Computational and Structural Biotechnology Journal 19 (2021) 5834–5844

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Computational methods for protein localization prediction

Yuexu Jiang, Duolin Wang, Weiwei Wang, Dong Xu *

Department of Electrical Engineering and Computer Science, Bond Life Sciences Center, University of Missouri, Columbia, MO, USA

ARTICLE INFO

Article history:
Received 31 March 2021
Received in revised form 12 October 2021
Accepted 13 October 2021
Available online 19 October 2021

Keywords:
Protein localization prediction
Computational methods
Review

ABSTRACT

The accurate annotation of protein localization is crucial in understanding protein function in tandem with a broad range of applications such as pathological analysis and drug design. Since most proteins do not have experimentally-determined localization information, the computational prediction of protein localization has been an active research area for more than two decades. In particular, recent machine-learning advancements have fueled the development of new methods in protein localization prediction. In this review paper, we first categorize the main features and algorithms used for protein localization prediction. Then, we summarize a list of protein localization prediction tools in terms of their coverage, characteristics, and accessibility to help users find suitable tools based on their needs. Next, we evaluate some of these tools on a benchmark dataset. Finally, we provide an outlook on the future exploration of protein localization methods.

Contents

* Corresponding author.
E-mail address: xudong@missouri.edu (D. Xu).

https://doi.org/10.1016/j.csbj.2021.10.023

1. Introduction

Cells contain well-organized compartments with different protein constituents. Although most proteins are synthesized in the cytosol, about half of them are transported into or across at least one cellular membrane to reach their functional destination [1–3]. The aberrant localization of proteins usually has harmful effects, including diseases in humans and animals and poor traits in plants [4–7]. Hence, studying the mechanism of protein localization is essential in a broad range of applications, such as plant breeding, pathological analysis, and the therapeutic modification of disease-related protein mislocalization [5,8]. Protein localization is a complicated biological process controlled by many factors, such as signal peptides, protein trafficking, protein–protein interactions, folding, and alternative splicing [5,9]. Among these, protein localization guided by targeting peptides is the most common mechanism [10] and includes pre-sequences and internal signals [11,12]. Pre-sequences are found at the N- or C-terminus of protein sequences with enrichment of charged or hydrophobic amino acids, while internal signals are located in the middle of a sequence. How precursor proteins are directed to their target organelles is only partially understood [11], and only a small number of targeting peptides (particularly internal signals) have been experimentally identified. According to UniProt annotation (release 2020_05), out of the reviewed 20,394 human proteins, 7348 (36.0%) proteins have localization annotation with experimental verification, while only 3608 (17.7%) proteins have known targeting peptides. Furthermore, limited sub-organelle compartment localization data are available. According to a recent search that we conducted on 16,213 human proteins in ten human organelles, 5882 (36.3%) proteins had experimentally verified organellar localization annotation, while only 3518 (21.7%) proteins had experimentally verified sub-organellar localization annotation. Targeting peptide and sub-organelle data for non-human species are even sparser.

Several experimental methods can be used for protein localization analysis. Quantitative mass spectrometric readouts allow for the identification of proteins across fractions [13–16]. Spatially and temporally resolved proteomic maps in living cells can be obtained by targetable peroxidase [17–19]. Techniques such as immunofluorescence and high-resolution confocal microscopy have enabled the visual estimation of protein localization within a single cell [20–24]. One problem with experimental methods is that their throughput is relatively low. In addition, experimental protein localization identification requires a great deal of time and resources. Importantly, experimental and computational protein localization identification approaches are complementary to each other. Experimental annotations are typically used as true labels for computational methods. Computational models are trained using these ground truth data to predict the localization of other proteins. Due to their cost-effective, automated, and high-throughput nature, computational methods are helpful for the large-scale characterization of protein subcellular locations.

Several papers have reviewed protein localization prediction methods. The review of [25] focuses on methods for bacterial protein localization prediction. Other reviews [26,27] mainly cover protein sequence features (such as targeting peptides) in localization prediction. The methods reviewed in [28] predict protein function taxonomies, such as the Functional Catalogue, Enzyme Commission, or Gene Ontology, rather than specific cellular components. Another review mainly discusses web-based prediction tools for human protein subcellular localization [29]. General methods and tools for protein localization prediction are introduced in the reviews of [30–33], which have a scope similar to ours. However, the most recent review in the literature [30–32] was published in 2014. Many new methods have been proposed since then that have greatly improved prediction accuracy, especially deep-learning methods. This review focuses on these new methods and tools in addition to previous representative methods. A less detailed review [33] was recently published. Compared to [33], this review separates the introduction of features, algorithms, and tools in greater detail so readers can better understand their relationships. Additionally, the applicability of the tools is considered, and only actively maintained tools are listed. Users can select the tools they need based on the information summarized and access them through the links provided. All the aforementioned features make this review unique and valuable. This review is organized as follows. In Sections 2 and 3, we analyze the features and classifiers that are often associated with different methods, respectively. Many of these methods provide standalone tools and/or web services that we summarize in Section 4. For each tool, information of target compartments, used algorithm, accessibility, etc. is given. In Section 5, a summary is provided together with promising directions for future protein localization prediction methods. The relationship of the data, features, and models used in computational protein localization prediction, as well as their outputs, are shown in Fig. 1. The features and main contributions of this review are summarized as follows:

- A systematic introduction of features, algorithms, methods, and tools, as well as their relationships related to protein localization.
- A comprehensive list of available protein localization prediction tools, many of which became available in recent years.
- Extensive evaluations of localization prediction tools/methods, providing insights on why some methods have better prediction performance than others.
- Significant discussion on the future direction of protein localization studies.

2. Data and features

2.1. Sequence-based features

Protein sequences are considered the most essential source of information for protein localization prediction, particularly terminal region sequences where targeting signals are likely to be found. Protein sequence information can be obtained from databases such as UniProt [34]. In addition, many types of features have been proposed based on protein sequences.

2.1.1. Amino acid composition

The simplest feature representing a protein sequence is likely amino acid (AA) composition [35]. Given a protein sequence P, the AA composition of P can be expressed by

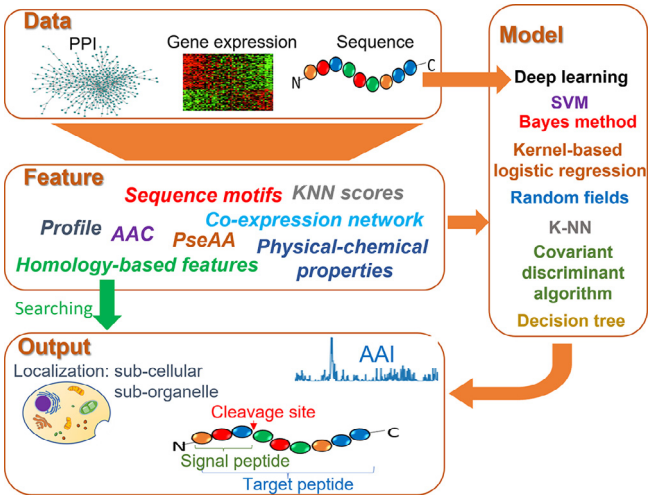

Fig. 1. Relationships among the data, features, models, and prediction outputs in the computational prediction of protein localization. Sequence data can be converted into different features before feeding the data to a classifier model. Some classification models take raw data (e.g., one-hot-encoding of protein sequences for deep learning) as input, while others use engineered features. Localization prediction (at the sub-cellular and/or suborganellar level) is the most common output. Some methods also provide side product predictions such as target peptides, signal peptide cleavage sites, and mechanism interpretability at amino-acid-level resolution (AAI). Homology-based methods are special in the sense that they can make predictions directly based on homology-based features, such as the GO terms of homologous proteins.

$$P = [f_1 f_2 \cdots f_{20}]^T, \tag{1}$$

where $f_u (u = 1, 2, \cdots, 20)$ are the normalized occurrence frequencies of the 20 native amino acids in protein P.

2.1.2. PseAA composition

The main shortcoming of using AA composition as a feature is its lack of protein sequence order information [31]. The concept of pseudo amino acid composition (PseAA) was proposed to address this problem [36] by representing a protein as a vector P:

$$P = [p_1 p_2 \cdots p_{20} p_{20+1} \cdots p_{20+\lambda}]^T, (\lambda < L \tag{2}$$

where the $20 + \lambda$ components are given by

$$P_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + w\sum_{k=1}^{\lambda} \tau_k}, & (1 \leqslant u \leqslant 20) \\ \dfrac{w\tau_{u-20}}{\sum_{i=1}^{20} f_i + w\sum_{k=1}^{\lambda} \tau_k}, & (20+1 \leqslant u \leqslant 20+\lambda) \end{cases} \tag{3}$$

where w is a weight factor set to 0.05 in the original paper [36], and τ_k is the k-th tier correlation factor, which reflects the sequence order correlation between all of the k-th most contiguous residues as formulated by

$$\tau_k = \frac{1}{L-k} \sum_{i=1}^{L-k} J_{i,i+k}, \ (K < L) \tag{4}$$

As in Eq. (2), the first 20 components are associated with the conventional amino acid composition of P, whereas the remaining components are the λ correlation factors that reflect the first tier, second tier, and so on up to the λ-th tier sequence order correlation patterns. These λfactors incorporate sequence order effects, and λ is a chosen hyperparameter (integer). The calculation of τ_k integrates the hydrophobicity values (H_1), hydrophilicity values (H_2), and side-chain masses (M) for amino acids i and $i + k$ as

$$J_{i,i+k} = \frac{1}{3}\left\{[H_1(R_{i+k}) - H_1(R_i)]^2 + [H_2(R_{i+k}) - H_2(R_i)]^2 + [M(R_{i+k}) - M(R_i)]^2\right\} \tag{5}$$

Note that Eq. (5) is just one form for deriving the correlation factors. Other information, such as physicochemical distance and amphiphilic patterns, can also derive different types of PseAA composition.

2.1.3. Homology information

As subcellular localization tends to be evolutionarily conserved [37], homology to a protein of known localization is often a good indicator of actual protein localization [38]. Such information can be derived via BLAST [39] or a more sensitive search method such as HHblits [40] against a database of proteins with known localization. One important source of known localization is the cellular component of Gene Ontology (GO) [41], which has been used to improve protein localization prediction performance [42–45]. Homology information can also be obtained through protein structure similarity, as did in C-I-Tasser [46], a template-based method for protein structure and function prediction. In C-I-Tasser, the function prediction of a query protein is obtained by matching its structural model with proteins in the BioLiP function library via structure and sequence profile comparisons. Each entry in BioLiP contains GO terms so that the GO cellular localization of the query protein can be inferred.

2.1.4. Evolutionary profiles

Evolutionary profiles, represented by Position-Specific Scoring Matrices (PSSMs), etc., provide informative input for protein localization prediction. PSSMs indicate the amino acid occurrence for each position in a protein multiple sequence alignment. PSSM scores are generally given as positive or negative values. A positive score means that the given amino acid substitution occurs more frequently in the alignment than expected by chance, while a negative score indicates that the substitution occurs less frequently than expected by chance. PSSMs can be created using PSI-BLAST, which finds similar protein sequences to a query sequence and then constructs a PSSM from the resulting alignment.

The BLOSUM (BLOcks SUbstitution Matrix) matrix [47] is a substitution matrix used for scoring alignments between evolutionarily divergent protein sequences. Several BLOSUM matrices exist using different alignment databases, which are named with sequence identity thresholds in the alignments. For example, BLOSUM62 is a matrix built using sequences with less than 62% similarity (sequences with ≥62% identity were clustered). BLOSUM62 is the default matrix for protein BLAST and is among the best for detecting weak protein similarities. Encoding with BLOSUM matrices is fast and provides a viable alternative if acquiring a PSSM is slow or unsuccessful [48,49].

One particular usage of a sequence profile is as the profile kernel of an SVM. A key feature of the SVM optimization problem is that it depends only on the inner products of the feature vectors representing the input data. Several kernel functions have been proposed to avoid the explicit transformation of input data to feature vectors, explained as follows. Let Φ represent a mapping from the input space of protein sequences into a (possibly high-dimensional) vector space called the feature space. A string kernel is defined by $K(x, y) = \langle \Phi(x), \Phi(y) \rangle$, where x and y are sequences, e.g., $x = x_1 x_2 \cdots x_N$ from the alphabet Σ of amino acids ($|\Sigma| = 20$, and the length $N = |x|$ depends on the sequence). Let $P(x) = \{p_i(a), a \in \Sigma\}_{i=1}^{N}$ represent a profile for sequence x, with $p_i(a)$ denoting the emission probability of amino acid a in position i and $\sum_{a \in \Sigma} p_i(a) = 1$ for each position i; a profile kernel is defined as $K(P(x), P(y))$. The Fisher-SVM method [50] is a profile-kernel method that represents each protein sequence as a vector of Fisher scores extracted from a profile Hidden Markov Model (HMM) for a protein family. Kuang *et al.* proposed profile-based string kernels that use probabilistic profiles, such as those produced by the PSI-

BLAST algorithm, to define position-dependent mutation neighborhoods along with protein sequences for inexact matching of k-length subsequences ("k-mers") [51]. Such profile kernels are used in LocTree2 [52], an SVM-based method for protein localization prediction.

2.1.5. Motifs

Certain sequence patterns may correlate with a specific subcellular localization due to localization signals or functional relationships [53]. This motif information can be retrieved from databases such as PROSITE [54] or by data mining. One special type of motifs represents targeting peptides, i.e., short sequences mainly present at protein termini that function like a postal code to specify an intracellular or extracellular destination [55]. Some methods predict the presence of targeting peptides as a side product in tandem with protein localization prediction [56,57], while other methods use targeting peptides as input features to predict protein localization [53].

A sequence pattern can also be extracted through a sliding window of a k-mer sequence. The motif length *k* is often set based on specific needs or prior biological knowledge. For example, TetraMito [58] uses over-represented tetrapeptides (four continuous amino acids believed to encode a particular structure) as features to predict submitochondrial protein localization. A similar idea is used for sub-Golgi protein localization prediction by SubGolgi 2.0 [59], which uses an SVM classifier trained with g-gap dipeptide compositions (two amino acids with g residues between them). LOCALIZER [60] is another k-mer-based method for predicting plant and effector protein localization to chloroplasts, mitochondria, and nuclei. The motif length *k* varies in LOCALIZER to capture the target signals on protein sequences.

2.1.6. Physical–chemical properties

As the name suggests, this feature uses AAs' physical and chemical properties to represent protein sequences. These previously calculated properties are stored in public databases. According to Venkatarajan and Braun [61], a comprehensive list of 237 physical–chemical properties of each amino acid was compiled from the SWISS-PROT [34] and dbGET [62] databases. They showed that the number of properties could be reduced while retaining approximately the same distribution of amino acids in the feature space. Notably, the correlation coefficient between the original and regenerated distances was more than 99% using the first five eigenvectors.

2.1.7. Pre-train sequence embedding

Evolutionary information significantly benefits model prediction performance; however, as the number of proteins in databases increases, retrieving such information is often time-consuming. Additionally, evolutionary information is less powerful for small protein families, e.g., for proteins from the Dark Proteome [63]. One promising sequence embedding method uses the pre-train model adopted from Natural Language Processing (NLP). The pre-train model utilizes large, unlabeled text-corpora such as Wikipedia to conceptualize syntax and semantics. Pre-train methods such as Transformer [64], ELMo [65], Word2Vec [66], and Bert [67] employ self-learning and predict either the next word in a sentence given all previous words, the current word from a window of surrounding context words (or using the current word to predict the surrounding window of context words), or masked-out words given all unmasked words. Once trained, language models can extract features, referred to as embeddings, to use as input for subsequent supervised learning (transfer-learning). A similar strategy has been used for protein sequence embedding. SeqVec [68] uses ELMo on UniRef50 for pre-train embedding and transfer-learning for subcellular localization prediction. ProtTrans [69] employs different pre-training embedding models on UniRef and BFD data containing 2.1 billion protein sequences, which can also be used for protein localization prediction. In addition, a recent study showed that the pre-training embedding from language models followed by an attention-based deep-learning architecture could yield excellent performance in protein localization prediction even without using evolutionary information [70].

2.2. Protein interactions

If two proteins interact, they are neighbors of each other in a protein–protein-interaction (PPI) network. The localizations of the neighbors in a PPI network carry information about the localization of un-annotated proteins. For example, if the majority of a protein's neighbors share the same localization, the protein is likely localized to the same location. The definition of protein interaction varies and can be based on physical connections or genetic regulations. Protein interaction data can be retrieved from databases such as MINT [71], DIP [72], BioGRID [73], and STRING [74].

2.3. Gene/protein expression

The rationale for using gene/protein expression as a feature is that genes/proteins in the same compartment at the organelle or suborganelle level tend to be co-expressed to perform related functions. Gene/protein expression information can be used in network form like the aforementioned protein interaction feature [75]. For example, an interaction is established if the expression correlation between two genes/proteins exceeds a predefined threshold. Gene/protein expression information can also be used to create features such as the k-nearest-neighbor (k-NN) scores in the MU-LOC method [76] or used as standalone features in the SLocX method [77]. Gene/protein expression data are widely available and can be downloaded from databases like the Gene Expression Omnibus (GEO) [78] and The Cancer Genome Atlas (TCGA) [79].

3. Classification algorithms

3.1. Support vector machine

Support vector machines (SVMs) [80] use kernel functions to map input vectors into high dimensional feature space and construct a hyperplane that maximizes the margin between different classes. SVMs can handle large feature spaces and effectively avoid overfitting.

The method proposed in [81] is an early SVM-based protein localization prediction approach. To deal with a multi-class classification problem, it uses AA composition as a feature to train SVM classifiers in a one-versus-rest fashion. pSLIP [82] employs the SVM method in conjunction with multiple physicochemical properties of amino acids to predict protein subcellular localization in eukaryotes across six different localizations. The Density-induced Support Vector Data Description (D-SVDD) is an extension of Conventional Support Vector Data Description (C-SVDD) that was introduced for a one-class classification task inspired by SVMs [83]. PLPD [84] uses AA-based and motif features to modify the D-SVDD for multi-class multi-label protein localization prediction, mainly from imbalanced training datasets. A two-level SVM system to predict protein localization is described in [85]. The first level consists of multiple SVMs using distinct AA-based features (AA composition and physical–chemical properties), and the SVM at the second level makes the final prediction. SLocX [77] uses an SVM to predict the subcellular localization of Arabidopsis proteins using gene expression and AA composition as features.

Recent SVM-based methods include SubMitoPred [86], which uses Pfam domain information to predict mitochondrial proteins and their sub-mitochondrial localization. ERPred [87] predicts ER-

resident proteins by training an SVM with a combination of amino acid compositions from different parts of proteins. SubNucPred [88] predicts protein localization for 10 sub-nuclear locations sequentially by combining the presence or absence of a unique Pfam domain and an amino acid composition-based SVM model. CELLO2GO [89] combines an SVM-based localization prediction method with BLAST homology search. When homologous proteins with known localizations are available, their GO terms are used as possible functional annotations for a queried protein. Otherwise, the SVM classifier provides localization prediction. MultiP-SChlo [90] is another SVM-based method that predicts subchloroplast protein localization with multiple labels based on features such as PseAAC and AA properties. MKLoc [91] is an SVM-based method for multi-label protein localization prediction where protein sequences are represented by a 30-dimensional feature vector consisting of PseAAC, physical–chemical properties, motifs from PROSITE, and GO annotations. LocTree3 [42] improves upon LocTree2 [52] by including information about homologs, if available, through a PSI-BLAST search. MitoFates [92] is a prediction method for cleavable N-terminal mitochondrial targeting signals and their cleavage sites. Besides classical features such as AA composition, sequence profiles, and physical–chemical properties, MitoFates introduces novel sequence features, including positively charged amphiphilicity and presequence motifs, and trains an SVM classifier using these features. SChloro [93] converts a protein sequence into a PSSM profile and Kyte-Doolittle scale (average hydrophobicity). Two layers of SVMs are designed to predict targeting signal and membrane protein information. The final output predicts six sub-chloroplastic localizations by integrating the predictions from previous layers.

3.2. Probabilistic methods

3.2.1. Bayes method

Probabilistic models, specifically Bayesian methods such as the Bayes Optimal Classifier or Bayesian Networks, make the most probable prediction for a new example. Bayesian methods use the Bayes Theorem [94] for calculating a conditional probability. They are also closely related to the Maximum a Posteriori (MAP), a probabilistic framework that finds the most probable hypothesis for a training dataset. In large real-world applications, the Bayes method usually assumes that different features are independent of each other, known as Naïve Bayes.

PSORT-B [53] and subsequent versions of it [95,96] (with higher prediction coverage and refined subcategories), construct six analytical modules based on features including homology, motifs, and signal peptides. A query protein undergoes each of the six analyses and the results are combined using a Bayesian Network to generate a final probability value for each localization site.

3.2.2. Kernel-based logistic regression

When determining the probability of a protein to be localized at a specific location given a PPI network, kernel-based logistic regression (KLR) considers the localization information of all the proteins in the network. The KLR model can be formulated as follows [97]. Given a protein–protein interaction network with N proteins $X_1, \cdots, X_N$, some of which have unknown localization, let

$$X_{[-i]} = (X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_N) \tag{6}$$

represent the protein set excluding protein i. Let

$$M_L(i) = \sum_{j \neq i, x_j known} K(i,j) I\{x_j = L\} \tag{7}$$

represent the summed distances of protein i to proteins targeting localization L, where $K(i,j)$ is the kernel function for calculating the distances between two proteins in the network. Then, the KLR model is given by$(20+1 \leqslant u \leqslant 20+\lambda)$

$$log \frac{\Pr(X_i = L|X_{[-i]}, \theta)}{1 - \Pr(X_i = L|X_{[-i]}, \theta)} = \gamma + \delta M_{!L}(i) + \eta M_L(i) \tag{8}$$

which means that the logit of $\Pr(X_i = L|X_{[-i]}, \theta)$, and the probability of protein i targeting location L is linear based on the summed distances of proteins targeting L or another location. Then, we have

$$\Pr(X_i = L|X_{[-i]}, \theta) = \frac{1}{1 + e^{-(\gamma + \delta M_{!L}(i) + \eta M_L(i))}} \tag{9}$$

Note that the probability of being in each localization is calculated separately as a binary classification problem.

NetLoc [75] applies KLR to protein networks based on different relationships, including physical PPI, genetic PPI, and coexpression. In NetLoc, networks with high connectivity and a high percentage of interacting protein pairs targeting the same location lead to better prediction performance.

3.2.3. Random Fields

Given a probability space, a random field $T(x)$ defined in $\mathbb{R}^n$ is a function such that for every fixed $x \in \mathbb{R}^n$, $T(x)$ is a random variable on the probability space [98]. Markov Random Fields (MRFs) and Conditional Random Fields (CRFs) have been used for protein localization prediction [56,99]. An MRF of a graph G is a set of random variables corresponding to the nodes in G (random field) with a joint distribution that is Markov-constrained for G. In other words, the joint probability distribution associated with the MRF is subject to the Markov constraint given by G: for any two variables, V_i and V_j, the value of V_i is conditionally independent of V_j given its neighbors B_i. In this case, the joint probability distribution $P(\{V_i\})$ factorizes according to G. In contrast, we can describe a CRF for a graph G as a set of random variables corresponding to the nodes in G, a subset $\{X_i\}_{i=1}^n$ of which are assumed always to be observed, and remaining variables $\{Y_i\}_{i=1}^m$ with a conditional distribution $P(\{Y_i\}_{i=1}^m|\{X_i\}_{i=1}^n)$ that is Markov-constrained for G. Both MRFs and CRFs typically fit a model that can be used for conditional inference in diverse settings. The main difference is that an MRF has no consistently designated "observed variables" and requires a joint distribution over all variables that adhere to the Markov constraints of G.

CRFs are used for signal peptide cleavage site prediction in DeepSig [99] and specific signal peptide prediction in SignalP 5.0 [56]. A tissue-specific subcellular localization prediction method is proposed in [100] using multi-label MRF. A tissue-specific network was constructed from generic physical PPI networks and tissue-specific functional associations, and tissue-specific localization annotations were obtained from HPA [101].

3.3. Distance-based methods

3.3.1. k-nearest Neighbors (k-NN) classification

The k-NN algorithm is a nonparametric method used for classification and regression [102]. In both cases, the input consists of the k closest training examples in the data set. The output depends on whether the k-NN model is used for classification or regression. In k-NN classification, the output is class membership. An object is classified by a plurality vote of its neighbors, and assigned to the most common class of its k nearest neighbors (k is typically a small positive integer). If k = 1, then the object is simply assigned to the class of the single nearest neighbor.

WoLF PSORT [103] converts protein amino acid sequences into numerical localization features such as targeting signals, amino acid composition, and functional motifs. After conversion, a k-NN classifier is used for prediction. An idea similar to k-NN is used in [104], where a physical interaction network was obtained from BioGRID [73], and GO Cellular Component annotation was mapped onto the network, if available, for the corresponding protein

(node). For a query protein, the percentage of its interactors associated with each target localization is calculated. The top two localizations are then reported as the prediction.

3.3.2. Covariant discriminant algorithm based on Mahalanobis distance

The Mahalanobis distance [105] is a measure of the distance between a point P and a distribution D. It is essentially a multidimensional generalization to measure how many standard deviations away P is from the mean of D. This distance is zero if P is at the mean of D and grows as P moves away from the mean along each principal component axis. If each of these axes is re-scaled to have unit variance, then the Mahalanobis distance corresponds to the standard Euclidean distance in the transformed space. The Mahalanobis distance is thus unitless and scale-invariant and takes the correlations in a data set into account.

The Mahalanobis distance of an observation $\vec{x} = (x_1, x_2, x_3, \cdots, x_N)^T$ from a set of observations with mean $\vec{\mu} = (\mu_1, \mu_2, \mu_3, \cdots, \mu_N)^T$ and covariance matrix S is defined as:

$$D_M\left(\vec{x}\right) = \sqrt{\left(\vec{x} - \vec{\mu}\right)^T S^{-1} (\vec{x} - \vec{\mu})} \tag{10}$$

The similarity between standard vector $\bar{X}^{\xi}$ (normalized occurrence frequencies of the 20 AA from class ξ) and protein X is characterized by the covariant discriminant, as defined by Liu and Chou in [106]:

$$F\left(X, \bar{X}^{\xi}\right) = D^2\left(X, \bar{X}^{\xi}\right) + \ln(\lambda_2^{\xi}\lambda_3^{\xi}\lambda_4^{\xi}\cdots\lambda_{20}^{\xi}) \tag{11}$$

where the first term is the squared Mahalanobis distance, and λ_i^{ξ} is the i-th eigenvalue of covariance matrix S.

The covariant discriminant algorithm is used in general protein localization prediction in [106], as well as in apoptosis protein localization prediction [107] and Golgi protein subtype prediction [108]. The features used in these methods are AA composition or Pseudo AAC.

3.4. Neural network/deep learning

An artificial neural network (ANN) is based on a collection of connected units or nodes called artificial neurons that loosely model the neurons in a biological brain. Each connection, like the synapses in a biological brain, can transmit a signal to other neurons. Each artificial neuron receives a signal and processes it, and the output of each neuron is computed by a non-linear function of the sum of its inputs. Increased GPU computing power and distributed computing allow the use of larger networks, which is known as "deep learning" [109]. Deep learning has become the hottest field in machine learning, and different architectures have been proposed, such as deep neural networks (DNNs) [110], convolutional neural networks (CNNs) [109], recurrent neural networks (RNNs) [111,112], and attention mechanisms [113]. These deep learning methods, as well as traditional ANNs, have been applied in protein localization prediction. Due to the abstract feature extraction capability of deep learning models, artificial feature engineering is sometimes not required. Raw protein sequences can be given as inputs for many deep learning localization prediction methods [114,115]. Among different deep learning architectures, RNNs are inherently suitable for processing protein sequences. Notably, a widely-adopted implementation of RNN, Long Short-Term Memory (LSTM), captures long-distance dependencies well [116]. LSTMs have been successfully applied in machine translation [117–119] and speech recognition [120,121]. The methods used for these tasks can be applied to protein localization prediction by considering protein sequences as sentences and amino acids as words. CNNs are most commonly applied to analyze visual imagery [122]. A CNN uses shared-weight convolution kernels to slide along input features and provide feature maps for downstream calculations. The pooling operation reduces data dimension by combining the outputs of neuron clusters at one layer into a single neuron in the next layer. It is often desirable to apply CNNs to long protein sequences at the cost of losing single residue resolution for improved computational efficiency [48,49,123]. Moreover, CNN filters can be used to build position-weight matrices (PWMs) of sequence motifs, which can improve model interpretability [123]. The attention mechanism technique mimics cognitive attention [113] as it enhances the essential parts of input data and fades out the rest. This increases the signal-to-noise ratio and elucidates the contribution of features to the final prediction [48,124], e.g., determines which amino acids are responsible for protein localization.

Several neural network/deep learning-based methods have been proposed for protein localization prediction. SCLpred [114] is an N-to-1 neural network for protein localization prediction capable of mapping a whole sequence into fixed-length properties so that no predefined feature is needed. A similar method was later used in SCLpred-EMS [125] to predict proteins in the endomembrane system and secretory pathway. DeepLoc [49] applies the CNN method, bidirectional LSTM [112], and the attention mechanism for predicting localization and detecting the regions in a protein sequence that contribute to localization prediction. The length of the embedding is the same as the input sequence, while the attention weight of each amino acid is a combination of several CNN filters of different receptive fields. This reduces the interpretation resolution of the model. The researchers also apply different embedding methods and illustrate that PSSM achieves significantly better performance than BLOSUM62 at the cost of increased computing time. MU-LOC [76] provides two models (SVM and DNN) to predict mitochondrial protein in plants. The features used include AA composition, PSSM, and gene expression. MULocDeep [48], developed from the same group that developed MU-LOC, is a recently developed deep learning method that extends target localization coverage to 10 main subcellular compartments and their suborganellar compartments with 44 localization classes in total. Its deep learning model consists of a bidirectional LSTM and a multi-head self-attention mechanism [124]. In addition to protein localization prediction, it sheds light on the mechanism of localization by highlighting regions on protein sequences as likely targeting peptides. DeepMito [126] is another deep learning method for sub-mitochondrial localization prediction using CNNs. Its features include physical–chemical properties and PSSM in addition to the one-hot encoding of raw sequences.

Some methods do not predict localization directly; rather, they predict the presence and location of targeting peptides from which the localization of corresponding proteins can be roughly inferred. For example, DeepSig [99] and SignalP 5.0 [56] predict signal peptides and their cleavage sites using deep-learning methods. DeepSig uses a CNN, while SignalP 5.0 applies a CNN, bidirectional LSTM, and a CRF for specific signal peptide prediction. TargetP 2.0 [57] is a deep learning model constructed by bidirectional LSTM and a multi-attention mechanism to predict N-terminal targeting signals that direct proteins to the secretory pathways, mitochondria, and chloroplasts, or other plastids. One attention head was assigned to each target class and trained as the second loss function to focus on the peptide cleavage site.

3.5. Decision tree-based methods

For prediction problems involving large-scale labeled data, neural networks tend to outperform other algorithms or frameworks. However, when it comes to small- to medium-sized data, decision tree-based algorithms are often considered optimal. A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute, where each branch represents the outcome

of the test, and each leaf node represents a class label. Decision tree-based methods have evolved over the years. For example, bagging (bootstrap aggregating) combines the predictions of multiple decision trees through a majority voting mechanism, random forests select only a subset of features at random to build a forest of decision trees, and boosting is achieved by sequentially minimizing the errors of previous models. Gradient boosting employs the gradient descent algorithm to minimize errors in sequential models. XGBoost [127] optimizes gradient boosting through parallel processing, tree-pruning, handling missing values, and regularization to avoid overfitting.

Decision-tree-based methods have also been applied to protein localization problems. Pang et al. developed a CNN-XGBoost model [128] to predict protein subcellular localization. A CNN acts as a feature extractor to automatically obtain features from a protein sequence, and an XGBoost classifier functions as a recognizer based on the output of the CNN. SubMito-XGBoost [129] extracts protein sequence-based features including g-gap dipeptide composition, PseAAC, and PSSM as feature vectors for boosting to predict protein submitochondrial localization. A similar study [130] extracts feature vectors of protein sequences using PSSM for a random forest model. Both [129] and [130] apply the synthetic minority oversampling technique (SMOTE) to balance samples [131].

4. Tools

Many of the aforementioned methods mention web servers or standalone tools, but some of these are inaccessible due to lack of maintenance. We summarize a list of available protein localization prediction tools regarding their coverage, algorithms, accessibility, and other characteristics. These localization prediction tools (at the subcellular or suborganellar level) are shown in Table 1. Note that the BUSCA [132] and SubCons [133] tools are web servers that integrate different computational tools for protein subcellular localization prediction. The localization coverage of some tools, e.g., DeepSig and SignalP 2.0, is marked as SP (secretory pathway) in Table 1 because they are signal peptide prediction tools. Signal peptides direct proteins toward the secretory pathway, where the proteins are either located inside certain organelles (the endoplasmic reticulum, Golgi, or endosomes), secreted from the cell, or inserted into cellular membranes. Thus, the specific localization of these proteins is not unique. Some tools consider the secretory pathway as a low-resolution localization. For example, TargetP 2.0 predicts the presence of signal peptides and also predicts the targeting peptide for mitochondrial proteins and plastid proteins where unique protein localization can be inferred.

To assess prediction tools, competitions can provide large-scale blind tests for objective evaluation. A well-known example is the CASP [134] in the protein structure prediction field. For protein localization prediction, the Critical Assessment of protein Function Annotation algorithms (CAFA) [135] is a good platform for such a purpose. CAFA requires a method to provide prediction in the form of cellular component ontology (CCO) terms. However, most methods reviewed in this paper predict UniProt's localization annotations rather than the CCO terms, and hence may not be assessed at CAFA directly. DeepLoc is a state-of-the-art method, and their dataset is often used by new methods for training and testing, as well as method comparison. Here, we used the DeepLoc dataset as a benchmark to evaluate some of the tools. The DeepLoc dataset was extracted from the UniProt database, release 2016_04. The protein dataset was filtered using the following criteria: eukaryotic, complete protein, encoded in the nucleus, longer than 40 amino acids, and experimentally verified (ECO:0000269) single localization annotation. Similar locations or subclasses of the same location were mapped to 10 main locations to increase the number of proteins per compartment (refer to Table 1 in [49] for details regarding the class distribution). A total of 13,858 proteins were obtained after the filtering process. PSI-CD-HIT [137] was used to cluster proteins with 30% identity or a 10^{-6} E-value cutoff, and the alignment was required to cover 80% of the shorter sequences, resulting in 8410 clusters for the whole dataset. The five-fold datasets generated had approximately the same number of proteins at each location. Four of the datasets were used for training and validation, and one was held out for testing. In this way, the redundancy between the training and testing datasets was reduced.

The DeepLoc, MULocDeep, SeqVec, ProtVec, and ProtTrans methods were stringently trained and tested using the training and testing samples in the DeepLoc dataset. LocTree2, MultiLoc2, CELLO, WoLF PSORT, YLoc, SherLoc2, and iLoc-Euk were run on the testing samples in the DeepLoc dataset. Thus, their performance is potentially overestimated because redundancy control was not performed. All the evaluated methods could be applied to proteins in eukaryotic cells. In the cases where a method predicted more than ten locations, the predicted locations were mapped onto the ten locations in the DeepLoc dataset. Overall accuracy is used as the evaluation criterion. The evaluation performance is directly cited from [48,49,68–70]. As shown in Fig. 2, the deep learning-based methods (DeepLoc, MULocDeep, ProtTrans, and SeqVec) have overall better performance than the other methods, except for ProtVec [138], which uses Word2Vec, a context-independent embedding method. DeepLoc_PSSM achieves better performance than DeepLoc_BLOSUM, indicating that evolutionary information enhances localization prediction. By comparing the performance of pre-trained methods (ProtTrans and SeqVec) with other deep learning methods (DeepLoc and MULocDeep), we find that a simple deep learning architecture with pre-train embedding can achieve competitive or even better performance than delicately designed deep-learning models using evolutionary profile features.

5. Discussion and outlook

The computational prediction of protein localization has significantly improved prediction accuracy and localization mechanism studies over past two decades, especially with deep learning. However, the current methods still have limitations. For example, an 80% overall prediction accuracy shown in Fig. 2 does not mean that the localization prediction problem is 80% solved. In particular, many suborganellar localizations do not have sufficient data to build reliable prediction models. In this section, we discuss several areas for future exploration of localization analysis methods.

Protein localization problems have several biological characteristics. Many proteins can localize to more than one compartment. Some proteins are tissue-/cell type-specific, meaning their localization varies between different tissues or cell types. Proteins expressed at the correct location but with altered efficiency or concentration can also lead to illness. Thus, quantitively measuring or predicting protein localization in different tissues or cell types are in great demand. Additionally, proteins may be mislocalized due to mutations, which may have disease consequences [5]. Predicting mislocalization due to mutations is also challenging because it requires more sensitive methods with individual residue resolution.

Researchers could also pay more attention to biological interpretability when designing future localization analysis models. The mechanism of protein localization is complicated. In addition to targeting peptides, which are considered in some existing methods, other phenomena can affect/control protein localization. The trafficking machinery in cells controls the transport of molecules across membranes of organelles. Dysregulation of the protein traf-

Table 1
Summary of protein localization prediction tools.

Tool	Cov_lv1	Cov_lv2	Species kingdom	Algorithm	Metrics	Year	Web server	Standalone
BUSCA [132]	1–4,7,11–14		Eu,Pro	Integrated method	F1, MCC	2018	http://busca.biocomp.unibo.it/	
CELLO2GO [89]	1–6,8–11,15		Eu,Pro,V	SVM and homology search	Acc	2014	http://cello.life.nctu.edu.tw/cello2go/	
MULocDeep [48]	1–10	1–10	Eu	LSTM + attention	Acc, MCC, Rec, Prec, ROC_AUC, P&R_AUC	2021	http://mu-loc.org/	√
DeepLoc [49]	1–10		Eu	CNN + LSTM + attention	Acc, MCC, Gorodkin measure	2017	https://services.healthtech.dtu.dk/service.php?DeepLoc-1.0	
TargetP 2.0 [57]	SP,4,7		Eu,Pro	LSTM + attention	Prec, Rec, F1, MCC	2019	https://services.healthtech.dtu.dk/service.php?TargetP-2.0	
MU-LOC [76]	4		P	SVM and neural network	Acc, Prec, F1, MCC	2018	http://136.32.161.178/	√
LocTree3 [42]	1–4,6–11		Eu,Pro	SVM and homology search	Acc, Std	2014	https://rostlab.org/services/loctree3/	
MitoFates [92]	4		Eu	SVM	Prec, Rec, MCC, ROC_AUC	2015	http://mitf.cbrc.jp/MitoFates/cgi-bin/top.cgi	√
LOCALIZER [60]	1,4,7		P	SVM	SN, SP, PPV, MCC, Acc	2017	http://localizer.csiro.au/	√
SignalP 5.0 [56]	SP		Eu,Pro	CNN, bidirectional LSTM, and CRF	MCC, Rec, Prec	2019	http://www.cbs.dtu.dk/services/SignalP/	√
DeepSig [99]	SP		Eu,Bac	CNN and CRF	MCC, FPR, F1	2018	https://deepsig.biocomp.unibo.it/welcome/default/index	√
PSORTb 3.0 [96]	2,3,14–16		Bac	SVM and homology search	Prec, Rec, Acc, MCC	2010	https://www.psort.org/psortb/	√
WoLF PSORT [103]	1–4,7,11		Eu	k-NN classifier	Acc	2007	https://wolfpsort.hgc.jp/	
SubCons [133]	1–4,6,8–11		Hum	Integrated method	F1, MCC	2017	https://subcons.bioinfo.se/	
TPpred 3.0 [136]	4,7		Eu	Integrated method	MCC, Prec, Rec	2015	https://tppred3.biocomp.unibo.it/tppred3	√
MultiLoc2 [44]	1–4,6–11		Eu	SVM	SN, SP, MCC	2009	https://abi-services.informatik.uni-tuebingen.de/multiloc2/webloc.cgi	√
YLoc [45]	1–4,6–11		Eu	Naïve Bayes and entropy-based discretization	F1, Acc	2010	https://abi-services.informatik.uni-tuebingen.de/yloc/webloc.cgi	√
SCLpred-EMS	SP		Eu	Neural network	SP, SN, FPR, MCC	2020	http://distilldeep.ucd.ie/SCLpred2/	
ERPred [87]	6		Eu	SVM	Acc, SN, SP, MCC	2017	http://proteininformatics.org/mkumar/erpred/index.html	√
SeqVec[68]	1–10		Eu	Language Model + FNN	Acc, MCC, FPR	2019	https://embed.protein.properties/	√
ProtTrans [69]	1–10		Eu	Language Model + FNN	Acc	2020	https://embed.protein.properties/	√
LA [70]	1–10		Eu	Language Model + attention	Acc	2021	https://embed.protein.properties/	√
DeepMito [126]		4	Eu	CNN	MCC, GCC	2019	http://busca.biocomp.unibo.it/deepmito/	√
SubGolgi v2 [59]	8	8	Eu	SVM	SN, Acc, MCC	2013	http://lin-group.cn/server/subGolgi2	
TetraMito [58]		4	Eu	SVM	SN, Acc, MCC	2013	http://lin-group.cn/server/TetraMito	
Schloro [93]		7	P	SVM	Acc, Rec, Prec, F1, ROC_AUC, MCC	2017	https://schloro.biocomp.unibo.it/welcome/default/index	√
SubMitoPred [86]	4	4	Eu	SVM	Acc	2017	http://proteininformatics.org/mkumar/submitopred/	√
SubNucPred [88]		1	Eu	SVM	Acc, SN, SP, MCC	2014	http://proteininformatics.org/mkumar/subnucpred/index.html	√

The localization coverage codes are: 1. nucleus; 2. cytoplasm; 3. extracellular; 4. mitochondrion; 5. cell membrane; 6. endoplasmic reticulum; 7. plastid/chloroplast; 8. Golgi apparatus; 9. lysosome/vacuole; 10. peroxisome; 11. plasma membrane; 12. organelle membrane; 13. endomembrane system; 14. outer membrane; 15. periplasmic; 16. cell wall; SP. secretory pathway.
Cov_lv1 represents subcellular localization coverage, and Cov_lv2 indicates that suborganellar localization predictions are provided for the organelle.
The species kingdom codes are: Eu (Eukaryota, including animal, plant, and fungi); Pro (Prokaryota, including Bacteria and Archaea); V (Virus); P (Plant); Bac (Bacteria); Hum (Human).
The metrics codes are: MCC (Matthews correlation coefficient), Acc (accuracy), SN (sensitivity), SP (specificity), Prec (precision), Rec (recall), ROC_AUC (area under receiver operating characteristic curve), P&R_AUC (area under precision & recall curve), GCC (Generalized Correlation Coefficient), PPV (positive predictive value), FPR (false positive rate).

ficking machinery can have dramatic effects on general protein transport processes [139]. For example, the homozygous mutation R391H in the nucleoporin NUP155 has been shown to reduce nuclear envelope permeability and affect the export of Hsp70 mRNA and import of HSP70 protein [140]. Another fairly common method that affects protein localization involves binding partners that carry bound proteins between compartments. This mechanism allows for indirect control of protein localization by regulating the localization and concentration levels of binding partners, similar to the role of import receptors [9]. However, the prediction of protein localization changes affected by other proteins has not been explored. Furthermore, some localization signals are not contained within the linear peptide sequence of a cargo protein but are formed by the arrangement of amino acid residues on its surface. One advantage of such an arrangement is that conformational changes induced by allosteric events can disrupt or reform the localization signal transiently in response to the state of the protein [9]. Making protein localization analysis methods inter-

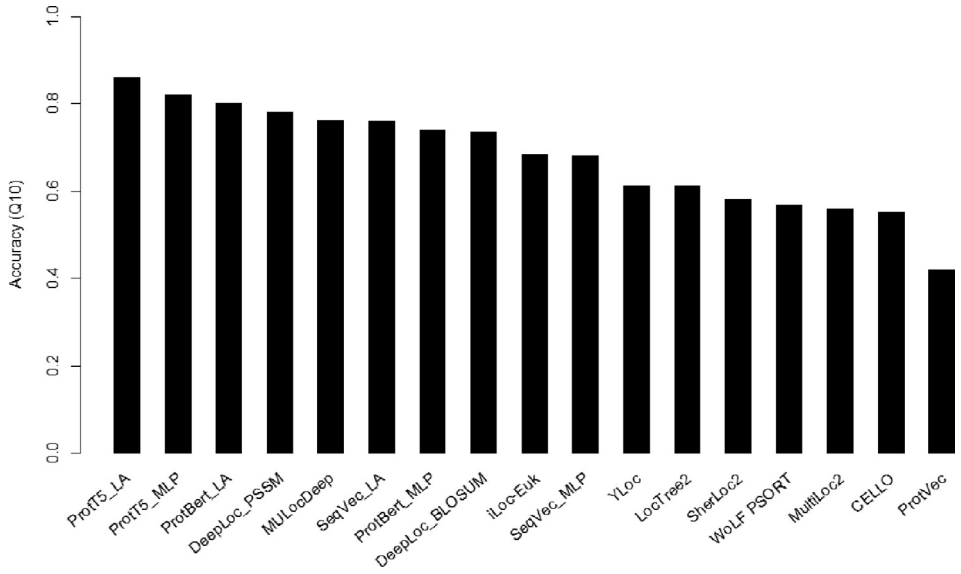

Fig. 2. Evaluations of protein localization methods/tools. The criterion is the overall prediction accuracy for 10 main localizations. DeepLoc_PSSM and DeepLoc_BLOSUM are DeepLoc methods with PSSM and BLOSUM62 embedding, respectively. ProtT5_MLP and ProtBert_MLP are simple feed-forward neural networks in the ProtTrans method but using pre-train embeddings by T5 and Bert, respectively. ProtT5_LA and ProtBert_LA use the same two pre-trained models as above but are followed by an attention-based neural network.

pretable would allow us to answer "how" besides "where" a protein localizes, which has implications in pathology and drug design. The corresponding training data for such methods is currently lacking but may become available in the near future.

CRediT authorship contribution statement

Yuexu Jiang: Conceptualization, Investigation, Visualization, Validation, Writing - original draft. **Duolin Wang:** Visualization, Writing - review & editing. **Weiwei Wang:** Validation. **Dong Xu:** Writing - review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the US National Institutes of Health grants R35-GM126985 and R21-LM012790. We like to thank Dr. Ian Max Møller for the helpful discussions.

References

[1] Schnell DJ, Hebert DN. Protein translocons: multifunctional mediators of protein translocation across membranes. Cell 2003;112(4):491–505.
[2] Wickner W, Schekman R. Protein translocation across biological membranes. Science 2005;310(5753):1452–6.
[3] Neupert W, Herrmann JM. Translocation of proteins into mitochondria. Annu Rev Biochem 2007;76:723–49.
[4] Davis JR, Kakar M, Lim CS. Controlling protein compartmentalization to overcome disease. Pharm Res 2007;24(1):17–27.
[5] Hung MC, Link W. Protein localization in disease and therapy. J Cell Sci 2011;124(Pt 20):3381–92.
[6] Rodriguez JA, Schüchner S, Au WW, Fabbro M, Henderson BR. Nuclear–cytoplasmic shuttling of BARD1 contributes to its proapoptotic activity and is regulated by dimerization with BRCA1. Oncogene 2004;23(10):1809–20.
[7] Marques-Bueno MM, Moreno-Romero J, Abas L, De Michele R, Martinez MC. A dominant negative mutant of protein kinase CK2 exhibits altered auxin responses in Arabidopsis. Plant J 2011;67(1):169–80.
[8] Thevissen K, de Mello TP, Xu D, Blankenship J, Vandenbosch D, Idkowiak-Baldys J, et al. The plant defensin RsAFP2 induces cell wall stress, septin mislocalization and accumulation of ceramides in Candida albicans. Mol Microbiol 2012;84(1):166–80.
[9] Bauer NC, Doetsch PW, Corbett AH. Mechanisms regulating protein localization. Traffic 2015;16(10):1039–61.
[10] Hagmann M. Protein zip codes make Nobel journey. Science 1999;286(5440):666.
[11] Chacinska A, Koehler CM, Milenkovic D, Lithgow T, Pfanner N. Importing mitochondrial proteins: machineries and mechanisms. Cell 2009;138(4):628–44.
[12] Schmidt O, Pfanner N, Meisinger C. Mitochondrial protein import: from proteomics to functional mechanisms. Nat Rev Mol Cell Biol 2010;11(9):655–67.
[13] Jakobsen L, Vanselow K, Skogs M, Toyoda Y, Lundberg E, Poser I, et al. Novel asymmetrically localizing components of human centrosomes identified by complementary proteomics methods. EMBO J 2011;30(8):1520–35.
[14] Christoforou A, Mulvey CM, Breckels LM, Geladaki A, Hurrell T, Hayward PC, et al. A draft map of the mouse pluripotent stem cell spatial proteome. Nat Commun 2016;7:8992.
[15] Itzhak DN, Tyanova S, Cox J, Borner GH. Global, quantitative and dynamic mapping of protein subcellular localization. Elife 2016;5.
[16] Orre LM, Vesterlund M, Pan Y, Arslan T, Zhu Y, Fernandez Woodbridge A, et al. Proteome-wide mapping of protein localization and relocalization. Mol Cell 2019;73(1):166–182 e167.
[17] Rhee HW, Zou P, Udeshi ND, Martell JD, Mootha VK, Carr SA, et al. Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging. Science 2013;339(6125):1328–31.
[18] Hung V, Zou P, Rhee HW, Udeshi ND, Cracan V, Svinkina T, et al. Proteomic mapping of the human mitochondrial intermembrane space in live cells via ratiometric APEX tagging. Mol Cell 2014;55(2):332–41.
[19] Lee SY, Kang MG, Park JS, Lee G, Ting AY, Rhee HW. APEX fingerprinting reveals the subcellular localization of proteins of interest. Cell Rep 2016;15(8):1837–47.
[20] Chong YT, Koh JL, Friesen H, Duffy SK, Cox MJ, Moses A, et al. Yeast proteome dynamics from single cell imaging and automated analysis. Cell 2015;161(6):1413–24.
[21] Barbe L, Lundberg E, Oksvold P, Stenius A, Lewin E, Bjorling E, et al. Toward a confocal subcellular atlas of the human proteome. Mol Cell Proteomics 2008;7(3):499–508.
[22] Stadler C, Skogs M, Brismar H, Uhlen M, Lundberg E. A single fixation protocol for proteome-wide immunofluorescence localization studies. J Proteomics 2010;73(6):1067–78.
[23] Thul PJ, Akesson L, Wiking M, Mahdessian D, Geladaki A, Ait Blal H, et al. A subcellular map of the human proteome. Science 2017;356(6340).
[24] Burns TJ, Frei AP, Gherardini PF, Bava FA, Batchelder JE, Yoshiyasu Y, et al. High-throughput precision measurement of subcellular localization in single cells. Cytometry A 2017;91(2):180–9.
[25] Gardy JL, Brinkman FS. Methods for predicting bacterial protein subcellular localization. Nat Rev Microbiol 2006;4(10):741–51.
[26] Nakai K. Protein sorting signals and prediction of subcellular localization. Adv Protein Chem 2000;54:277–344.

[27] Imai K, Nakai K. Tools for the recognition of sorting signals and the prediction of subcellular localization of proteins from their amino acid sequences. Front Genet 2020;11:607812.
[28] Bonetta R, Valentino G. Machine learning techniques for protein function prediction. Proteins 2020;88(3):397–413.
[29] Shen Y, Ding Y, Tang J, Zou Q, Guo F. Critical evaluation of web-based prediction tools for human protein subcellular localization. Brief Bioinform 2020;21(5):1628–40.
[30] Donnes P, Hoglund A. Predicting protein subcellular localization: past, present, and future. Genomics Proteomics Bioinformatics 2004;2(4):209–15.
[31] Chou KC, Shen HB. Recent progress in protein subcellular location prediction. Anal Biochem 2007;370(1):1–16.
[32] Wang Z, Zou Q, Jiang Y, Ju Y, Zeng X. Review of protein subcellular localization prediction. Curr Bioinform 2014;9(3):331–42.
[33] Kumar R, Dhanda SK. Bird eye view of protein subcellular localization prediction. Life (Basel) 2020;10(12).
[34] UniProt C. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 2019;47(D1):D506–15.
[35] Chou KC. A novel approach to predicting protein structural classes in a (20–1)-D amino acid composition space. Proteins 1995;21(4):319–44.
[36] Chou K. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins: Struct Funct Bioinf 2001;43(3):246–55.
[37] Nair R, Rost B. Sequence conserved for subcellular localization. Protein Sci 2002;11(12):2836–47.
[38] Joshi T, Xu D. Quantitative assessment of relationship between sequence similarity and function similarity. BMC Genomics 2007;8(1):1–10.
[39] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Mol Biol 1990;215(3):403–10.
[40] Remmert M, Biegert A, Hauser A, Soding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods 2011;9(2):173–5.
[41] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25(1):25–9.
[42] Goldberg T, Hecht M, Hamp T, Karl T, Yachdav G, Ahmed N, et al. LocTree3 prediction of localization. Nucleic Acids Res 2014;42(Web Server issue):W350–5.
[43] Briesemeister S, Blum T, Brady S, Lam Y, Kohlbacher O, Shatkay H. SherLoc2: a high-accuracy hybrid method for predicting subcellular localization of proteins. J Proteome Res 2009;8(11):5363–6.
[44] Blum T, Briesemeister S, Kohlbacher O. MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. BMC Bioinf 2009;10(1):274.
[45] Briesemeister S, Rahnenfuhrer J, Kohlbacher O: YLoc–an interpretable web server for predicting subcellular localization. *Nucleic Acids Res* 2010, **38**(Web Server issue):W497-502.
[46] Zheng W, Zhang C, Li Y, Pearce R, Bell EW, Zhang Y: Folding non-homologous proteins by coupling deep-learning contact maps with I-TASSER assembly simulations. *Cell Reports Methods* 2021:100014.
[47] Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci 1992;89(22):10915–9.
[48] Jiang Y, Wang D, Yao Y, Eubel H, Künzler P, Møller IM, et al. MULocDeep: A deep-learning framework for protein subcellular and suborganellar localization prediction with residue-level interpretation. Comput Struct Biotechnol J 2021;19:4825–39.
[49] Almagro Armenteros JJ, Sonderby CK, Sonderby SK, Nielsen H, Winther O. DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics 2017;33(21):3387–95.
[50] Jaakkola T, Diekhans M, Haussler D. A discriminative framework for detecting remote protein homologies. J Comput Biol 2000;7(1–2):95–114.
[51] Kuang R, Ie E, Wang K, Wang K, Siddiqi M, Freund Y, et al. Profile-based string kernels for remote homology detection and motif extraction. J Bioinf Comput Biol 2005;3(03):527–50.
[52] Goldberg T, Hamp T, Rost B. LocTree2 predicts localization for all domains of life. Bioinformatics 2012;28(18):i458–65.
[53] Gardy JL, Spencer C, Wang K, Ester M, Tusnady GE, Simon I, et al. PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. Nucleic Acids Res 2003;31(13):3613–7.
[54] Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, et al. PROSITE: a documented database using patterns and profiles as motif descriptors. Briefings Bioinf 2002;3(3):265–74.
[55] Blobel G, Dobberstein B. Transfer of proteins across membranes. I. Presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane-bound ribosomes of murine myeloma. J Cell Biol 1975;67(3):835–51.
[56] Almagro Armenteros JJ, Tsirigos KD, Sonderby CK, Petersen TN, Winther O, Brunak S, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol 2019;37(4):420–3.
[57] Almagro Armenteros JJ, Salvatore M, Emanuelsson O, Winther O, von Heijne G, Elofsson A, et al. Detecting sequence signals in targeting peptides using deep learning. Life Sci Alliance 2019;2(5).
[58] Lin H, Chen W, Yuan L, Li Z, Ding H. Using over-represented tetrapeptides to predict protein submitochondria locations. Acta Biotheor 2013;61(2):259–68.
[59] Ding H, Guo S-H, Deng E-Z, Yuan L-F, Guo F-B, Huang J, et al. Prediction of Golgi-resident protein types by using feature selection technique. Chemometr Intell Lab Syst 2013;124:9–13.
[60] Sperschneider J, Catanzariti A-M, DeBoer K, Petre B, Gardiner DM, Singh KB, et al. LOCALIZER: subcellular localization prediction of both plant and effector proteins in the plant cell. Sci Rep 2017;7(1):1–14.
[61] Venkatarajan MS, Braun W. New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical–chemical properties. Mol Model Annual 2001;7(12):445–53.
[62] Kawashima S, Ogata H, Kanehisa M. AAindex: amino acid index database. Nucleic Acids Res 1999;27(1):368–9.
[63] Perdigão N, Heinrich J, Stolte C, Sabir KS, Buckley MJ, Tabor B, et al. Unexpected features of the dark proteome. Proc Natl Acad Sci 2015;112(52):15898–903.
[64] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I: Attention is all you need. *arXiv preprint arXiv:03762* 2017.
[65] Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. Deep contextualized word representations. *arXiv preprint arXiv:05365* 2018.
[66] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. p. 3111–9.
[67] Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:04805* 2018.
[68] Heinzinger M, Elnaggar A, Wang Y, Dallago C, Nechaev D, Matthes F, et al. Modeling aspects of the language of life through transfer-learning protein sequences. BMC Bioinf 2019;20(1):1–17.
[69] Elnaggar A, Heinzinger M, Dallago C, Rihawi G, Wang Y, Jones L, et al. ProtTrans: towards cracking the language of Life's code through self-supervised deep learning and high performance computing. *arXiv preprint arXiv:06225* 2020.
[70] Stärk H, Dallago C, Heinzinger M, Rost B. Light attention predicts protein location from the language of life. 2021:2021.2004.2025.441334.
[71] Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, et al. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res* 2012;**40**(Database issue):D857-861.
[72] Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res 2002;30(1):303–5.
[73] Oughtred R, Stark C, Breitkreutz BJ, Rust J, Boucher L, Chang C, et al. The BioGRID interaction database: 2019 update. Nucleic Acids Res 2019;47(D1):D529–41.
[74] Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 2019;47(D1):D607–13.
[75] Ananda MM, Hu J. NetLoc: Network based protein localization prediction using protein-protein interaction and co-expression networks. In: *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM): 2010*. IEEE: 142-148.
[76] Zhang N, Rao RSP, Salvato F, Havelund JF, Moller IM, Thelen JJ, et al. MU-LOC: A machine-learning method for predicting mitochondrially localized proteins in plants. Front Plant Sci 2018;9:634.
[77] Ryngajllo M, Childs L, Lohse M, Giorgi FM, Lude A, Selbig J, et al. SLocX: Predicting subcellular localization of Arabidopsis proteins leveraging gene expression data. Front Plant Sci 2011;2:43.
[78] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;30(1):207–10.
[79] Tomczak K, Czerwinska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn) 2015;19(1A):A68–77.
[80] Cortes C, Vapnik V. Support-vector networks. Machine Learning 1995;20(3):273–97.
[81] Hua S, Sun Z. Support vector machine approach for protein subcellular localization prediction. Bioinformatics 2001;17(8):721–8.
[82] Sarda D, Chua GH, Li KB, Krishnan A. pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties. BMC Bioinf 2005;6:152.
[83] Tax DM, Duin RP. Support vector data description. Machine Learning 2004;54(1):45–66.
[84] Lee K, Kim DW, Na D, Lee KH, Lee D. PLPD: reliable protein localization prediction from imbalanced and overlapped datasets. Nucleic Acids Res 2006;34(17):4655–66.
[85] Yu CS, Chen YC, Lu CH, Hwang JK. Prediction of protein subcellular localization. Proteins 2006;64(3):643–51.
[86] 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.
[87] 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.
[88] Kumar R, Jain S, Kumari B, Kumar M. Protein sub-nuclear localization prediction using SVM and Pfam domain information. PLoS ONE 2014;9(6).

[89] Yu CS, Cheng CW, Su WC, Chang KC, Huang SW, Hwang JK, et al. CELLO2GO: a web server for protein subCELLular LOcalization prediction with functional gene ontology annotation. PLoS ONE 2014;9(6):e99368.
[90] Wang X, Zhang W, Zhang Q, Li GZ. MultiP-SChlo: multi-label protein subchloroplast localization prediction with Chou's pseudo amino acid composition and a novel multi-label classifier. Bioinformatics 2015;31 (16):2639–45.
[91] Hasan MAM, Ahmad S, Molla MKI. Protein subcellular localization prediction using multiple kernel learning based support vector machine. Mol BioSyst 2017;13(4):785–95.
[92] Fukasawa Y, Tsuji J, Fu SC, Tomii K, Horton P, Imai K. MitoFates: improved prediction of mitochondrial targeting sequences and their cleavage sites. Mol Cell Proteomics 2015;14(4):1113–26.
[93] Savojardo C, Martelli PL, Fariselli P, Casadio R. SChloro: directing Viridiplantae proteins to six chloroplastic sub-compartments. Bioinformatics 2017;33 (3):347–53.
[94] Joyce J. Bayes' theorem. *The Stanford Encyclopedia of Philosophy* 2003.
[95] Gardy JL, Laird MR, Chen F, Rey S, Walsh CJ, Ester M, et al. PSORTb vol 2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. Bioinformatics 2005;21 (5):617–23.
[96] Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, et al. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics 2010;26(13):1608–15.
[97] Lee H, Tu Z, Deng M, Sun F, Chen T. Diffusion kernel-based logistic regression models for protein function prediction. OMICS 2006;10(1):40–55.
[98] Chung MK. Introduction to random fields. *arXiv preprint arXiv:09660* 2020.
[99] Savojardo C, Martelli PL, Fariselli P, Casadio R. DeepSig: deep learning improves signal peptide detection in proteins. Bioinformatics 2018;34 (10):1690–6.
[100] Zhu L, Hofestädt R, Ester M. Tissue-specific subcellular localization prediction using multi-label Markov random fields. IEEE/ACM Trans Comput Biol Bioinf 2019;16(5):1471–82.
[101] Thul PJ, Lindskog C. The human protein atlas: A spatial map of the human proteome. Protein Sci 2018;27(1):233–44.
[102] Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat 1992;46(3):175–85.
[103] Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams Collier CJ, et al. WoLF PSORT: protein localization predictor. Nucleic Acids Res 2007;35(Web Server issue):W585–7.
[104] Garapati HS, Male G, Mishra K. Predicting subcellular localization of proteins using protein-protein interaction data. Genomics 2020;112(3):2361–8.
[105] Chandra MP. On the generalised distance in statistics. In: Proceedings of the National Institute of Sciences of India. p. 49–55.
[106] Chou K-C, Elrod DW. Protein subcellular location prediction. Protein Eng 1999;12(2):107–18.
[107] Zhou GP, Doctor K. Subcellular location prediction of apoptosis proteins. Proteins 2003;50(1):44–8.
[108] Ding H, Liu L, Guo F-B, Huang J, Lin H. Identify Golgi protein types with modified mahalanobis discriminant algorithm and pseudo amino acid composition. Protein Peptide Letters 2011;18(1):58–63.
[109] Goodfellow I, Bengio Y, Courville A, Bengio Y. Deep learning, vol. 1. Cambridge: MIT Press; 2016.
[110] Bengio Y. Learning deep architectures for AI. Now Publishers Inc; 2009.
[111] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986;323(6088):533–6.
[112] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8):1735–80.
[113] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint* 2014.
[114] Mooney C, Wang YH, Pollastri G. SCLpred: protein subcellular localization prediction by N-to-1 neural networks. Bioinformatics 2011;27(20):2812–9.
[115] Wang X, Jin Y, Zhang Q. DeepPred-SubMito: A novel submitochondrial localization predictor based on multi-channel convolutional neural network and dataset balancing treatment. Int J Mol Sci 2020;21(16):5710.
[116] Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent neural networks. In: *Proceedings of the 30th International Conference on Machine Learning; Proceedings of Machine Learning Research*: Edited by Sanjoy D, David M. PMLR 2013: 1310–1318.
[117] Kalchbrenner N, Blunsom P. Recurrent continuous translation models. In: Proceedings of the 2013 conference on empirical methods in natural language processing. p. 1700–9.
[118] Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:03762* 2014.
[119] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. p. 3104–12.
[120] Sak H, Senior AW, Beaufays F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. Interspeech 2014:338–42.
[121] Li X, Wu X. Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP): 2015*. IEEE: 4520-4524.
[122] Valueva MV, Nagornov N, Lyakhov PA, Valuev GV, Chervyakov N. Application of the residue number system to reduce hardware costs of the convolutional neural network implementation. Math Comput Simul 2020;177:232–43.
[123] Wang D, Zhang Z, Jiang Y, Mao Z, Wang D, Lin H, et al. DM3Loc: multi-label mRNA subcellular localization prediction and analysis based on multi-head self-attention mechanism. Nucleic Acids Res 2021.
[124] Lin Z, Feng M, Santos CNd, Yu M, Xiang B, Zhou B, Bengio Y: A structured self-attentive sentence embedding. *arXiv preprint* 2017.
[125] Kaleel M, Zheng Y, Chen J, Feng X, Simpson JC, Pollastri G, et al. SCLpred-EMS: Subcellular localization prediction of endomembrane system and secretory pathway proteins by deep N-to-1 convolutional neural networks. Bioinformatics 2020;36(11):3343–9.
[126] Savojardo C, Bruciaferri N, Tartari G, Martelli PL, Casadio R. DeepMito: accurate prediction of protein submitochondrial localization using convolutional neural networks. Bioinformatics 2019;36(1):56–64.
[127] Chen T, Guestrin C: Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining: 2016*. 785-794.
[128] Pang L, Wang J, Zhao L, Wang C, Zhan H. A novel protein subcellular localization method with CNN-XGBoost model for Alzheimer's disease. Front Genet 2019;9:751.
[129] Yu B, Qiu W, Chen C, Ma A, Jiang J, Zhou H, et al. SubMito-XGBoost: predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting. Bioinformatics 2020;36(4):1074–81.
[130] Wu L, Huang S, Wu F, Jiang Q, Yao S, Jin X. Protein subnuclear localization based on radius-SMOTE and kernel linear discriminant analysis combined with random forest. Electronics 2020;9(10):1566.
[131] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321–57.
[132] Savojardo C, Martelli PL, Fariselli P, Profiti G, Casadio R. BUSCA: an integrative web server to predict subcellular localization of proteins. Nucleic Acids Res 2018;46(W1):W459–66.
[133] Salvatore M, Warholm P, Shu N, Basile W, Elofsson A. SubCons: a new ensemble method for improved human subcellular localization predictions. Bioinformatics 2017;33(16):2464–70.
[134] Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)—Round XIII. Proteins: Struct Funct Bioinf 2019;87(12):1011–20.
[135] Zhou N, Jiang Y, Bergquist TR, Lee AJ, Kacsoh BZ, Crocker AW, et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. Genome Biol 2019;20(1):244.
[136] Savojardo C, Martelli PL, Fariselli P, Casadio R. TPpred3 detects and discriminates mitochondrial and chloroplastic targeting peptides in eukaryotic proteins. Bioinformatics 2015;31(20):3269–75.
[137] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22(13):1658–9.
[138] Asgari E, Mofrad MRJPo. Continuous distributed representation of biological sequences for deep proteomics and genomics. 2015, **10**(11):e0141287.
[139] Chahine MN, Pierce GN. Therapeutic targeting of nuclear protein import in pathological cell conditions. Pharmacol Rev 2009;61(3):358–72.
[140] Zhang X, Chen S, Yoo S, Chakrabarti S, Zhang T, Ke T, et al. Mutation in nuclear pore component NUP155 leads to atrial fibrillation and early sudden cardiac death. Cell 2008;135(6):1017–27.