Bird Eye View of Protein Subcellular Localization Prediction

Ravindra Kumar [1],* and **Sandeep Kumar Dhanda** [2],*

1   Biometric Research Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute, NIH, 9609 Medical Center Drive, Rockville, MD 20850, USA
2   Department of Oncology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA
*   Correspondence: ravindra.kumar@nih.gov (R.K.); sandeep.dhanda@stjude.org (S.K.D.)

Abstract: Proteins are made up of long chain of amino acids that perform a variety of functions in different organisms. The activity of the proteins is determined by the nucleotide sequence of their genes and by its 3D structure. In addition, it is essential for proteins to be destined to their specific locations or compartments to perform their structure and functions. The challenge of computational prediction of subcellular localization of proteins is addressed in various in silico methods. In this review, we reviewed the progress in this field and offered a bird eye view consisting of a comprehensive listing of tools, types of input features explored, machine learning approaches employed, and evaluation matrices applied. We hope the review will be useful for the researchers working in the field of protein localization predictions.

1. Introduction

Proteins are localized into different cellular compartments and sub-compartments inside the cell. Each subcellular compartment has a distinct well-defined function in the cell and has a characteristic physicochemical environment, which drives proper functioning of the proteins. Each subcellular compartment has a distinct, well defined function in the cell and is considered to have evolved from the prokaryotic cell. Typical eukaryotic cells have two types of DNAs (i) chromosomal nuclear DNA and (ii) organelle DNA, which is present in mitochondria and chloroplast while prokaryotic cells have only single type of DNA called nucleoid. The nuclear DNA encodes the majority of proteins while only a small number of proteins are encoded by organelle DNA. Eukaryotic cells can synthesize up to 100,000 different types of protein [1], which are destined for one or more predetermined subcellular locations. Figure 1 depicts various protein localization prediction methods available for different cellular compartments.

The protein synthesis occurs in the cytoplasm and then the newly synthesized proteins are further transported to their destined compartment to execute their function. Protein must be targeted to the right compartment in cells to perform their function and mis-localization of the proteins leads to functional loss or disorder, which contributes to many human diseases including cardiovascular, neurodegenerative disease and cancers [2,3]. Assigning subcellular localization for protein is a significant step to elucidate its interaction partners and predict their functions or potential roles in the cellular machinery [4]. There are a number of sequences that are deposited every year in the UniProt Knowledgebase (UniProtKB) but only a few of them were manually annotated and reviewed (UniProtKB/SwissProt), which explains the gap between the deposited sequence and annotated sequence is increasing every year (Figure 2). Therefore, there is a need of computational methods to predict

subcellular localization with high quality and accuracy, which is of great significance in understanding cellular proteome and also helpful in designing the drug or targets. To date, many efforts have been made in this regard. Based on different kinds of characteristics, several machine learning approaches have been developed such as neural networks [5,6], hidden Markov models [7–9], support vector machines [10–12], deep learning [13–15], random forest [16], and extreme gradient boosting [17] for prediction of subcellular localization of proteins.

The purpose of this review is not only to make a complete list of protein localization prediction methods but also describe those methods that imply significant developments in this field. In this review we primarily focus on the eukaryotic prediction method however, some of the mentioned subcellular localization prediction methods could be used for both the prokaryotic and eukaryotic cell. This review focuses on (1) available computational methods for subcellular localization prediction, (2) different algorithms used in the development of these methods, (3) various features used in prediction purposes, and (4) future aspect and importance of subcellular localization in the biological field.
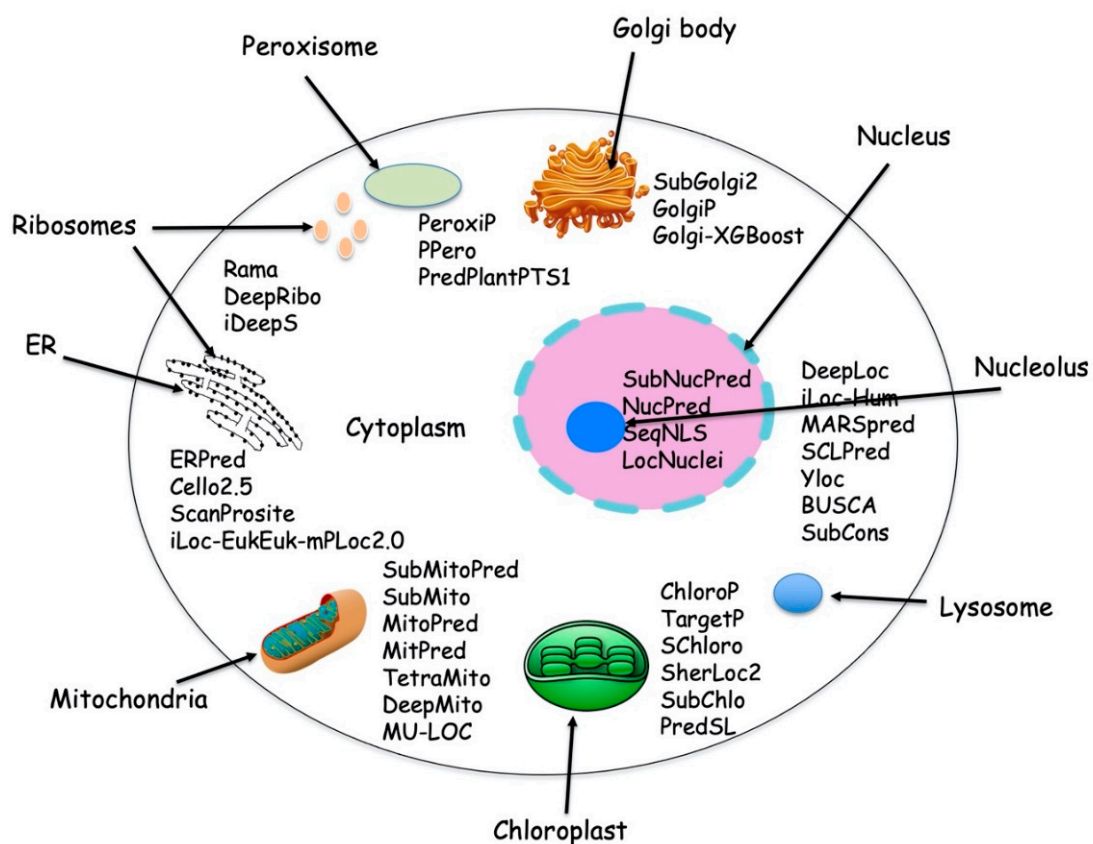

Figure 1. Typical cell with different subcellular location and with available protein localization prediction tools.
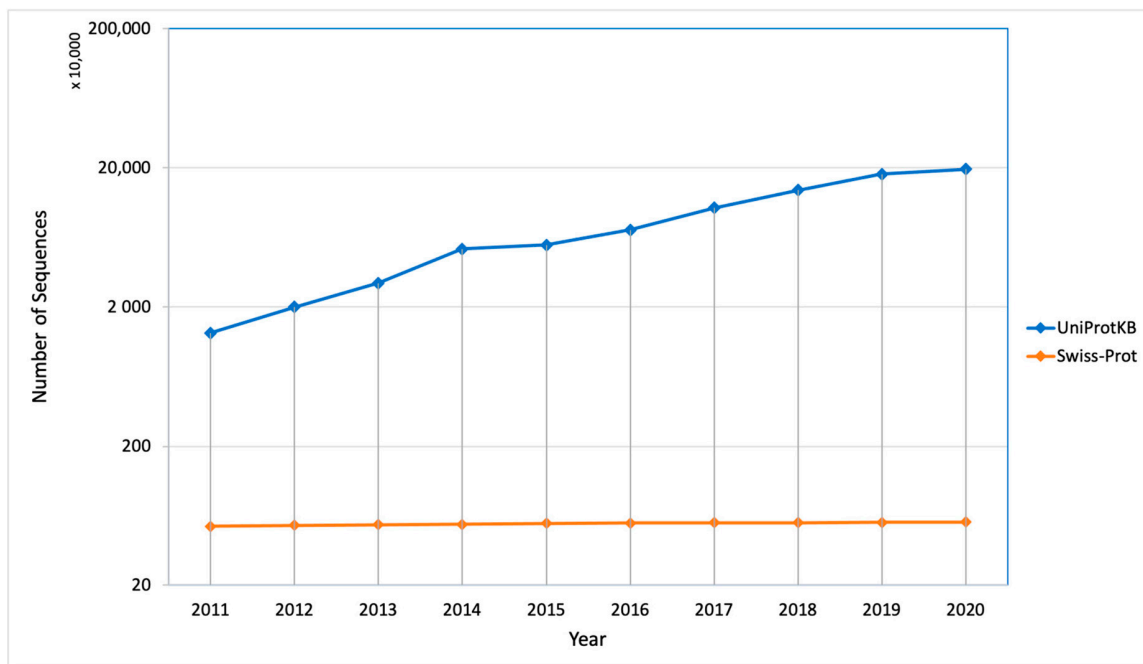
Figure 2. Number of sequences deposited and manually annotated proteins in the UniProt database in the last 10 years.

2. Experimental Approaches for Protein Localization

Several experimental methods are available for determining protein localization, but the most common method is to label the protein of interest with fluorescent probes and then visualize the distribution of the protein within cells under a fluorescence microscope such as immunofluorescence microscopy, immunolocalization, mass spectrometry, co-expression of fluorescent proteins, and electron microscopy. Fractionation based approach such as gradient centrifugation and 2D gel electrophoresis are also a widely used method to experimentally establish the localization of a protein. These experimental methods are relatively expensive and time consuming, which explains for a large information gap existing between known protein and their location information. Consequently, various computational methods have been developed to help fill this void. In this review, we focused only on the computational approaches and tools for prediction of protein localization.

3. Computational Approaches for Protein Localization

With the rapid development of advanced genome sequencing methods, the complete genome sequences are increasing day by day and the challenges for computational biologists are to manage, analyze, and annotate this plethora of unprocessed raw biological data. To now, a number of computational methods have been developed to solve this problem. While many have attempted to explore uncharacterized protein information, others have used the whole proteome sequence information to develop new machine learning algorithms for different things such as the prediction of motifs, prediction of ligand binding sites, etc. Based on protein sequence information, the computational method can be divided into the following categories: (1) sequence feature-based methods, (2) homology-based methods, (3) protein domain and motif information-based methods, (4) signal peptide-based methods, (5) non-sequence derived features-based methods, and (6) integrated methods, which could use a combination of two or more methods.

3.1. Sequence Feature-Based Method

Sequence features are commonly used in localization prediction since some differences in the sequence features are empirically known to be correlated with different localization sites.

Nishikawa and Ooi [18] first noted the correlation of amino acid composition to its biological and functional character in 1982. After that in 1983 Nakashima developed the first sequence-based method for subcellular localization [19]. They used amino acid composition to discriminate between intracellular and extracellular proteins. Later several research groups successfully used amino acid composition as a tool for subcellular localization predictions [16,20,21].

In sequence feature-based methods, the complete sequence of proteins is transformed into a numerical feature vector, which is then used to predict the subcellular location. There are different types of sequence feature-based methods available: (i) amino acid composition based method, in which the frequency of 20 different amino acids is calculated but it ignores the sequence order information of each residue. (ii) Chou's Pseudo amino acid composition (PseAAC) [22], which considers the amino acid composition along with the potential interaction among the adjacent residues. This can be further categorized into different modes of PseAAC such as the gene ontology mode, functional domain mode and sequential evolution mode. (iii) Hybrid method, which allows the integration of different parameters or features for the prediction and usually results in an increased the prediction performance [4].

3.2. Homology Based Method

This is the most common way to predict the uncharacterized protein on the basis of the presence of homologous sequences of known function with an assumption that function is evolutionarily conserved [23]. This approach first identifies for homologous sequences in the proteins with known subcellular location and then extrapolates to predict the location of unknown proteins, hence this approach is also known as "Annotation by Homology Transfer". Homology is a qualitative term, which attributes evolutionary relationships among different protein sequences. Orthologous proteins also typically have similar sequences and thus similar subcellular localization patterns. Proteins with a highly similar sequence correlate well with the cellular localization site while those with dissimilar sequences indicate that they are distant and may or may not be colocalized. In 2002, Nair and Rost [24] showed the correlation between sequence similarity with subcellular localization. They considered 11 different compartments and observed sequence conservation among the major compartments. BLAST, PSI-BLAST, and hidden Markov models (HMM) are routinely used for searching homologous sequences. The limitation of homology-based methods is more pronounced in cases where no homology is found between the query sequence and the annotated proteins sequence. Additionally, it is known that a single amino acid substitution in localization signals can change the localization of a protein [25–27]. Thus, sequence homology is a noncausal feature for the localization prediction and should be used with caution when applied to nonnative sequences or in case when homology is less [28].

3.3. Functional Motifs, Domains, and other Signatures Based Method

Proteins have evolved in different compartments, which limit their interactions with other proteins and ultimately impact their functions. Some of these proteins preserved some sequential or structural patterns or motifs. Though not all of these motifs and domains are specific to subcellular localization, many preferentially occur in some specific compartments and such domains can be used to predict the localization of any proteins. Studying proteins at a domain/motif level allows more accurate functional inference [29]. In 2002, Mott et al. [30] first used 300 Simple Modular Architecture Research Tool (SMART) domains to predict three subcellular locations viz secreted, cytoplasm, and nucleus. After that, several works have been used for the protein motif and domains as features for protein localization predictions [31,32].

These motifs are not just limited to sequence patterns, but also extended to the structural information. There are a couple of tools such as PROSITE [33] and MEME [34] that employed this feature to use for protein localization. While the structure is not available for a big chunk of protein sequences, this gap is filled by several proteins structure predictions servers, like I-TASSER and C-I-TASSER servers [35]

3.4. Signal Peptide Based Method

Signal peptides are short amino acid sequences in the amino terminus of the newly synthesized proteins and are found in all organisms including bacteria, archaea, and eukaryotes. The function of the signal peptide is to enable the transport machinery to translocate the proteins to different subcellular locations. They are present in secretory proteins and in transmembrane proteins and the protein residing in different eukaryotic organelles have different types of signal peptide sequences [36]. The signal peptide is followed by a stretch of amino acids that form the cleavage site recognized by peptidases and the signal peptide is removed after translocation, except in the case of transmembrane proteins. In case of transmembrane proteins, this signal peptide serves as signal anchor sequences. The importance of various signal peptide sequences in proteins in their subcellular localization has led to attempts to predict the subcellular location on the basis of the signal peptide present in proteins. The prediction of the signal peptide involves two main tasks: (1) discriminating between the signal peptide and signal anchor sequences and (2) also predicting the position of the signal peptide cleavage site [37]. The major challenge in signal peptide prediction is discriminating between true signal sequences and other hydrophobic regions. In addition to it, the accurate prediction of the cleavage site is also very important due to the high variability of the signal sequence length and the absence of sequence motifs that unambiguously mark the position of the cutting site [38]. A number of prediction methods are available that recognize and predict the subcellular location on the basis of signal peptides (Table 1). SignalP was the first publicly available method [15] and there are many versions available, which were developed based on different methods. Version-1 [39] was based on artificial neural networks, while version-2 [40] was based on hidden Markow models, version-3 [41] has an improved cleavage site prediction, version-4 [42] has improved discrimination of signal peptides and TM helices, and version-5 [38] is a deep neural network-based method combined with a conditional random field classification and an optimized transfer learning for improved signal peptide prediction.

Table 1. Some useful signal peptide-based methods.

Method	Tools Used	Performance Matrix	Locations/Organism	Availability	Year
SignalP-5.0 *	convolutional and recurrent (LSTM) neural networks	MCC, precision and recall	Archaea, Gram-positive Bacteria, Gram-negative Bacteria and Eukarya	http://www.cbs.dtu.dk/services/SignalP/	2019
TargetP 2.0 *	recurrent neural networks (RNNs) network	Precision, recall, F1-score, MCC	mitochondrial, chloroplastic, secretory pathway	http://www.cbs.dtu.dk/services/TargetP/	2019
SigUNet	Convolutional neural network	MCC, precision, recall, F1 measure	Eukaryotes, Gram-positive and Gram-negative bacteria	https://github.com/mbilab/SigUNet	2019
DeepSig	Convolutional Neural Networks	MCC, False Positive Rate, precision and recall	Eukaryotes, Gram-positive bacteria and Gram-negative bacteria	https://deepsig.biocomp.unibo.it	2018
SChloro	SVM	Accuracy, Recall, Precision, F1-score, and MCC	six chloroplastic sub-compartments	http://schloro.biocomp.unibo.it	2017
PredSL	combination of neural networks, Markov chains, scoring matrices (PrediSi), and HMMs,	Accuracy	Eukaryotic subcellular location	http://bioinformatics.biol.uoa.gr/PredSL/	2006
TatP	HMM/artificial neural networks.	S-score and the C-score, Y-score, D-score	bacteria	http://www.cbs.dtu.dk/services/TatP/	2005

Table 1. *Cont.*

Method	Tools Used	Performance Matrix	Locations/Organism	Availability	Year
ChloroP	Neural network	MCC, sensitivity, specificity	chloroplast transit peptides	http://www.cbs.dtu.dk/services/ChloroP/	1999

* There are different versions of the software available, but here we mentioned only the recent one.

The signal peptide-based method is a good approach to predict the proteins that contain the signal peptide, but it has some drawbacks, which make these methods not able to be applied for proteome scale prediction. (i) Not all proteins contain signal peptides. There are many proteins that do not have any reported signal peptide sequence and despite this are still translocated to their respective subcellular location. (ii) Many proteins follow the "piggyback import" mechanism during protein translocation, which means these proteins do not have any specific signal peptide for the localization, but they interact and bind to different proteins that have a signal peptide for translocation and then are co-imported to specific target locations [43,44].

3.5. Non-Sequence Derived Features

A variety of non-sequence derived features have been used to predict subcellular localization. For example, LOC3D [45], which used the structural information for identification and prediction of proteins subcellular locations. There are a number of non-sequence derived features that have been used in an automated classifiers including immunohistochemistry [3,46,47], fluorescence microscopy image [48,49], protein–protein interaction (PPI) data [50], expression data [51], and recommendation systems [52].

3.6. Integrated Method

The different strategies for predicting protein localization have their own strengths and weaknesses. To enhance the performance of prediction methods, it is important to combine multi-characteristic strategies, which give more complete information to understand the relationship between protein localization with its sequence, structure, physicochemical properties, and function. Hence a combination of different input vectors and different tools will be the successful strategy in protein subcellular localization prediction. Many methods have successfully utilized the combination of protein features to enhance the performance of protein subcellular localization predictions. The Protein Subcellular Localization Prediction Tool (PSORT) family method is one of the integrated methods, which contains several tools for localization prediction. The family includes a number of tools: (i) PSORT [53], the first integrated method of the PSORT family (http://psort.org) for the plant and bacterial protein, (ii) PSORT II [54] for yeast and animal proteins, (iii) iPSORT [55] for N-terminal sorting signals for plant or non-plants; (iv) PSORTb [56–58] for bacterial and archaeal proteins, and (v) WoLF PSORT [59] for eukaryotic proteins including plants, animals, and fungi.

A similar approach was taken by many researchers where they integrated biological or empirical sequence features correlated with subcellular location with a variety of machine-learning algorithm such as KNN, SVM, and deep learning: MultiLoc, integration of the phylogenetic profile and GO terms of retrieved homologues such as MultiLoc2, CELLO2.5, SherLoc2, YLoc, iLoc-Euk, Loctree3, DeepLoc, etc. People are also integrating different computational tools for predicting subcellular localization. The Bologna Unified Subcellular Component Annotator (BUSCA) [60] is an example of such an integrated tool where the author combines methods for identifying signal and transit peptides (DeepSig and TPpred3), GPI-anchoes (PredGPI), and transmembrane domains (ENSEMBLE3.0 and BetAware) with tools for discriminating subcellular localization of both globular and membrane proteins (BaCelLo, MemLoci, and SChloro). This integrated method performs better than the other methods based on single feature approaches. There are a number of recently developed subcellular localization methods available, which are used by a wide range of researchers (Table 2 and Table S1).

Table 2. List of subcellular localization methods.

Method	Tools Used	Performance Matrix	Feature Based	Locations/Organism	Availability	Year
DeepPred-SubMito	Convolutional neural network	Accuracy, MCC	Sequence information	Mitochondrial and submitochondrial proteins	https://github.com/ jinyinping/DeepPred-SubMito.git	2020
SubMito-XGBoost	Extreme gradient boosting (XGBoost)	Sensitivity, Specificity, False positive rate, MCC, F1-measure, precision	Sequence information	Submitochondrial proteins	https://github.com/ QUST-AIBBDRC/ SubMito-XGBoost/	2020
mRNALoc	SVM	Sensitivity, Specificity, Accuracy, MCC	Sequence information	eukaryotic	http: //proteininformatics. org/mkumar/mrnaloc	2020
SCLpred-EMS	Convolutional neural network	Sensitivity, Specificity, False positive rate, MCC	Sequence information	endomembrane system and secretory pathway	http://distilldeep.ucd. ie/SCLpred2/	2020
BUSCA	Integrated method of DeepSig, TPpred3, PredGPI, BetAware and ENSEMBLE3.0	Precision, recall, F1-score, MCC	Sequence information, signal and transit peptides, glycophosphatidylinositol (GPI) anchors and transmembrane domains	Gram-positive, gram-negative, fungi, plant, animal	http://busca.biocomp. unibo.it	2018
SubMitoPred	SVM	Sensitivity, Specificity, Accuracy, MCC	Sequence and domain information	Mitochondrial and submitochondrial proteins	http: //proteininformatics. org/mkumar/ submitopred/	2018
pLoc-mEuk	ML-GKR (multi-label Gaussian kernel regression) classifier	Coverage, Accuracy, Absolute true, Absolute false	Gene Ontology and Chou's general PseAAC	22 different subcellular localizations of eukaryotic proteins	http: //www.jci-bioinfo.cn/ pLoc-mEuk/	2018
ERPred	SVM	Sensitivity, Specificity, Accuracy, MCC	Sequence information,	ER Proteins	http: //proteininformatics. org/mkumar/erpred/ index.html	2017
DeepLoc	deep recurrent neural networks	Accuracy, MCC	Sequence information	10 different location of eukaryotic proteins	http://www.cbs.dtu. dk/services/DeepLoc	2017

Method	Tools Used	Performance Matrix	Feature Based	Locations/Organism	Availability	Year
SubNucPred	SVM	Sensitivity, Specificity, Accuracy, MCC	Sequence and domain information	Nuclear and subnuclear protein	http://proteininformatics.org/mkumar/subnucpred/	2014
LocTree3	SVM and homology	Accuracy, recall, standard deviation, standard error	Homology-based, Gene Ontology	18 classes for eukaryotes, in six for bacteria and in three for archaea	http://www.rostlab.org/services/loctree3	2014
PlantLoc	localization motif search	accuracy	localization motif information	11 different location of plant proteins	http://cal.tongji.edu.cn/PlantLoc/	2013
iLoc-Cell, package of predictors for subcellular locations of proteins. It includes iLoc-Hum, iLoc-Animal, iLoc-Plant, iLoc-Euk, iLoc-Virus, iLoc-Gpos, iLoc-Gneg	multi-label learning, multi-label KNN	Accuracy, Precision, Recall, Absolute-true rate, Absolute-false rate,	Sequence information, gene ontology, PSSM,	Different subcellular location of Human, animals, plants, eukaryotic, Virus, gram-positive, gram-negatives	http://www.jci-bioinfo.cn/iLoc-Cell	2011, 2012, 2013
MARSpred	SVM	Sensitivity, specificity, Accuracy, MCC	Sequence information, PSSM	cytosolic and mitochondrial aminoacyl tRNA synthetase	http://www.imtech.res.in/raghava/marspred/	2012
SCLPred	Neural Network	Sensitivity, specificity, False positive rate, MCC	primary sequence and multiple sequence alignments	four classes for animals and fungi and five classes for plants	http://distill.ucd.ie/distill/	2011
AtSubP	SVM	Sensitivity, specificity, error rate, MCC, ROC curve	Sequence information, PSSM	subcellular localization of Arabidopsis	http://bioinfo3.noble.org/AtSubP	2010

Table 2. *Cont.*

Method	Tools Used	Performance Matrix	Feature Based	Locations/Organism	Availability	Year
Euk-mPLoc 2.0	OET-KNN (Optimized Evidence-Theoretic K-Nearest Neighbor) classifiers	accuracy	gene ontology information, functional domain information, and sequential evolutionary information	eukaryotic proteins among the following 22 locations	http://www.csbio.sjtu.edu.cn/bioinf/euk-multi-2/	2010
PSORTb	SVM	Precision, recall, accuracy, MCC	Sequence information	Different subcellular location of Gram-negative, Gram-positive, archaea	http://www.psort.org/psortb	2010
YLoc	naïve Bayes alongside entropy-based discretization	overall accuracy, F1-score	Sequences information, GO-term and motif	animal, fungal and plant proteins	www.multiloc.org/YLoc	2010
SubChlo	evidence-theoretic *K*-nearest neighbor (ET-KNN) algorithm	overall accuracy, accuracy	Sequences information (PseAAC),	chloroplast proteins	http://bioinfo.au.tsinghua.edu.cn/subchlo	2009
MultiLoc2	SVM	Sensitivity, specificity, Accuracy, MCC	phylogenetic profiles and gene ontology terms	Plant, Animal, Fungal	https://abi-services.informatik.uni-tuebingen.de/multiloc2/webloc.cgi	2009
AAIndexLoc	SVM	Sensitivity, specificity, Accuracy, MCC	Sequence information and physicochemical properties	Animal, Fungal and plants	http://aaindexloc.bii.a-star.edu.sg	2008
Cell-PLoc package of predictors for subcellular locations of proteins. It includes Euk-mPLoc, Hum-mPLoc, Plant-PLoc, Gpos-PLoc, Gneg-PLoc, Virus-PLoc	KNN or OET-KN algorithm	Accuracy and F1 score	GO and functional domain information	22 subcellular location of eukaryotic, human, plant, Gram-positive bacterial, Gram-negative bacterial and viral proteins	http://chou.med.harvard.edu/bioinf/Cell-PLoc	2008

Table 2. *Cont.*

Method	Tools Used	Performance Matrix	Feature Based	Locations/Organism	Availability	Year
ProLoc-GO	SVM-GO, k-NN-GO and fuzzy k-NN-GO	MCC	GO term information	eukaryotic, human,	http://iclab.life.nctu.edu.tw/prologo	2008
ProLoc	SVM	Accuracy	physicochemical composition	subnuclear localizations	http://iclab.life.nctu.edu.tw/proloc	2007
SherLoc	SVM	Sensitivity, specificity, MCC	Sequence information	eukaryotic proteins	http://www-bs.informatik.uni-tuebingen.de/Services/SherLoc/	2007
MitPred	SVM	Sensitivity, specificity, Accuracy, MCC	Sequence information	Mitochondrial proteins	http://www.imtech.res.in/raghava/mitpred/	2006
BaCelLo	SVM	Coverage, Normalized Accuracy, geometric average, overall accuracy, Generalized Correlation	Sequence information	Plant, Animal, Fungal	http://www.biocomp.unibo.it/bacell/	2006
HSLpred	SVM	Accuracy, MCC, Reliability index	Sequence information	Human Protein	http://www.imtech.res.in/raghava/hslpred/	2005
PSLpred	SVM	Accuracy, MCC, Reliability index	Sequence information	gram-negative bacterial proteins	http://www.imtech.res.in/raghava/pslpred/	2005
ESLpred	SVM	Accuracy, MCC, Reliability index	Sequence information and PSSM	eukaryotic proteins	http://www.imtech.res.in/raghava/eslpred/	2004

4. Machine Learning Tools Used in Protein Prediction

A number of machine learning methods have been successfully employed for protein subcellular localization predictions using the sequence information for all the proteins with known subcellular localization information to learn and develop a meaningful representation or representative model of the biological data. In the machine learning, the biological data is first categorized into classes based on the Figure 3, which represents the flowchart of a typical machine learning approach for subcellular protein localization. The performance of machine learning model has to be tested using various parameters. Below are a few parameters that are widely used to analyze machine learning performances (Table 3).
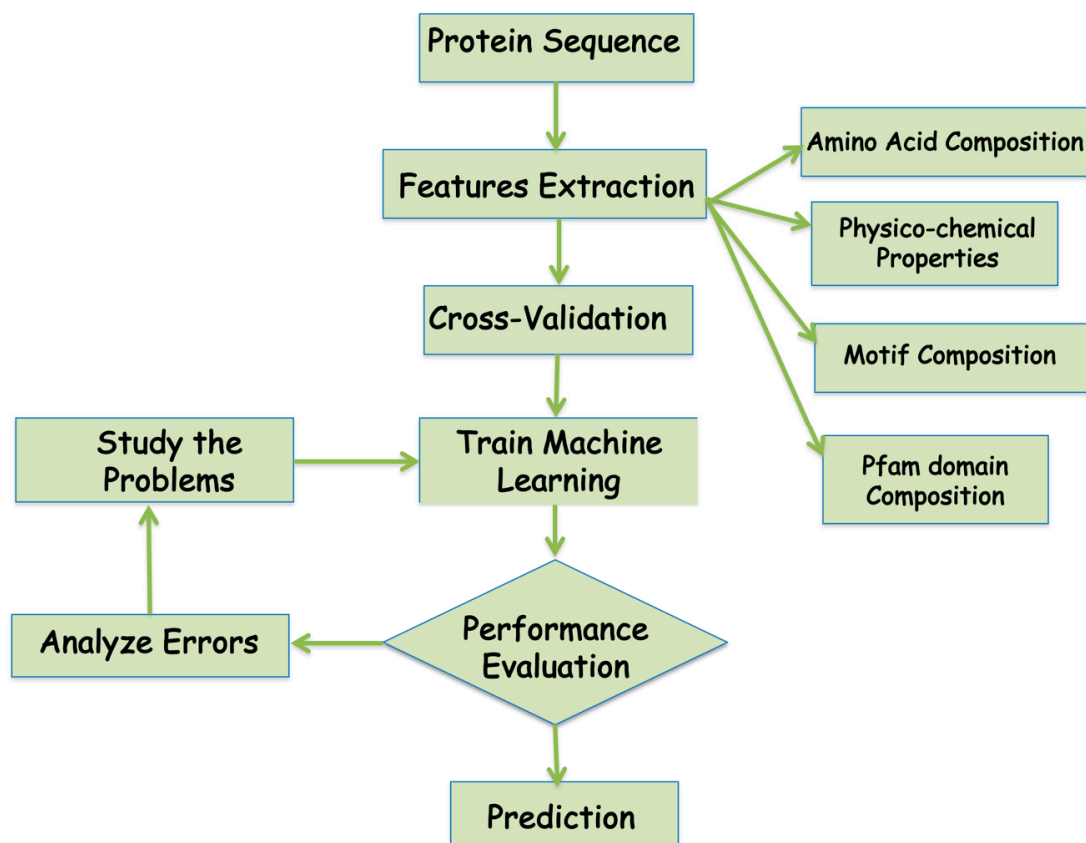
Figure 3. Flowchart of the development of the machine learning algorithm for the prediction of protein localization.

Table 3. List of commonly used metrics found in protein subcellular localization prediction.

Metric	Explanation	Formula	References
Accuracy	It is the ratio of the number of correct predictions to the total number of input samples. It works well if there are an equal number of samples belonging to each class.	TP + TN/(TP + FP + TN + FN)	[61,62]
Specificity/True Negative Rate	Proportion of negatives that are correctly identified	TN/(TN + FP)	[61,62]

Table 3. *Cont.*

Metric	Explanation	Formula	References
Precision/Positive Predictive Value	It tells us how many of the correctly predicted cases actually turned out to be positive.	TP/(TP + FP)	[61]
Recall/Sensitivity/True Positive Rate	It tells us how many of the actual positive cases we were able to predict correctly with our model.	TP/(TP + FN)	[61,62]
F1-Score	F1 score is the harmonic mean between precision and recall. Greater the F1 score, better the performance of the model. It tells how precise the classifier is and how robust it is.	F1 = 2 * precision * Recall/Precision + Recall	[63]
Negative Predictive Value	It is defined as the proportion of predicted negatives, which are real negatives. It reflects the probability that a predicted negative is a true negative.	TN/(TN + FN)	[64]
False Positive Rate	Number of incorrect positive predictions divided by the total number of negatives.	FP/(FP + TN)	[61]

Note: TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative respectively.

4.1. Support Vector Machine

Support vector machine (SVM) is a supervised machine learning tool, which is used for classification and regression purposes for complex biological problems. It is based on statistical learning theory and was developed by Vapnik in the year 1995 [65]. It can solve linear and non-linear issues well for many problems such as subcellular localization prediction, different protein family classification prediction. The objective of an SVM is to find a hyperplane in a N-dimensional space that classifies the data into different classes with maximum margin. There are a number of protein subcellular localization methods available that used SVM as a prediction tool [7,11,21,66]. For implementation of SVM, people used different packages of SVM like SMV_light, LIBSVM, Caret in R, and Scikit-Learn for python.

4.2. Random Forest

Random forest (RF) is an ensemble learning method for classification and regression analysis. The key element of RF is to build multiple decision trees and merge them together to get a more accurate and stable prediction. This is the most popular choice for bioinformaticians to analyze the complex biological data. Apart from subcellular localization prediction [2], RF is applied in a variety of problems such as gene expression classification [67,68], biomarker discovery [69–72], protein–protein interaction [73,74], and cancer drug predictions [75–78].

4.3. Neural Network and Deep Learning

The successes of neural networks have led to the development of various programming frameworks to build and train neural network models. The traditional neural network architecture is the feed forward neural network with one hidden layer, in which each input neuron is connected to each neuron in the hidden layer and which is further connected to each neuron in the output layer. Apart from the subcellular localization prediction [13,79], deep learning is successfully applied in many other biological fields such as the prediction of splicing pattern prediction [80,81], protein secondary structure prediction [82,83], different types of cancer and drug-target interactions [84–87], and the patterns in the biomedical imaging datasets [88].

There are some popular python libraries, which implements deep learning such as PyTorch, Caffe, TensorFlow, Theano, Keras, and Lasagne [89]. Keras is the most powerful and easy to use python library for developing and evaluating deep learning models.

5. Importance and Future Aspects of Protein Subcellular Localization

Protein subcellular localization is getting more attention as the function of the protein requires its localization to a destined compartment. With increase in genomics and proteomics sequencing has led to the accumulation of a large number of sequences in different databases. We believe that the computational methods developed using the integration of different sources of information will be the key factor to resolve these issues. Most of the current prediction methods are feature based methods, which used sequence information or physicochemical properties. The drawback of these feature-based methods is that it does not utilize rich network information of the proteins such as the gene co-expression network, genetic interaction, and metabolic network [90,91]. There are few methods available that also use PPI [92] and the metabolic network [93] for protein localization predictions. The proteins interacting with each other tend to localize within the same subcellular compartments and thus, the PPI information could be useful in predicting subcellular localizations [25]. Additionally, it is important to explore more interacting partners information to develop a more accurate localization prediction method. Another thing that we need to improve in development of protein localization prediction methods is the basis of most of the methods is that a particular protein can be destined to a single subcellular location, which lacks the capability to predict the proteins present in multiple locations. More efforts are needed for development of protein localization methods to address proteins present in multiple locations. There is still some room for major improvements in the multicellular localization prediction with high reliability while in many ways the field still holds interesting challenges for the bioinformatician.

References

1. Harper, J.W.; Bennett, E.J. Proteome complexity and the forces that drive proteome imbalance. *Nature* **2016**, *537*, 328–338. [CrossRef] [PubMed]
2. Zhao, L.; Wang, J.; Nabil, M.M.; Zhang, J. Deep Forest-based Prediction of Protein Subcellular Localization. *Curr. Gene Ther.* **2018**, *18*, 268–274. [CrossRef] [PubMed]

3.   Xue, Z.-Z.; Wu, Y.; Gao, Q.-Z.; Zhao, L.; Xu, Y.-Y. Automated classification of protein subcellular localization in immunohistochemistry images to reveal biomarkers in colon cancer. *BMC Bioinform.* **2020**, *21*, 1–15. [CrossRef] [PubMed]

4.   Li, B.; Cai, L.; Liao, B.; Fu, X.; Bing, P.; Yang, J. Prediction of Protein Subcellular Localization Based on Fusion of Multi-view Features. *Molecules* **2019**, *24*, 919. [CrossRef]

5.   Mooney, C.; Wang, Y.; Pollastri, G. SCLpred: Protein subcellular localization prediction by N-to-1 neural networks. *Bioinformatics* **2011**, *27*, 2812–2819. [CrossRef]

6.   Emanuelsson, O.; Nielsen, H.; Von Heijne, G. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci.* **1999**, *8*, 978–984. [CrossRef]

7.   Kumar, R.; Jain, S.; Kumari, B.; Kumar, M. Protein Sub-Nuclear Localization Prediction Using SVM and Pfam Domain Information. *PLoS ONE* **2014**, *9*, e98345. [CrossRef]

8.   Kumar, M.; Raghava, G. Prediction of nuclear proteins using SVM and HMM models. *BMC Bioinform.* **2009**, *10*, 22. [CrossRef]

9.   Chen, Y.; Yu, P.; Luo, J.; Jiang, Y. Secreted protein prediction system combining CJ-SPHMM, TMHMM, and PSORT. *Mamm. Genome* **2003**, *14*, 859–865. [CrossRef]

10.   Li, G.-P.; Du, P.-F.; Shen, Z.-A.; Liu, H.-Y.; Luo, T. DPPN-SVM: Computational Identification of Mis-Localized Proteins in Cancers by Integrating Differential Gene Expressions With Dynamic Protein-Protein Interaction Networks. *Front. Genet.* **2020**, *11*, 600454. [CrossRef]

11.   Kumar, R.; Kumari, B.; Kumar, M. Proteome-wide prediction and annotation of mitochondrial and sub-mitochondrial proteins by incorporating domain information. *Mitochondrion* **2018**, *42*, 11–22. [CrossRef] [PubMed]

12.   Garg, A.; Singhal, N.; Kumar, R.; Kumar, M. mRNALoc: A novel machine-learning based in-silico tool to predict mRNA subcellular localization. *Nucleic Acids Res.* **2020**, *48*, W239–W243. [CrossRef]

13.   Armenteros, J.J.A.; Sønderby, C.K.; Sønderby, S.K.; Nielsen, H.; Winther, O. DeepLoc: Prediction of protein subcellular localization using deep learning. *Bioinformatics* **2017**, *33*, 3387–3395. [CrossRef] [PubMed]

14.   Kaleel, M.; Zheng, Y.; Chen, J.; Feng, X.; Simpson, J.C.; Pollastri, G.; Mooney, C. SCLpred-EMS: Subcellular localization prediction of endomembrane system and secretory pathway proteins by Deep N-to-1 Convolutional Neural Networks. *Bioinformatics* **2020**, *36*, 3343–3349. [CrossRef] [PubMed]

15.   Savojardo, C.; Bruciaferri, N.; Tartari, G.; Martelli, P.L.; Casadio, R. DeepMito: Accurate prediction of protein sub-mitochondrial localization using convolutional neural networks. *Bioinformatics* **2020**, *36*, 56–64. [CrossRef]

16.   Lv, Z.; Jin, S.; Ding, H.; Zou, Q. A Random Forest Sub-Golgi Protein Classifier Optimized via Dipeptide and Amino Acid Composition Features. *Front. Bioeng. Biotechnol.* **2019**, *7*, 215. [CrossRef]

17.   Yu, B.; Qiu, W.; Chen, C.; Ma, A.; Jiang, J.; Zhou, H.; Ma, Q. SubMito-XGBoost: Predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting. *Bioinformatics* **2020**, *36*, 1074–1081. [CrossRef]

18.   Nishikawa, K.; Ooi, T. Correlation of the Amino Acid Composition of a Protein to Its Structural and Biological Characters1. *J. Biochem.* **1982**, *91*, 1821–1824. [CrossRef]

19.   Nishikawa, K.; Kubota, Y.; Ooi, T. Classification of Proteins into Groups Based on Amino Acid Composition and Other Characters. II. Grouping into Four Types. *J. Biochem.* **1983**, *94*, 997–1007. [CrossRef]

20.   Behbahani, M.; Nosrati, M.; Moradi, M.; Mohabatkar, H. Using Chou's General Pseudo Amino Acid Composition to Classify Laccases from Bacterial and Fungal Sources via Chou's Five-Step Rule. *Appl. Biochem. Biotechnol.* **2020**, *190*, 1035–1048. [CrossRef]

21.   Kumar, R.; Kumari, B.; Kumar, M. Prediction of endoplasmic reticulum resident proteins using fragmented amino acid composition and support vector machine. *PeerJ* **2017**, *5*, e3561. [CrossRef] [PubMed]

22.   Chou, K.-C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins Struct. Funct. Bioinform.* **2001**, *43*, 246–255. [CrossRef] [PubMed]

23.   Cozzetto, D.; Jones, D.T. Computational Methods for Annotation Transfers from Sequence. In *The Gene Ontology Handbook Methods in Molecular Biology*; Dessimoz, C.Š.N., Ed.; Humana Press: New York, NY, USA, 2017.

24.   Nair, R.; Rost, B. Sequence conserved for subcellular localization. *Protein Sci.* **2002**, *11*, 2836–2847. [CrossRef] [PubMed]

25.   Silver, P.A.; Chiang, A.; Sadler, I. Mutations that alter both localization and production of a yeast nuclear protein. *Genes Dev.* **1988**, *2*, 707–717. [CrossRef]

26. Freeman, B.T.; Sokolowski, M.; Roy-Engel, A.M.; Smither, M.E.; Belancio, V.P. Identification of charged amino acids required for nuclear localization of human L1 ORF1 protein. *Mob. DNA* **2019**, *10*, 20. [CrossRef]

27. Laurila, K.; Vihinen, M. Prediction of disease-related mutations affecting protein localization. *BMC Genom.* **2009**, *10*, 122. [CrossRef]

28. Nakai, K.; Horton, P. Computational Prediction of Subcellular Localization. *Methods Mol. Biol.* **2007**, *390*, 429–466. [CrossRef]

29. Loewenstein, Y.; Raimondo, D.; Redfern, O.C.; Watson, J.; Frishman, D.; Linial, M.; Orengo, C.; Thornton, J.; Tramontano, A. Protein function annotation by homology-based inference. *Genome Biol.* **2009**, *10*, 1–8. [CrossRef]

30. Mott, R.; Schultz, J.; Bork, P.; Ponting, C.P. Predicting Protein Cellular Localization Using a Domain Projection Method. *Genome Res.* **2002**, *12*, 1168–1174. [CrossRef]

31. Guda, C.; Subramaniam, S. TARGET: A new method for predicting protein subcellular localization in eukaryotes. *Bioinformatics* **2005**, *21*, 3963–3969. [CrossRef]

32. Nair, R.; Rost, B. Mimicking Cellular Sorting Improves Prediction of Subcellular Localization. *J. Mol. Biol.* **2005**, *348*, 85–100. [CrossRef] [PubMed]

33. Sigrist, C.J.A.; Cerutti, L.; Hulo, N.; Gattiker, A.; Falquet, L.; Pagni, M.; Bairoch, A.; Bucher, P. PROSITE: A documented database using patterns and profiles as motif descriptors. *Brief. Bioinform.* **2002**, *3*, 265–274. [CrossRef] [PubMed]

34. Bailey, T.L.; Williams, N.; Misleh, C.; Li, W.W. MEME: Discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* **2006**, *34*, W369–W373. [CrossRef] [PubMed]

35. Yang, J.; Zhang, Y. I-TASSER server: New development for protein structure and function predictions. *Nucleic Acids Res.* **2015**, *43*, W174–W181. [CrossRef] [PubMed]

36. Armenteros, J.J.A.; Tsirigos, K.D.; Sønderby, C.K.; Petersen, T.N.; Winther, O.; Brunak, S.; Von Heijne, G.; Nielsen, H. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* **2019**, *37*, 420–423. [CrossRef] [PubMed]

37. Nielsen, H.; Tsirigos, K.D.; Brunak, S.; Von Heijne, G. A Brief History of Protein Sorting Prediction. *Protein J.* **2019**, *38*, 200–216. [CrossRef] [PubMed]

38. Savojardo, C.; Martelli, P.L.; Fariselli, P.; Casadio, R. DeepSig: Deep learning improves signal peptide detection in proteins. *Bioinformatics* **2018**, *34*, 1690–1696. [CrossRef]

39. Nielsen, H.; Engelbrecht, J.; Brunak, S.; Von Heijne, G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **1997**, *10*, 1–6. [CrossRef]

40. Nielsen, H.; Krogh, A. Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc. Int. Conf. Intell. Syst. Mol. Boil.* **1998**, *6*, 122–130.

41. Bendtsen, J.D.; Nielsen, H.; Von Heijne, G.; Brunak, S. Improved Prediction of Signal Peptides: SignalP 3.0. *J. Mol. Biol.* **2004**, *340*, 783–795. [CrossRef]

42. Petersen, T.N.; Brunak, S.; Von Heijne, G.; Nielsen, H. SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nat. Methods* **2011**, *8*, 785–786. [CrossRef] [PubMed]

43. Thoms, S. Import of proteins into peroxisomes: Piggybacking to a new home away from home. *Open Biol.* **2015**, *5*, 150148. [CrossRef] [PubMed]

44. Tessier, T.M.; MacNeil, K.M.; Mymryk, J.S. Piggybacking on Classical Import and Other Non-Classical Mechanisms of Nuclear Import Appear Highly Prevalent within the Human Proteome. *Biology* **2020**, *9*, 188. [CrossRef] [PubMed]

45. Nair, R. LOC3D: Annotate sub-cellular localization for protein structures. *Nucleic Acids Res.* **2003**, *31*, 3337–3340. [CrossRef]

46. Kumar, A.; Rao, A.; Bhavani, S.; Newberg, J.Y.; Murphy, R.F. Automated analysis of immunohistochemistry images identifies candidate location biomarkers for cancers. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 18249–18254. [CrossRef]

47. Xu, Y.-Y.; Shen, H.-B.; Murphy, R.F. Learning complex subcellular distribution patterns of proteins via analysis of immunohistochemistry images. *Bioinformatics* **2020**, *36*, 1908–1914. [CrossRef]

48. Tahir, M.; Khan, A.; Kaya, H. Protein subcellular localization in human and hamster cell lines: Employing local ternary patterns of fluorescence microscopy images. *J. Theor. Biol.* **2014**, *340*, 85–95. [CrossRef]

49. Xiao, X.; Shao, S.; Ding, Y.; Huang, Z.; Chou, K.-C. Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. *Amino Acids* **2006**, *30*, 49–54. [CrossRef]

50. Garapati, H.S.; Male, G.; Mishra, K. Predicting subcellular localization of proteins using protein-protein interaction data. *Genomics* **2020**, *112*, 2361–2368. [CrossRef]

51. Ryngajłło, M.; Childs, L.H.; Lohse, M.; Giorgi, F.M.; Elude, A.; Selbig, J.; Usadel, B. SLocX: Predicting subcellular localization of Arabidopsis proteins leveraging gene expression data. *Front. Plant Sci.* **2011**, *2*, 43. [CrossRef]

52. Mehrabad, E.M.; Hassanzadeh, R.; Eslahchi, C. PMLPR: A novel method for predicting subcellular localization based on recommender systems. *Sci. Rep.* **2018**, *8*, 12006. [CrossRef] [PubMed]

53. Nakai, K.; Kanehisa, M. Expert system for predicting protein localization sites in gram-negative bacteria. *Proteins* **1991**, *11*, 95–110. [CrossRef] [PubMed]

54. Horton, P.; Nakai, K. Better prediction of protein cellular localization sites with the k nearest neighbors classifier. *Proc. Int. Conf. Intell. Syst. Mol. Boil.* **1997**, *5*, 147–152.

55. Bannai, H.; Tamada, Y.; Maruyama, O.; Nakai, K.; Miyano, S. Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics* **2002**, *18*, 298–305. [CrossRef] [PubMed]

56. Gardy, J.L.; Spencer, C.; Wang, K.; Ester, M.; Tusnády, G.E.; Simon, I.; Hua, S.; Defays, K.; Lambert, C.; Nakai, K.; et al. PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res.* **2003**, *31*, 3613–3617. [CrossRef]

57. Gardy, J.L.; Laird, M.R.; Brinkman, F.S.L.; Chen, F.; Rey, S.; Walsh, C.J.; Ester, M. PSORTb v.2.0: Expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics* **2005**, *21*, 617–623. [CrossRef]

58. Yu, N.Y.; Wagner, J.R.; Laird, M.R.; Melli, G.; Rey, S.; Lo, R.; Dao, P.; Sahinalp, S.C.; Ester, M.; Foster, L.J.; et al. PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* **2010**, *26*, 1608–1615. [CrossRef]

59. Horton, P.; Park, K.-J.; Obayashi, T.; Fujita, N.; Harada, H.; Adams-Collier, C.; Nakai, K. WoLF PSORT: Protein localization predictor. *Nucleic Acids Res.* **2007**, *35*, W585–W587. [CrossRef]

60. Savojardo, C.; Martelli, P.L.; Fariselli, P.; Profiti, G.; Casadio, R. BUSCA: An integrative web server to predict subcellular localization of proteins. *Nucleic Acids Res.* **2018**, *46*, W459–W466. [CrossRef]

61. Langlois, R.; Frank, J. A clarification of the terms used in comparing semi-automated particle selection algorithms in Cryo-EM. *J. Struct. Biol.* **2011**, *175*, 348–352. [CrossRef]

62. Wang, H.; Zheng, H. Model Cross-Validation. In *Encyclopedia of Systems Biology*; Dubitzky, W., Wolkenhauer, O., Cho, K.H., Yokota, H., Eds.; Springer: New York, NY, USA, 2013.

63. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 1–13. [CrossRef] [PubMed]

64. Wang, H.; Zheng, H. Negative Predictive Value. In *Encyclopedia of Systems Biology*; Dubitzky, W., Wolkenhauer, O., Cho, K.H., Yokota, H., Eds.; Springer: New York, NY, USA, 2013.

65. Vapnik, V. *The Nature of Statical Learning Theory*; Springer: New York, NY, USA, 1995.

66. Yadav, A.K.; Singla, D. VacPred: Sequence-based prediction of plant vacuole proteins using machine-learning techniques. *J. Biosci.* **2020**, *45*, 1–9. [CrossRef]

67. Kong, Y.; Yu, T. A Deep Neural Network Model using Random Forest to Extract Feature Representation for Gene Expression Data Classification. *Sci. Rep.* **2018**, *8*, 1–9. [CrossRef] [PubMed]

68. Clayton, E.A.; Pujol, T.A.; McDonald, J.; Qiu, P. Leveraging TCGA gene expression data to build predictive models for cancer drug response. *BMC Bioinform.* **2020**, *21*, 1–11. [CrossRef] [PubMed]

69. Adnan, N.; Lei, C.; Ruan, J. Robust edge-based biomarker discovery improves prediction of breast cancer metastasis. *BMC Bioinform.* **2020**, *21*, 1–18. [CrossRef]

70. Guan, X.; Runger, G.; Liu, L. Dynamic incorporation of prior knowledge from multiple domains in biomarker discovery. *BMC Bioinform.* **2020**, *21*, 1–10. [CrossRef]

71. Guo, L.; Wang, Z.; Du, Y.; Mao, J.; Zhang, J.; Yu, Z.; Guo, J.; Zhao, J.; Zhou, H.; Wang, H.; et al. Random-forest algorithm based biomarkers in predicting prognosis in the patients with hepatocellular carcinoma. *Cancer Cell Int.* **2020**, *20*, 1–12. [CrossRef]

72. Tang, J.; Mou, M.; Wang, Y.; Luo, Y.; Zhu, F. MetaFS: Performance assessment of biomarker discovery in metaproteomics. *Brief. Bioinform.* **2020**. [CrossRef]

73. Yu, X.; Lai, S.; Chen, H.; Chen, M. Protein–protein interaction network with machine learning models and multiomics data reveal potential neurodegenerative disease-related proteins. *Hum. Mol. Genet.* **2020**, *29*, 1378–1387. [CrossRef]

74. Niu, B.; Liang, C.; Lu, Y.; Zhao, M.; Chen, Q.; Zhang, Y.; Zheng, L.; Chou, K.-C. Glioma stages prediction based on machine learning algorithm combined with protein-protein interaction networks. *Genomics* **2020**, *112*, 837–847. [CrossRef]

75. Su, R.; Liu, X.; Wei, L.; Zou, Q. Deep-Resp-Forest: A deep forest model to predict anti-cancer drug response. *Methods* **2019**, *166*, 91–102. [CrossRef] [PubMed]

76. Shi, H.; Liu, S.; Chen, J.; Li, X.; Ma, Q.; Yu, B. Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure. *Genomics* **2019**, *111*, 1839–1852. [CrossRef] [PubMed]

77. Lind, A.P.; Anderson, P.C. Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties. *PLoS ONE* **2019**, *14*, e0219774. [CrossRef] [PubMed]

78. Kim, Y.; Kim, D.; Cao, B.; Carvajal, R.; Kim, M. PDXGEM: Patient-derived tumor xenograft-based gene expression model for predicting clinical response to anticancer therapy in cancer patients. *BMC Bioinform.* **2020**, *21*, 1–21. [CrossRef] [PubMed]

79. Long, W.; Yang, Y.; Shen, H.-B. ImPLoc: A multi-instance deep learning model for the prediction of protein subcellular localization based on immunohistochemistry images. *Bioinformatics* **2020**, *36*, 2244–2250. [CrossRef] [PubMed]

80. Jaganathan, K.; Panagiotopoulou, S.K.; McRae, J.F.; Darbandi, S.F.; Knowles, D.; Li, Y.I.; Kosmicki, J.A.; Arbelaez, J.; Cui, W.; Schwartz, G.B.; et al. Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **2019**, *176*, 535–548.e24. [CrossRef]

81. Zhang, Z.; Pan, Z.; Ying, Y.; Xie, Z.; Adhikari, S.; Phillips, J.; Carstens, R.P.; Black, D.L.; Wu, Y.; Xing, Y. Deep-learning augmented RNA-seq analysis of transcript splicing. *Nat. Methods* **2019**, *16*, 307–310. [CrossRef]

82. Guo, Y.; Li, W.; Wang, B.; Liu, H.; Zhou, D. DeepACLSTM: Deep asymmetric convolutional long short-term memory neural models for protein secondary structure prediction. *BMC Bioinform.* **2019**, *20*, 1–12. [CrossRef]

83. Wardah, W.; Khan, M.; Sharma, A.; Rashid, M.A. Protein secondary structure prediction using neural networks and deep learning: A review. *Comput. Biol. Chem.* **2019**, *81*, 1–8. [CrossRef]

84. Baptista, D.; Ferreira, P.G.; Rocha, M. Deep learning for drug response prediction in cancer. *Brief. Bioinform.* **2020**. [CrossRef]

85. You, J.; McLeod, R.D.; Hu, P. Predicting drug-target interaction network using deep learning model. *Comput. Biol. Chem.* **2019**, *80*, 90–101. [CrossRef] [PubMed]

86. Issa, N.T.; Stathias, V.; Schürer, S.; Dakshanamurthy, S. Machine and deep learning approaches for cancer drug repurposing. *Semin. Cancer Biol.* **2020**. [CrossRef] [PubMed]

87. Kuenzi, B.M.; Park, J.; Fong, S.H.; Sanchez, K.S.; Lee, J.; Kreisberg, J.F.; Ma, J.; Ideker, T. Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells. *Cancer Cell* **2020**, *38*, 672–684.e6. [CrossRef] [PubMed]

88. González, G.; Evans, C.L. Biomedical Image Processing with Containers and Deep Learning: An Automated Analysis Pipeline: Data architecture, artificial intelligence, automated processing, containerization, and clusters orchestration ease the transition from data acquisition to insights in medium-to-large datasets. *BioEssays* **2019**, *41*, 1900004. [CrossRef]

89. Jurtz, V.I.; Johansen, A.R.; Nielsen, M.; Armenteros, J.J.A.; Nielsen, H.; Sønderby, C.K.; Winther, O.; Sønderby, S.K. An introduction to deep learning on biological sequence data: Examples and solutions. *Bioinformatics* **2017**, *33*, 3685–3690. [CrossRef]

90. Ananda, M.M.; Hu, J. NetLoc: Network based protein localization prediction using protein-protein interaction and co-expression networks. In Proceedings of the 2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Hong Kong, China, 18–21 December 2010; Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, USA, 2010; pp. 142–148.

91. Mondal, A.M.; Lin, J.-R.; Hu, J. Network based subcellular localization prediction for multi-label proteins. In Proceedings of the 2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW), Atlanta, GA, USA, 12–15 November 2011; Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, USA, 2011; pp. 473–480.

92. Lee, K.; Chuang, H.-Y.; Beyer, A.; Sung, M.-K.; Huh, W.-K.; Lee, B.; Ideker, T. Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species. *Nucleic Acids Res.* **2008**, *36*, e136. [CrossRef]

93. Mintz-Oron, S.; Aharoni, A.; Ruppin, E.; Shlomi, T. Network-based prediction of metabolic enzymes' subcellular localization. *Bioinformatics* **2009**, *25*, i247–i1252. [CrossRef]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.