Review

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.)

Received: 24 November 2020; Accepted: 11 December 2020; Published: 14 December 2020

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.

Keywords: protein localization; signal peptide-based method; sequence compositional information-based method; machine learning based method; integrated method

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](image-url). Typical cell with different subcellular location and with available protein localization prediction tools.
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.

Figure 2. Number of sequences deposited and manually annotated proteins in the UniProt database in the last 10 years.
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 Markov 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>
<thead>
<tr>
<th>Method</th>
<th>Tools Used</th>
<th>Performance Matrix</th>
<th>Locations/Organism</th>
<th>Availability</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>SignalP-5.0</td>
<td>convolutional and recurrent (LSTM) neural networks</td>
<td>MCC, precision and recall</td>
<td>Archaea, Gram-positive Bacteria, Gram-negative Bacteria and Eukarya</td>
<td><a href="http://www.cbs.dtu.dk/services/SignalP/">http://www.cbs.dtu.dk/services/SignalP/</a></td>
<td>2019</td>
</tr>
<tr>
<td>TargetP 2.0</td>
<td>recurrent neural networks (RNNs) network</td>
<td>Precision, recall, F1-score, MCC</td>
<td>mitochondrial, chloroplastic, secretory pathway</td>
<td><a href="http://www.cbs.dtu.dk/services/TargetP/">http://www.cbs.dtu.dk/services/TargetP/</a></td>
<td>2019</td>
</tr>
<tr>
<td>SigUNet</td>
<td>Convolutional neural network</td>
<td>MCC, precision, recall, F1 measure</td>
<td>Eukaryotes, Gram-positive and Gram-negative bacteria</td>
<td><a href="https://github.com/mbilab/SigUNet">https://github.com/mbilab/SigUNet</a></td>
<td>2019</td>
</tr>
<tr>
<td>DeepSig</td>
<td>Convolutional Neural Networks</td>
<td>MCC, False Positive Rate, precision and recall</td>
<td>Eukaryotes, Gram-positive bacteria and Gram-negative bacteria</td>
<td><a href="https://deepsig.biocomp.unibo.it">https://deepsig.biocomp.unibo.it</a></td>
<td>2018</td>
</tr>
<tr>
<td>SChloro</td>
<td>SVM</td>
<td>Accuracy, Recall, Precision, F1-score, and MCC</td>
<td>six chloroplastic sub-compartments</td>
<td><a href="http://schloro.biocomp.unibo.it">http://schloro.biocomp.unibo.it</a></td>
<td>2017</td>
</tr>
<tr>
<td>PredSL</td>
<td>combination of neural networks, Markov chains, scoring matrices (PrediSi), and HMMs,</td>
<td>Accuracy</td>
<td>Eukaryotic subcellular location</td>
<td><a href="http://bioinformatics.biol.uoa.gr/PredSL/">http://bioinformatics.biol.uoa.gr/PredSL/</a></td>
<td>2006</td>
</tr>
<tr>
<td>TatP</td>
<td>HMM/artificial neural networks.</td>
<td>S-score and the C-score, Y-score, D-score</td>
<td>bacteria</td>
<td><a href="http://www.cbs.dtu.dk/services/TatP/">http://www.cbs.dtu.dk/services/TatP/</a></td>
<td>2005</td>
</tr>
</tbody>
</table>
Table 1. Cont.

<table>
<thead>
<tr>
<th>Method</th>
<th>Tools Used</th>
<th>Performance Matrix</th>
<th>Locations/Organism</th>
<th>Availability</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChloroP</td>
<td>Neural network</td>
<td>MCC, sensitivity, specificity</td>
<td>chloroplast transit peptides</td>
<td><a href="http://www.cbs.dtu.dk/services/ChloroP/">http://www.cbs.dtu.dk/services/ChloroP/</a></td>
<td>1999</td>
</tr>
</tbody>
</table>

* 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 archaeanal 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 TTPred), GPI-anchoeees (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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Tools Used</th>
<th>Performance Matrix</th>
<th>Feature Based</th>
<th>Locations/Organism</th>
<th>Availability</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepPred-SubMito</td>
<td>Convolutional neural network</td>
<td>Accuracy, MCC</td>
<td>Sequence information</td>
<td>Mitochondrial and submitochondrial proteins</td>
<td><a href="https://github.com/jinyinping/DeepPred-SubMito.git">https://github.com/jinyinping/DeepPred-SubMito.git</a></td>
<td>2020</td>
</tr>
<tr>
<td>mRNALoc</td>
<td>SVM</td>
<td>Sensitivity, Specificity, Accuracy, MCC</td>
<td>Sequence information</td>
<td>eukaryotic</td>
<td><a href="http://proteininformatics.org/mkumar/mrnaloc">http://proteininformatics.org/mkumar/mrnaloc</a></td>
<td>2020</td>
</tr>
<tr>
<td>SCLpred-EMS</td>
<td>Convolutional neural network</td>
<td>Sensitivity, Specificity, False positive rate, MCC</td>
<td>Sequence information</td>
<td>endomembrane system and secretory pathway</td>
<td><a href="http://distilldeep.ucd.ie/SCLpred2/">http://distilldeep.ucd.ie/SCLpred2/</a></td>
<td>2020</td>
</tr>
<tr>
<td>BUSCA</td>
<td>Integrated method of DeepSig, T5pred3, PredGPI, BetAware and ENSEMBLE3.0</td>
<td>Precision, recall, F1-score, MCC</td>
<td>Sequence information, signal and transit peptides, glycoprophosphatidylinositol (GPI) anchors and transmembrane domains</td>
<td>Gram-positive, gram-negative, fungi, plant, animal</td>
<td><a href="http://busca.biocomp.unibo.it">http://busca.biocomp.unibo.it</a></td>
<td>2018</td>
</tr>
<tr>
<td>SubMitoPred</td>
<td>SVM</td>
<td>Sensitivity, Specificity, Accuracy, MCC</td>
<td>Sequence and domain information</td>
<td>Mitochondrial and submitochondrial proteins</td>
<td><a href="http://proteininformatics.org/mkumar/submitopred/">http://proteininformatics.org/mkumar/submitopred/</a></td>
<td>2018</td>
</tr>
<tr>
<td>pLoc-mEuk</td>
<td>ML-GKR (multi-label Gaussian kernel regression) classifier</td>
<td>Coverage, Accuracy, Absolute true, Absolute false</td>
<td>Gene Ontology and Chou’s general PseAAC</td>
<td>22 different subcellular localizations of eukaryotic proteins</td>
<td><a href="http://www.jcibiinfo.cn/pLoc-mEuk/">http://www.jcibiinfo.cn/pLoc-mEuk/</a></td>
<td>2018</td>
</tr>
<tr>
<td>ERPred</td>
<td>SVM</td>
<td>Sensitivity, Specificity, Accuracy, MCC</td>
<td>Sequence information</td>
<td>ER Proteins</td>
<td><a href="http://proteininformatics.org/mkumar/erpred/index.html">http://proteininformatics.org/mkumar/erpred/index.html</a></td>
<td>2017</td>
</tr>
</tbody>
</table>
### Table 2. Cont.

<table>
<thead>
<tr>
<th>Method</th>
<th>Tools Used</th>
<th>Performance Matrix</th>
<th>Feature Based</th>
<th>Locations/Organism</th>
<th>Availability</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>SubNucPred</td>
<td>SVM</td>
<td>Sensitivity, Specificity, Accuracy, MCC</td>
<td>Sequence and domain information</td>
<td>Nuclear and subnuclear protein</td>
<td><a href="http://proteininformatics.org/mkumar/subnucpred/">http://proteininformatics.org/mkumar/subnucpred/</a></td>
<td>2014</td>
</tr>
<tr>
<td>LocTree3</td>
<td>SVM and homology</td>
<td>Accuracy, recall, standard deviation, standard error</td>
<td>Homology-based, Gene Ontology</td>
<td>18 classes for eukaryotes, in six for bacteria and in three for archaea</td>
<td><a href="http://www.rostlab.org/services/loctree3">http://www.rostlab.org/services/loctree3</a></td>
<td>2014</td>
</tr>
<tr>
<td>PlantLoc</td>
<td>localization motif search</td>
<td>accuracy</td>
<td>localization motif information</td>
<td>11 different location of plant proteins</td>
<td><a href="http://cal.tongji.edu.cn/PlantLoc/">http://cal.tongji.edu.cn/PlantLoc/</a></td>
<td>2013</td>
</tr>
<tr>
<td>MARSpred</td>
<td>SVM</td>
<td>Sensitivity, specificity, Accuracy, MCC</td>
<td>Sequence information, PSSM</td>
<td>cytosolic and mitochondrial aminoacyl tRNA synthetase</td>
<td><a href="http://www.imtech.res.in/raghava/marspred/">http://www.imtech.res.in/raghava/marspred/</a></td>
<td>2012</td>
</tr>
<tr>
<td>SCLPred</td>
<td>Neural Network</td>
<td>Sensitivity, specificity, False positive rate, MCC</td>
<td>primary sequence and multiple sequence alignments</td>
<td>four classes for animals and fungi and five classes for plants</td>
<td><a href="http://distill.ucd.ie/distill/">http://distill.ucd.ie/distill/</a></td>
<td>2011</td>
</tr>
<tr>
<td>AtSubP</td>
<td>SVM</td>
<td>Sensitivity, specificity, error rate, MCC, ROC curve</td>
<td>Sequence information, PSSM</td>
<td>subcellular localization of Arabidopsis</td>
<td><a href="http://bioinfo3.noble.org/AtSubP">http://bioinfo3.noble.org/AtSubP</a></td>
<td>2010</td>
</tr>
</tbody>
</table>
Table 2. Cont.

<table>
<thead>
<tr>
<th>Method</th>
<th>Tools Used</th>
<th>Performance Matrix</th>
<th>Feature Based</th>
<th>Locations/Organism</th>
<th>Availability</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euk-mPLoc 2.0</td>
<td>OET-KNN (Optimized Evidence-Theoretic K-Nearest Neighbor) classifiers</td>
<td>accuracy</td>
<td>gene ontology information, functional domain information, and sequential evolutionary information</td>
<td>eukaryotic proteins among the following 22 locations</td>
<td><a href="http://www.csbio.sjtu.edu.cn/bioinf/euk-multi-2/">http://www.csbio.sjtu.edu.cn/bioinf/euk-multi-2/</a></td>
<td>2010</td>
</tr>
<tr>
<td>PSORTb</td>
<td>SVM</td>
<td>Precision, recall, accuracy, MCC</td>
<td>Sequence information</td>
<td>Different subcellular location of Gram-negative, Gram-positive, archaea</td>
<td><a href="http://www.psort.org/psortb">http://www.psort.org/psortb</a></td>
<td>2010</td>
</tr>
<tr>
<td>YLoc</td>
<td>naïve Bayes alongside entropy-based discretization</td>
<td>overall accuracy, F1-score</td>
<td>Sequences information, GO-term and motif</td>
<td>animal, fungal and plant proteins</td>
<td><a href="http://www.multiloc.org/YLoc">www.multiloc.org/YLoc</a></td>
<td>2010</td>
</tr>
<tr>
<td>MultiLoc2</td>
<td>SVM</td>
<td>Sensitivity, specificity, Accuracy, MCC</td>
<td>phylogenetic profiles and gene ontology terms</td>
<td>Plant, Animal, Fungal</td>
<td><a href="https://abi-services.informatik.uni-tuebingen.de/multiloc2/webloc.cgi">https://abi-services.informatik.uni-tuebingen.de/multiloc2/webloc.cgi</a></td>
<td>2009</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Tools Used</th>
<th>Performance Matrix</th>
<th>Feature Based</th>
<th>Locations/Organism</th>
<th>Availability</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>SherLoc</td>
<td>SVM</td>
<td>Sensitivity, specificity, MCC</td>
<td>Sequence information</td>
<td>eukaryotic proteins</td>
<td><a href="http://www-bs.informatik.uni-tuebingen.de/Services/SherLoc">http://www-bs.informatik.uni-tuebingen.de/Services/SherLoc</a></td>
<td>2007</td>
</tr>
<tr>
<td>MitPred</td>
<td>SVM</td>
<td>Sensitivity, specificity, Accuracy, MCC</td>
<td>Sequence information</td>
<td>Mitochondrial proteins</td>
<td><a href="http://www.imtech.res.in/raghava/mitpred/">http://www.imtech.res.in/raghava/mitpred/</a></td>
<td>2006</td>
</tr>
<tr>
<td>HSLpred</td>
<td>SVM</td>
<td>Accuracy, MCC, Reliability index</td>
<td>Sequence information</td>
<td>Human Protein</td>
<td><a href="http://www.imtech.res.in/raghava/hslpred/">http://www.imtech.res.in/raghava/hslpred/</a></td>
<td>2005</td>
</tr>
<tr>
<td>PSLpred</td>
<td>SVM</td>
<td>Accuracy, MCC, Reliability index</td>
<td>Sequence information</td>
<td>gram-negative bacterial proteins</td>
<td><a href="http://www.imtech.res.in/raghava/pslpred/">http://www.imtech.res.in/raghava/pslpred/</a></td>
<td>2005</td>
</tr>
<tr>
<td>ESLpred</td>
<td>SVM</td>
<td>Accuracy, MCC, Reliability index</td>
<td>Sequence information and PSSM</td>
<td>eukaryotic proteins</td>
<td><a href="http://www.imtech.res.in/raghava/eslpred/">http://www.imtech.res.in/raghava/eslpred/</a></td>
<td>2004</td>
</tr>
</tbody>
</table>
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).

![Flowchart of the development of the machine learning algorithm for the prediction of protein localization.](image)

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

<table>
<thead>
<tr>
<th>Metric</th>
<th>Explanation</th>
<th>Formula</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>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.</td>
<td>$\frac{TP + TN}{TP + FP + TN + FN}$</td>
<td>[61,62]</td>
</tr>
<tr>
<td>Specificity/True Negative Rate</td>
<td>Proportion of negatives that are correctly identified</td>
<td>$\frac{TN}{TN + FP}$</td>
<td>[61,62]</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Metric</th>
<th>Explanation</th>
<th>Formula</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision/Positive Predictive Value</td>
<td>It tells us how many of the correctly predicted cases actually turned out to be positive.</td>
<td>TP/(TP + FP)</td>
<td>[61]</td>
</tr>
<tr>
<td>Recall/Sensitivity/True Positive Rate</td>
<td>It tells us how many of the actual positive cases we were able to predict correctly with our model.</td>
<td>TP/(TP + FN)</td>
<td>[61,62]</td>
</tr>
<tr>
<td>F1-Score</td>
<td>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.</td>
<td>$F_1 = 2 \cdot \text{precision} \cdot \text{Recall}/\text{Precision} + \text{Recall}$</td>
<td>[63]</td>
</tr>
<tr>
<td>Negative Predictive Value</td>
<td>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.</td>
<td>TN/(TN + FN)</td>
<td>[64]</td>
</tr>
<tr>
<td>False Positive Rate</td>
<td>Number of incorrect positive predictions divided by the total number of negatives.</td>
<td>FP/(FP + TN)</td>
<td>[61]</td>
</tr>
</tbody>
</table>

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 feedforward 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.

Supplementary Materials: The following are available online at http://www.mdpi.com/2075-1729/10/12/347/s1,
Table S1: Comparative performance of available subcellular localization methods.

Author Contributions: Conceptualization, R.K. and S.K.D.; Data Curation, R.K.; Formal Analysis, R.K.; Writing—Original Draft Preparation, R.K.; Writing—Review and Editing, R.K. and S.K.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding, and the APC was waived off by the Journal.

Acknowledgments: Authors acknowledge the support from the Life journal in organizing a special issue on this topic. We gratefully acknowledge Rajrani Ruhel for critically reading the manuscript. No funding is received to write and compile this review article.

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

References


43. Thomas, 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.; Myymyk, 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]
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]
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]


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]


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]


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]


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

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).