q., and Guo, F. (2020). Critical evaluation of web-based prediction tools for human protein subcellular localization. Brief. Bioinform. *21*, 1628–1640.
30. Gardy, J.L., and Brinkman, F.S.L. (2006). Methods for predicting bacterial protein subcellular localization. Nat. Rev. Microbiol. *4*, 741–751.
31. Dönnes, P., and Höglund, A. (2004). Predicting protein subcellular localization: past, present, and future. Dev. Reprod. Biol. *2*, 209–215.
32. Kumar, R., and Dhanda, S.K. (2020). Bird eye view of protein subcellular localization prediction. Life *10*, 347.
33. Feng, Z.-P. (2002). An overview on predicting the subcellular location of a protein. In Silico Biol. *2*, 291–303.
34. Bannasch, D., Mehrle, A., Glatting, K.H., Pepperkok, R., Poustka, A., and Wiemann, S. (2004). LIFEdb: a database for functional genomics experiments integrating information from external sources, and serving as a sample tracking system. Nucleic Acids Res. *32*, D505–D508.
35. Eppig, J.T., Bello, S.M., Kadin, J.A., Bult, C.J., Blake, J.A., Richardson, J.E., Baldarelli, R.M., Anagnostopoulos, A., Beal, J.S., Baya, M., et al. (2005). The Mouse Genome Database (MGD): from genes to mice—a community resource for mouse biology. Nucleic Acids Res. *33*, D471–D475.
36. Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., et al. (2005). The universal protein resource (UniProt). Nucleic Acids Res. *33*, D154–D159.
37. Hubbard, T.J.P., Aken, B.L., Beal, K., Ballester, B., Cáccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., et al. (2007). Ensembl 2007. Nucleic Acids Res. *35*, D610–D617.
38. Lagos-Quintana, M., Rauhut, R., Yalcin, A., Meyer, J., Lendeckel, W., and Tuschl, T. (2002). Identification of tissue-specific microRNAs from mouse. Curr. Biol. *12*, 735–739.
39. Chen, A.-J., Paik, J.-H., Zhang, H., Shukla, S.A., Mortensen, R., Hu, J., Ying, H., Hu, B., Hurt, J., Farny, N., et al. (2012). STAR RNA-binding protein Quaking suppresses cancer via stabilization of specific miRNA. Genes Dev. *26*, 1459–1472.
40. Cheng, X., Zhao, S.-G., Lin, W.-Z., Xiao, X., and Chou, K.-C. (2017). pLoc-mAnimal: predict subcellular localization of animal proteins with both single and multiple sites. Bioinformatics *33*, 3524–3531.
41. Ding, Y., Tang, J., and Guo, F. (2020). Human protein subcellular localization identification via fuzzy model on kernelized neighborhood representation. Appl. Soft Comput. *96*, 106596.

42. Cheng, X., Lin, W.-Z., Xiao, X., and Chou, K.-C. (2019). pLoc_bal-mAnimal: predict subcellular localization of animal proteins by balancing training dataset and PseAAC. Bioinformatics *35*, 398–406.

43. Xiao, X., Cheng, X., Chen, G., Mao, Q., and Chou, K.-C. (2019). pLoc_bal-mGpos: predict subcellular localization of Gram-positive bacterial proteins by quasi-balancing training dataset and PseAAC. Genomics *111*, 886–892.

44. Chou, K.-C., Cheng, X., and Xiao, X. (2019). pLoc_bal-mHum: predict subcellular localization of human proteins by PseAAC and quasi-balancing training dataset. Genomics *111*, 1274–1282.

45. Xiao, X., Cheng, X., Chen, G., Mao, Q., and Chou, K.-C. (2019). pLoc_bal-mVirus: predict subcellular localization of multi-label virus proteins by Chou's general PseAAC and IHTS treatment to balance training dataset. Med. Chem. *15*, 496–509.

46. Liu, H., Zhang, Y., Lu, S., Chen, H., Wu, J., Zhu, X., Zou, B., and Hua, J. (2021). Identifying protein subcellular location with embedding features learned from networks. New Phytol. *231*, 646–660.

47. Pan, X., Li, H., Zeng, T., Li, Z., Chen, L., Huang, T., and Cai, Y.-D. (2020). Identification of protein subcellular localization with network and functional embeddings. Front. Genet. *11*, 626500.

48. Sahu, S.S., Loaiza, C.D., and Kaundal, R. (2020). Plant-mSubP: a computational framework for the prediction of single-and multi-target protein subcellular localization using integrated machine-learning approaches. AoB Plants *12*, plz068.

49. Cao, Z., Pan, X., Yang, Y., Huang, Y., and Shen, H.-B. (2018). The lncLocator: a subcellular localization predictor for long non-coding RNAs based on a stacked ensemble classifier. Bioinformatics *34*, 2185–2194.

50. Su, Z.-D., Huang, Y., Zhang, Z.-Y., Zhao, Y.-W., Wang, D., Chen, W., Chou, K.-C., and Lin, H. (2018). iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. Bioinformatics *34*, 4196–4204.

51. Ahmad, A., Lin, H., Shatabda, S., and Locate, -R. (2020). Subcellular localization of long non-coding RNAs using nucleotide compositions. Genomics *112*, 2583–2589.

52. Tang, Q., Nie, F., Kang, J., and Chen, W. (2021). mRNALocater: enhance the prediction accuracy of eukaryotic mRNA subcellular localization by using model fusion strategy. Mol. Ther. *29*, 2617–2623.

53. Li, J., Zhang, L., He, S., Guo, F., Zou, Q., and SubLocEP. (2021). A novel ensemble predictor of subcellular localization of eukaryotic mRNA based on machine learning. Brief. Bioinform. *22*, bbaa401.

54. Zhang, Z.-Y., Sun, Z.-J., Yang, Y.-H., and Lin, H. (2022). Towards a better prediction of subcellular location of long non-coding RNA. Front. Comput. Sci. *16*, 165903.

55. Jeon, Y.-J., Hasan, M.M., Park, H.W., Lee, K.W., and Manavalan, B. (2022). TACOS: a novel approach for accurate prediction of cell-specific long noncoding RNAs subcellular localization. Brief. Bioinform. *23*, bbac243.

56. Yuan, G.-H., Wang, Y., Wang, G.-Z., and Yang, L. (2023). RNAlight: a machine learning model to identify nucleotide features determining RNA subcellular localization. Brief. Bioinform. *24*, bbac509.

57. Choudhury, S., Bajiya, N., Patiyal, S., and Raghava, G.P. (2023). A hybrid approach for predicting multi-label subcellular localization of mRNA at genome scale. Preprint at bioRxiv. https://doi.org/10.1101/2023.01.17.524365.

58. Nakashima, H., and Nishikawa, K. (1994). Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. J. Mol. Biol. *238*, 54–61.

59. Hua, S., and Sun, Z. (2001). Support vector machine approach for protein subcellular localization prediction. Bioinformatics *17*, 721–728.

60. Chou, K.-C., Cheng, X., and Xiao, X. (2019). pLoc_bal-mEuk: predict subcellular localization of eukaryotic proteins by general PseAAC and quasi-balancing training dataset. Med. Chem. *15*, 472–485.

61. Nakai, K., and Kanehisa, M. (1992). A knowledge base for predicting protein localization sites in eukaryotic cells. Genomics *14*, 897–911.

62. Horton, P., and Nakai, K. (1997). Better prediction of protein cellular localization sites with the it k nearest neighbors classifier. Proc. Int. Conf. Intell. Syst. Mol. Bio. *5*, 147–152.

63. Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., and Miyano, S. (2002). Extensive feature detection of N-terminal protein sorting signals. Bioinformatics *18*, 298–305.

64. Drawid, A., and Gerstein, M. (2000). A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. J. Mol. Biol. *301*, 1059–1075.

65. Mitchell, T., Buchanan, B., DeJong, G., Dietterich, T., Rosenbloom, P., and Waibel, A. (1990). Machine learning. Annu. Rev. Comput. Sci. *4*, 417–433.

66. Gnip, P., and Drotár, P. (2019). Ensemble methods for strongly imbalanced data: bankruptcy prediction. In 2019 IEEE 17th International Symposium on Intelligent Systems and Informatics (SISY) (IEEE).

67. Reid, R.L. (1986). The psychology of the near miss. J. Gambling Stud. *2*, 32–39.

68. Lin, W.-C., Tsai, C.-F., Hu, Y.-H., and Jhang, J.-S. (2017). Clustering-based undersampling in class-imbalanced data. Inf. Sci. *409-410*, 17–26.

69. Cox, M.G., and Siebert, B.R.L. (2006). The use of a Monte Carlo method for evaluating uncertainty and expanded uncertainty. Metrologia *43*, S178–S188.

70. Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. (2002). SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. *16*, 321–357.

71. He, H., Bai, Y., Garcia, E.A., and Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence) (IEEE).

72. Chen, Z., Zhao, P., Li, F., Marquez-Lago, T.T., Leier, A., Revote, J., Zhu, Y., Powell, D.R., Akutsu, T., Webb, G.I., et al. (2020). iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. Brief. Bioinform. *21*, 1047–1057.

73. Chen, Z., Zhao, P., Li, F., Leier, A., Marquez-Lago, T.T., Wang, Y., Webb, G.I., Smith, A.I., Daly, R.J., and Chou, K.-C. (2018). iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics *34*, 2499–2502.

74. Meyer, P.E., Kontos, K., Lafitte, F., and Bontempi, G. (2007). Information-theoretic inference of large transcriptional regulatory networks. EURASIP J. Bioinform. Syst. Biol. *2007*, 79879.

75. He, S., Guo, F., and Zou, Q. (2020). MRMD2. 0: a python tool for machine learning with feature ranking and reduction. Curr. Bioinform. *15*, 1213–1221.

76. Daffertshofer, A., Lamoth, C.J.C., Meijer, O.G., and Beek, P.J. (2004). PCA in studying coordination and variability: a tutorial. Clin. Biomech. *19*, 415–428.

77. Schuldt, C., Laptev, I., and Caputo, B. (2004). Recognizing human actions: a local SVM approach. In Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004 (IEEE).

78. Belgiu, M., and Drăguţ, L. (2016). Random forest in remote sensing: a review of applications and future directions. ISPRS J. Photogrammetry Remote Sens. *114*, 24–31.

79. Schapire, R.E. (1999). A brief introduction to boosting. In Ijcai.

80. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., and Chen, K. (2015). Xgboost: extreme gradient boosting, pp. 1–4. R package version 0.4-2.

81. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: a highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. *30*.

82. Sagi, O., and Rokach, L. (2018). Ensemble learning: a survey. WIREs Data Mining Knowl. Discov. *8*, e1249.

83. Williams, N., Zander, S., and Armitage, G. (2006). A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification. ACM SIGCOMM Comput. Commun. Rev. *36*, 5–16.

84. Yan, J., Koç, M., and Lee, J. (2004). A prognostic algorithm for machine performance assessment and its application. Prod. Plann. Control *15*, 796–801.

85. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

86. Cai, Y.-D., and Chou, K.-C. (2000). Using neural networks for prediction of subcellular location of prokaryotic and eukaryotic proteins. Mol. Cell Biol. Res. Commun. *4*, 172–173.

87. Emanuelsson, O., Nielsen, H., Brunak, S., and Von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J. Mol. Biol. *300*, 1005–1016.

88. Bodén, M., and Hawkins, J. (2005). Prediction of subcellular localization using sequence-biased recurrent networks. Bioinformatics *21*, 2279–2286.

89. Pärnamaa, T., and Parts, L. (2017). Accurate classification of protein subcellular localization from high-throughput microscopy images using deep learning. G3 *7*, 1385–1392.

90. Yang, F., Liu, Y., Wang, Y., Yin, Z., and Yang, Z. (2019). MIC_Locator: a novel image-based protein subcellular location multi-label prediction model based on multi-scale monogenic signal representation and intensity encoding strategy. BMC Bioinformatics *20*, 522.

91. Long, W., Yang, Y., and Shen, H.-B. (2020). ImPLoc: a multi-instance deep learning model for the prediction of protein subcellular localization based on immunohistochemistry images. Bioinformatics *36*, 2244–2250.

92. Zhang, P., Zhang, M., Liu, H., and Yang, Y. (2022). Prediction of protein subcellular localization based on microscopic images via multi-task multi-instance learning. Chin. J. Electron. *31*, 888–896.

93. Kobayashi, H., Cheveralls, K.C., Leonetti, M.D., and Royer, L.A. (2022). Self-supervised deep learning encodes high-resolution features of protein subcellular localization. Nat. Methods *19*, 995–1003.

94. Aggarwal, S., Gupta, S., Gupta, D., Gulzar, Y., Juneja, S., Alwan, A.A., and Nauman, A. (2023). An artificial intelligence-based stacked ensemble approach for prediction of protein subcellular localization in confocal microscopy images. Sustainability *15*, 1695.

95. Ullah, M., Hadi, F., Song, J., and Yu, D.-J. (2023). PScL-2LSAESM: bioimage-based prediction of protein subcellular localization by integrating heterogeneous features with the two-level SAE-SM and mean ensemble method. Bioinformatics *39*, btac727.

96. Ding, J., Xu, J., Wei, J., Tang, J., and Guo, F. (2023). A multi-scale multi-model deep neural network via ensemble strategy on high-throughput microscopy image for protein subcellular localization. Expert Syst. Appl. *212*, 118744.

97. Gudenas, B.L., and Wang, L. (2018). Prediction of LncRNA subcellular localization with deep learning from sequence features. Sci. Rep. *8*, 16385.

98. Yan, Z., Lécuyer, E., and Blanchette, M. (2019). Prediction of mRNA subcellular localization using deep recurrent neural networks. Bioinformatics *35*, i333–i342.

99. Fukunaga, T., Iwakiri, J., Ono, Y., and Hamada, M. (2019). LncRRIsearch: a web server for lncRNA-RNA interaction prediction integrated with tissue-specific expression and subcellular localization data. Front. Genet. *10*, 462.

100. Lin, Y., Pan, X., and Shen, H.-B. (2021). lncLocator 2.0: a cell-line-specific subcellular localization predictor for long non-coding RNAs with interpretable deep learning. Bioinformatics *37*, 2308–2316.

101. Wang, D., Zhang, Z., Jiang, Y., Mao, Z., Wang, D., Lin, H., and Xu, D. (2021). DM3Loc: multi-label mRNA subcellular localization prediction and analysis based on multi-head self-attention mechanism. Nucleic Acids Res. *49*, e46.

102. Zeng, M., Wu, Y., Lu, C., Zhang, F., Wu, F.-X., and Li, M. (2022). DeepLncLoc: a deep learning framework for long non-coding RNA subcellular localization prediction based on subsequence embedding. Brief. Bioinform. *23*, bbab360.

103. Li, M., Zhao, B., Yin, R., Lu, C., Guo, F., and Zeng, M. (2023). GraphLncLoc: long non-coding RNA subcellular localization prediction using graph convolutional networks based on sequence to graph transformation. Brief. Bioinform. *24*, bbac565.

104. Wang, S., Shen, Z., Liu, T., Long, W., Jiang, L., and Peng, S. (2023). DeepmRNALoc: a novel predictor of eukaryotic mRNA subcellular localization based on deep learning. Molecules *28*, 2284.

105. Cai, J., Wang, T., Deng, X., Tang, L., Liu, L., and GM-lncLoc. (2023). LncRNAs subcellular localization prediction based on graph neural network with meta-learning. BMC Genom. *24*, 52.

106. Andrade, M.A., O'Donoghue, S.I., and Rost, B. (1998). Adaptation of protein surfaces to subcellular location. J. Mol. Biol. *276*, 517–525.

107. Kumar, A., Rao, A., Bhavani, S., Newberg, J.Y., and Murphy, R.F. (2014). Automated analysis of immunohistochemistry images identifies candidate location biomarkers for cancers. Proc. Natl. Acad. Sci. USA *111*, 18249–18254.