frontiers in Genetics

REVIEW
published: 25 November 2020
doi: 10.3389/fgene.2020.607812

Tools for the Recognition of Sorting Signals and the Prediction of Subcellular Localization of Proteins From Their Amino Acid Sequences

Kenichiro Imai[1] *and Kenta Nakai*[2]*

[1]*Cellular and Molecular Biotechnology Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan,* [2]*The Institute of Medical Science, The University of Tokyo, Tokyo, Japan*

At the time of translation, nascent proteins are thought to be sorted into their final subcellular localization sites, based on the part of their amino acid sequences (i.e., sorting or targeting signals). Thus, it is interesting to computationally recognize these signals from the amino acid sequences of any given proteins and to predict their final subcellular localization with such information, supplemented with additional information (e.g., *k*-mer frequency). This field has a long history and many prediction tools have been released. Even in this era of proteomic atlas at the single-cell level, researchers continue to develop new algorithms, aiming at accessing the impact of disease-causing mutations/cell type-specific alternative splicing, for example. In this article, we overview the entire field and discuss its future direction.

Keywords: protein sorting/targeting, subcellular loalization, sorting/targeting signals, prediction methods, bacteria, archaea, eukarya

OPEN ACCESS

Edited by:
Shuai Cheng Li,
City University of Hong Kong, Hong Kong

Reviewed by:
Litao Sun,
Sun Yat-sen University, China
Marti Aldea,
Instituto de Biología Molecular de Barcelona (IBMB), Spain

***Correspondence:**
Kenta Nakai
knakai@ims.u-tokyo.ac.jp

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Genetics

Received: *18 September 2020*
Accepted: *03 November 2020*
Published: *25 November 2020*

Citation:
Imai K and Nakai K (2020) Tools for the Recognition of Sorting Signals and the Prediction of Subcellular Localization of Proteins From Their Amino Acid Sequences. Front. Genet. 11:607812. doi: 10.3389/fgene.2020.607812

INTRODUCTION

Although we should not underestimate the importance of non-coding genes, the main players of the genetic system of living organisms are still regarded as protein-coding genes, which specify amino acid sequence information. Thus, in principle, we should be able to infer the *in vivo* fate of any protein from its amino acid sequence, if its environmental conditions, such as the cell type where it is synthesized, are appropriately given. For example, we should be able to predict the three-dimensional structure of a protein from its sequence or to design novel amino acid sequences that take a desired three-dimensional structure (Baker, 2019), as well as to predict how it binds/interacts with other proteins/small molecule ligands (Vakser, 2020). Another important information to be predicted is which kind of post-translational modifications, if any, it will take [at which residue(s); Audagnotto and Dal Peraro, 2017]. Also, it may be possible to predict the half-life of a given protein/peptide-based on the degradation signals (degrons) and/or other properties (Mathur et al., 2018; Eldeeb et al., 2019). Finally, the prediction of subcellular localization of a protein based on its amino acid sequence is a challenging field in bioinformatics. It is well accepted that the protein sorting for subcellular localization is regulated by so-called protein sorting (or targeting) signals, which are typically represented as a short stretch(es) of its amino acid sequence. Nowadays, many of the protein localization mechanisms/pathways that recognize and utilize such signals have been clarified. Therefore, many predictors

have been developed for the recognition of such sorting signals and attempts have been done to combine such predictors, leading to the comprehensive prediction of the final localization site. However, not all such signals have been clarified. Moreover, not all proteins are equipped with such typical signals and use some alternative (minor/exceptional) pathways. Adding the knowledge of such exceptional cases will make the prediction system gradually more realistic but the objective assessment of its performance, like the ones commonly used in the field of machine learning, will become difficult because the knowledge of exceptional cases are quite unlikely to be generalized (in other words, any sequence features of such exceptional proteins, which are nothing to do with their sorting mechanisms, would work as clues for their prediction). It should be also noted that the practical value of subcellular localization predictors has been degraded because the localization information is being comprehensively determined with subcellular proteomics experiments (Harvey Millar and Taylor, 2014). However, the rise of synthetic biology as well as precision medicine will demand prediction tools that enable the prediction against artificial proteins and/or the prediction of the impact of mutations/polymorphic variations on potential sorting signals.

In this review article, we will introduce the outline of this field, emphasizing its recent progress. The readers are recommended to refer to additional reviews by other authors and ourselves, too (Imai and Nakai, 2010, 2019; Du and Xu, 2013; Nielsen, 2017; Nielsen et al., 2019).

PREDICTION OF SUBCELLULAR LOCALIZATION SITES FOR BACTERIAL/ARCHAEAL PROTEINS

Even in the simplest type of organisms, which are unicellular organisms without any subcellular compartments, proteins can be localized at either the cytoplasmic space, the cellular membrane, or the extracellular space (i.e., secreted). This is basically the case for so-called Gram-positive bacteria and archaea, but, in reality, they also have a cell wall for another localization site. The basic prediction strategy for these proteins is to combine two kinds of predictors: a predictor for N-terminal signal peptides and that for transmembrane segments. Namely, a protein that neither has an N-terminal (and cleavable) signal peptide nor any hydrophobic transmembrane segment(s) is predicted to be localized at the cytoplasmic space; a protein that has any transmembrane segment(s) (including an N-terminal uncleavable segment) is predicted to be localized at the cellular membrane; and finally, a protein that has a cleavable N-terminal signal peptide but does not have any transmembrane segment(s) is predicted to be secreted to the extracellular space or to be localized at the cell wall. In Gram-positive bacteria, proteins that are anchored to the cell wall are characterized with the existence of the LPXTG-motif, followed by a hydrophobic domain and a tail of positively-charged residues (for recent review, see Siegel et al., 2017). On the other hand, Gram-negative bacteria contain one more membrane, the outer membrane, instead of the cell wall. Therefore, their possible localization sites are the cytoplasmic space, the inner membrane (which is equivalent to the membrane of Gram-positive bacteria), the periplasm, the outer membrane, and the extracellular space. Generally speaking, proteins that are localized at the latter three sites (the periplasm, the outer membrane, and the extracellular space) have an N-terminal cleavable signal peptide but do not have any hydrophobic transmembrane segment(s). Proteins that are integrated into the outer membrane are typically β-barrel proteins (Bakelar et al., 2017). To distinguish these three types of proteins, their difference in amino acid composition and/or *k*-mer frequency as well as motif/homology-based methods are often used.

A pioneering work to propose the above formalism is published in 1991 (Nakai and Kanehisa, 1991), where the predictor was named PSORT (I). In 2003, its approach was inherited and elaborated by Fiona Brinkman's group (Gardy et al., 2003); their software is named PSORTb (or PSORT-B). Its latest version is PSORTb 3.0 (Yu et al., 2010). The group published an excellent review of bacterial protein subcellular localization in 2006 (Gardy and Brinkman, 2006). According to the assessment shown in the review, PSORTb was the best predictor at that time. The group also releases PSORTdb, which contains a collection of experimentally-determined information of subcellular localization as well as systematic outputs of PSORTb applied to thousands of bacterial proteomes [its latest reference reports v. 3.0: (Peabody et al., 2016) but its latest version is v. 4.0]. The same group also proposes PSORTm, a variant of PSORTb designed for the prediction of metagenomic data (Peabody et al., 2020). The basic idea of PSORTm is to first identify the taxonomy of each read based on a reference database of microbial proteins. From the estimated taxonomy, the read is automatically classified with cell envelope types and then it is subject to a variant of PSORTb, which uses various types of analyses (such as motif/profile analysis) for its subcellular localization prediction. Although the assessment of its precise accuracy would be difficult, they report an assessment using artificial data and the comparison with the prediction against pre-assembled data. In view of the rapid growth of microbiome analyses, the need of characterizing metagenome data should increase even more and thus the field looks promising. Of course, other groups have developed a variety of predictors for bacterial/archaeal proteins, among which PSO-LocBact (Lertampaiporn et al., 2019), GPos-ECC-mPLoc/Gneg-ECC-mPLoc (Wang et al., 2015), BUSCA (Savojardo et al., 2018b), which will be introduced below, and ClubSub-P (Paramasivam and Linke, 2011) are released relatively recently. Some of them claim that they can deal with proteins with multiple-locations. Although once a database for (eukaryotic) proteins with multiple subcellular localizations is released (Zhang et al., 2008), it still seems difficult to classify multiple localizations objectively and quantitatively because the data come from different sources which rely on different experimental conditions (but see the discussion below).

Beyond the basic scheme described above, there are several issues to be further explored. One is the prediction of several specialized localization sites, such as host-associated, type III

secretion, fimbrial, flagellar, and spore. In PSORTb, they are treated as subcategories. Of course, it is favorable that a predictor can deal with such localization sites but it is questionable if such a predictor can also deal with artificial proteins that are transported to such locations. In other words, it is likely that such predictions are easily done with simple homology transfer from known examples. Another issue is how to deal with the proteins that are transported with minor pathways. For the users' convenience, it is desirable that a predictor can inform users which pathway the input protein will use. For example, it is surely useful if a predictor informs us that the input protein will be transported *via* the twin-arginine translocation pathway (Palmer and Stansfeld, 2020) or the lipoprotein signal peptidase II-dependent pathway (El Arnaout and Soulimane, 2019). This can already be done with several predictors, including SignalP-5.0 (Almagro Armenteros et al., 2019, see below). Hopefully, more knowledge of various protein sorting pathways should be incorporated into predictors, even if the objective assessment of their predictability would become difficult. In this sense, more benchmarking efforts/systematic analysis of subcellular localization from various viewpoints would be valuable (Stekhoven et al., 2014; Orioli and Vihinen, 2019; see below).

PREDICTION OF SUBCELLULAR LOCALIZATION SITES FOR EUKARYOTIC PROTEINS

So far, many prediction methods of eukaryotic protein subcellular localization have been developed. They are mainly based on biological/empirical sequence features related to subcellular localization. In these methods, a variety of machine learning algorithms, such as the *k*-nearest neighbor (*k*-NN) classifier, the Random Forest classifier, the support vector machine (SVM), and the deep learning, have been used. Those methods usually target 10 main localization sites, where subcompartments of localization sites are merged into 10 major sites in order to increase the number of proteins per localization site (see **Table 1**). As further explained below, for the prediction of subcellular localization sites, three types of prediction features are generally used: targeting signal features, sequence-based features, and annotation-based features (**Figure 1**). The features associated with targeting signals are most powerful, when available, and many subcellular localization predictors based on

TABLE 1 | Representative subcellular locations covered by predictors for eukaryotic proteins.

Main location	Representative subcompartments
Nucleus	inner and outer membranes, matrix, chromosome, nucleus speckle, etc.
Mitochondrion	inner and outer membranes, matrix, intermembrane space
Endoplasmic reticulum (ER)	ER membrane and lumen, microsome, rough ER, smooth ER, etc.
Plastid	inner and outer membranes, stroma, thylakoid, etc.
Golgi apparatus	Golgi apparatus membrane, lumen
Lysosome/Vacuole	vacuole lumen and membrane, lysosome lumen and membrane, etc.
Peroxisome	matrix, membrane
Cytoplasm	cytosol, cytoskeleton
Cell membrane	cell membrane, cell projection, apical, basal, etc.
Extracellular	–

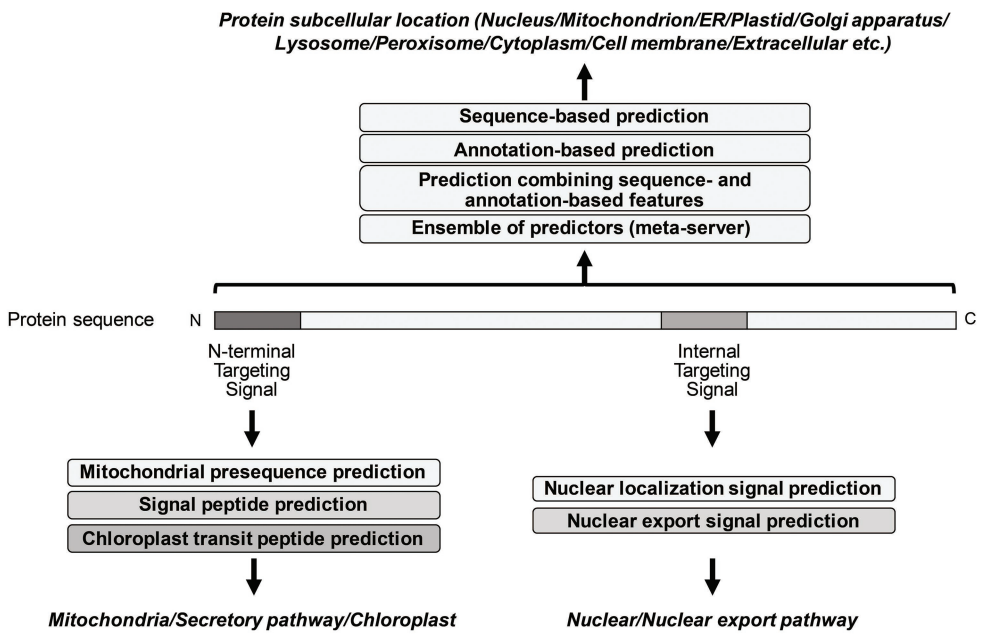

FIGURE 1 | Summary of representative prediction approaches of different subcellular localization.

targeting signal features have been developed. Thus, we first overview the representative targeting-signal predictors and then predictors for localization sites.

Prediction of Targeting Signals

The targeting signals are roughly grouped into two categories: N-terminal targeting signals and non-N-terminal targeting signals. The mitochondrial targeting signal (presequences), the signal sequence for the secretory pathway (signal peptides), and the transit signal for chloroplast (transit peptides) are well-known as N-terminal targeting signals, while the nuclear localization signal (NLS) and the nuclear export signal (NES) are internal signal sequences. Peroxisome matrix proteins contain peroxisomal targeting signal type 1 (PTS1) in the C-terminus.

Prediction of Mitochondrial Targeting Signal

Mitochondria have been estimated to host 1,000 to 1,500 distinct proteins. Approximately, 99% of mitochondrial proteins are encoded in the nuclear genome and are imported by translocases in the mitochondrial outer and inner membranes. Approximately 60% of mitochondrial proteins possess an N-terminal cleavable targeting signal (presequence; Vögtle et al., 2009). These presequences are typically recognized by the translocase of the outer membrane (TOM) receptors, which consist of Tom20 and Tom22, in the TOM complex. Then, they direct the translocation of signal-containing proteins through the main protein translocation channel, Tom40 (Pfanner et al., 2019). Upon translocation across the outer membrane, the presequence-containing proteins are transferred across the inner membrane by the translocase of the inner membrane complex (TIM23) with the presequence translocase-associated motor (PAM). The length of presequences is 20–60 amino acid residues (Calvo et al., 2017). The representative features of presequences are high and low composition of arginine residues and negatively-charged residues, respectively (von Heijne, 1986; Schneider et al., 1998). Positively charged amphiphilicity (amphiphilic α-helical structure with hydrophobic residues on one face and positively-charged residues on the opposite face) is also a well-characterized feature (Chacinska et al., 2009; Fukasawa et al., 2015). Recently, the TOM complex structure was revealed by cryo-electron microscopy and it provided structural insights into the import path of precursor protein containing presequence through the TOM complex (Araiso et al., 2019). Presequence is typically cleaved by three mitochondrial peptidases in the matrix (MPP, Icp55, and Oct1; Mossmann et al., 2012). The cleavage by MPP occurs after the position of two amino acids of C-terminal to an arginine (the R-2 motif). Icp55 and Oct1 subsequently cleave off one amino acid and eight amino acids from the newly-emerged N-terminus, respectively. Therefore, proteins processed by MPP and Icp55 have an arginine at position -3 (the R-3 motif) in the presequence, while proteins processed by MPP and Oct1 have an arginine at position -10 (the R-10 motif).

MitoProtII (Claros, 1995), TargetP (Emanuelsson et al., 2000), Predotar (Small et al., 2004), TPpred3.0 (Savojardo et al., 2015), and MitoFates (Fukasawa et al., 2015) were widely used presequence prediction methods. Those are developed using machine-learning techniques with these features of presequences. Those tools are also capable of predicting the existence of presequence as well as their cleavage site. MitoProtII and MitoFates are specific predictors for (mitochondrial) presequences, while TargetP, Predotar, and TPpred3.0 can also predict other N-terminal targeting signals, such as secretory signal sequence and chloroplastic targeting signal. Recently, TargetP2.0 is developed as a deep learning model, using bidirectional long-short-term memory (LSTM) and a multi-attention mechanism (Armenteros et al., 2019). Among existing tools, three of them (MitoFates, TPpred3.0, and TargetP2.0) perform better in the prediction of both the presequence existence and its cleavage site. MitoFates employs an SVM classifier by combining amino acid composition and physicochemical properties with positively charged amphiphilicity, discovered presequence motifs, and position-weight matrices of cleavage site patterns. TPpred3.0 is a combination of a Grammatical Restrained Hidden Conditional Random Field, N-to-1 Extreme Learning Machines, and SVMs. We compared the performance of those three methods, using recent proteomic data of the N-termini of mouse mitochondrial proteins (we omitted proteins whose length of cleaved N-terminal sequences is shorter than 10 or longer than 100 amino acids in the comparison; Calvo et al., 2017). The recalls of presequence prediction by TPpred3.0, MitoFates, and TargetP2.0 are 63.2, 75.9, and 79.9%, respectively. Whereas the recalls of the cleavage prediction by TPpred3.0, MitoFates, and TargetP2.0 are 27.0, 28.8, and 45.5%, respectively. MitoFates and TargetP2.0 show better performance on the presequence prediction. In the cleavage site prediction, TargetP2.0 far outperformed other methods, though the cleavage site prediction is still a challenging task. About 20% of mouse cleavage site data does not match with the R-2, R-3, and R-10 motifs (Calvo et al., 2017). It will be necessary to better characterize these untypical presequences.

Prediction of Signal Sequence

The targeting signal sequence for the secretory pathway (signal peptides) is located at the N-terminal of protein sequence in both eukaryotes and prokaryotes. The length of signal peptides is 16–30 amino acid residues. It is estimated that about 10–20% of eukaryotic proteome and 10% of bacterial proteome have the signal peptide at N-terminus (Kanapin et al., 2003; Ivankov et al., 2013). In eukaryotic cells, the signal recognition particle (SRP) co-translationally recognizes signal peptides upon their emergence from the ribosome and transfers them to the Sec61 translocon in the endoplasmic reticulum (ER) membrane *via* the SRP receptor (Nilsson et al., 2015). The signal peptidase cleaves off signal peptides and thus mature proteins are generated. Signal peptides share several characteristic features (von Heijne, 1990); they have tripartite architecture: a positively charged N-terminus (n-region), a hydrophobic segment (h-region), and a cleavage site for signal peptidase (c-region). The cleavage site is characterized by the (-1, -3) rule; amino acids with

small, uncharged side chains at the -1 and -3 position relative to the cleavage site.

For predicting signal peptides and their cleavage sites, many prediction methods, such as SignalP 4.0 (Petersen et al., 2011), SPEPlip (Fariselli et al., 2003), Phobius (Krogh et al., 2007), and DeepSig (Savojardo et al., 2018a), have been developed. The discrimination between secretory and non-secretory proteins based on the signal peptide prediction has been most successful in targeting signal predictions because SignalP 3.0 has already achieved the best Matthews' Correlation Coefficient (MCC) of 0.76 in eukaryotic data sets in an assessment study in 2009 (Choo et al., 2009). Recently, SignalP has been further improved as a deep neural network-based method, combining with conditional random field classification and optimized transfer learning (SignalP-5.0; Almagro Armenteros et al., 2019). According to their benchmark results, SignalP-5.0 outperforms other methods in predicting both the signal peptide existence and the cleavage site: the MCC was 0.88 in the signal peptide prediction and the recall of cleavage site detection was 72.9%.

Prediction of Chloroplastic Targeting Signal

The translocons at the outer and the inner membranes of chloroplasts, the TOC and TIC complexes mediate the targeting and import of ~3,500 different nuclear-encoded proteins. Those proteins are imported from the cytoplasm *via* interaction between their cleavable, N-terminal chloroplast targeting signal (transit peptides), and the TOC–TIC import systems (Li and Chiu, 2010; Paila et al., 2015). The transit peptide is removed off by the activity of stroma processing peptidase (SPP), which is related to the mitochondrial peptidase, MPP. SPP does not interact stably with the TOC–TIC import system, thus the cleavage event occurs after protein translocation or upon the emergence of the transit peptide cleavage site into the stroma. Chloroplast transit peptides are mostly unstructured but can form α-helical structures in hydrophobic environments (Bruce, 2001; Jarvis, 2008). In addition, chloroplast transit peptides have a high content of hydroxylated amino acids (e.g., serine residues) and positively charged amino acids and a very low content of negatively charged amino acids (Bhushan et al., 2006). Transit peptides and presequences are therefore similar in several aspects. In spite of the similarities, chloroplast transit peptides direct precursor proteins specifically to chloroplasts. Ge et al. (2014) demonstrated that transit peptides and presequences can be discriminated by their charge properties and hydrophobicity. Also, the analysis of 916 chloroplast proteins revealed an N-terminal domain beginning with Met-Ala and the low composition of arginine in the N-terminal portion (Zybailov et al., 2008). Moreover, Lee et al. (2019) recently showed that mitochondrial or chloroplast targeting specificities are characterized by the N-terminal regions of these targeting signals: an N-terminal multiple-arginine motif was identified as the mitochondrial specificity factor and chloroplast evasion signal. Cleavage sites of transit peptides are characterized by higher content of Ala, Ile, Cys, and Val residues (Gavel and von Heijne, 1990). The three motifs, [V,I][R,A]↓[A,C]AAE, S[V,I][R,S,V]↓[C,A]A, and [A,V]N↓A[A,M]AG[E,D], are derived by a set of 198 cleavage sites (Savojardo et al., 2015).

The existing prediction tools for the chloroplastic targeting signal deal with cleavable N-terminal transit peptides. Widely used prediction methods have been integrated as a part of prediction of N-terminal targeting signals in general: e.g., TargetP (Emanuelsson et al., 2000), iPSORT (Bannai et al., 2002), Predotar (Small et al., 2004), and TPpred3 (Savojardo et al., 2015). Among those tools, TPpred3 achieved better performance for transit peptide prediction (46% precision and 64% recall). As mentioned above, TargetP is recently updated to version 2.0 as a deep learning model (TargetP2.0; Armenteros et al., 2019). In their comparison, the precision and recall of chloroplastic transit peptide identification of TargetP2.0 are 90 and 86%, respectively, while those of TPpred3 are 76 and 69%. In the cleavage site prediction, the recalls of TargetP2.0 and TPpred3 are 49 and 30%, respectively. Like mitochondrial presequence prediction, the cleavage site prediction of chloroplastic targeting signal is a difficult problem. Comparing with the data size of signal peptides, that of transit peptides is quite small and thus the lower performance could have been caused by this reason. Larger-scale N-terminal proteomics data of chloroplastic proteins would be necessary for the improvement of their cleavage site prediction.

Prediction of Nuclear Localization Signals and Nuclear Export Signals

Nuclear proteins are transported into or out of the nuclei through the nuclear pore complex by the importin-β (Impβ) family nucleocytoplasmic transport receptors (Kimura and Imamoto, 2014). The human proteome contains 20 Impβ family proteins: 10 are nuclear import receptors (importin-β, transportin-1, -2, -SR, importin-4, -5 (RanBP5), -7, -8, -9 and -11), seven are export receptors (exportin-1 (CRM1), -2(CAS/CSE1L), -5, -6, -7, -t, and RanBP17), two are bi-directional receptors (imporin-13 and exportin-4), while the function of remaining RanBP6 is undetermined (Kimura and Imamoto, 2014). Those nucleocytoplasmic transport receptors are thought to recognize specific targeting signals on those cargo proteins. Several types of NLSs and NESs have been reported, so far. The most studied NLS is the classical NLS (cNLS) that binds to Impα, which is a cargo-binding adaptor exclusively used for Impβ (Lange et al., 2007). Sequences similar to the Impβ binding (IBB)-domain in Impα act as NLSs that bind directly to Impβ. Other known NLSs/NESs that bind directly to Impβ family are: the PY-NLS for Trn1 and Trn2 (Lee et al., 2006), the Leu-rich NES for CRM1 (Hutten and Kehlenbach, 2007), the SR-domain for TrnSR (Maertens et al., 2014), and the β-like importin binding (BIB)-domain, which binds to several nucleocytoplasmic transport receptors (Jäkel and Görlich, 1998). In addition, the RG/RGG-rich segment for Trn1 and the RSY-rich segment for TrnSR were reported recently (Bourgeois et al., 2020). However, these known NLSs/NESs do not explain all of the cargo recognition sites. Moreover, recent proteomic analysis for the identification of cargo proteins of 12 nucleocytoplasmic transport receptors (10 nuclear import

receptors and 2 bi-directional receptors; Kimura et al., 2017) also pointed out that about 30% of identified cargos are shared by multiple receptors. The degree of multiplicity and diversity of cargo recognition by nucleocytoplasmic transport receptors are still controversial.

Among known nuclear targeting signals, cNLS and NES of CRM1 are well characterized. Thus, existing prediction methods of NLSs and NESs mainly target these two types. cNLSs are grouped into monopartite and bipartite NLSs. Monopartite NLS is characterized with a single stretch of basic residues (e.g., KR[K/R]R and K[K/R]RK), while bipartite NLS has two clusters of basic residues, separated by a spacer region of 10–12 amino acids (e.g., $KRX_{10-12}K[K/R][K/R]$; Kosugi et al., 2009). Lisitsyna et al. (2017) assessed the prediction performance of widely used methods, Nucpred (Brameier et al., 2007), cNLSmapper (Kosugi et al., 2008a), NLStradamus (Ba et al., 2009), NucImport (Mehdi et al., 2011), and SeqNLS (Lin and Hu, 2013), using a human NLS dataset (Lisitsyna et al., 2017). NucPred, seqNLS, and NLStradamus showed better MCCs (~0.3); however, the recalls of those methods were still ~45%. Recently, Guo et al. (2020) reported INSP, which is a NLS predictor based on a multivariate regression model integrating PSSM-based conservation score, protein language-based SVM learning score, disorder-based structural score, and amino acid physical chemistry property-based score. On their test dataset, INSP showed 50.6% precision at 67.0% recall, whereas seqNLS, NLStradamus, and cNLSmapper obtained 60.6% precision at 36.4% recall, 53.9% precision at 35.6% recall, and 50.9% precision at 50.9% recall, respectively. INSP showed a favorable balance between the prediction precision and recall, but NLS prediction seems to be still difficult because the cNLS sequence patterns are often observed in non-nuclear protein sequences (i.e., false positives).

Nuclear export signals function as essential regulators for the export of hundreds of distinct cargo proteins by interacting with CRM1. So far, 11 consensus patterns of NES have been proposed by a peptide-library study and structure analyses of CRM1-NES (Kosugi et al., 2008b; Fung et al., 2015, 2017). In general, NESs are represented by $\Phi 0\text{-}x_{1\text{-}2}\text{-}\Phi 1\text{-}(x)_{2\text{-}3}\text{-}\Phi 2\text{-}(x)_{2\text{-}3}\text{-}\Phi 3\text{-}x\text{-}\Phi 4$ (Φ1-4 denote Leu, Val, Ile, Phe, and Met while x is any amino acid. Φ0 is not restricted to the hydrophobic amino acids). Those hydrophobic residues in Φ0–Φ4 are bound to the corresponding hydrophobic pockets in CRM1. Based on the pattern of these Φ's and spacing sequences, the NES motifs are classified into seven classes and four additional reverse classes, representing binding in the opposite direction. Several prediction tools for NESs, such as NetNES (La Cour et al., 2004), NESsential (Fu et al., 2011), NESmapper (Kosugi et al., 2014), Wregex (Prieto et al., 2014), LocNES (Xu et al., 2015), and NoLogo (Liku et al., 2018) have been developed, representing the consensus sequences with regular expressions or PSSMs as well as biophysical properties (disorder propensity, solvent accessibility, and secondary structure information). Among those tools, LocNES outperformed other prediction tools; however, the precision is ~50% at 20% recall. The low performance is caused by high false-positive rates. As mentioned above, the NES consensus patterns are simple and commonly observed in other protein sequences. Thus, it seems to be difficult to improve the prediction performance by only using the sequence information. Recently, Lee et al. (2019) provided a comprehensive table for cargo proteins, containing the location of the NES motifs with the disordered propensity, the predicted secondary structures, and the conserved domain information. They also proposed a structure modeling-based prediction which predicts the binding energy of the NES peptide bound to the binding groove of CRM1, using multiple structures of CRM1-NES peptide complex as templates (Lee et al., 2019). The structure-based methods performed at the same level as LocNES in recall rate but outperformed LocNES in specificity and false-positive rate. Thus, combining sequence-based and structure-based predictions seems promising in significantly improving the NES prediction. Moreover, NLSdb, which is a database containing NLSs and NESs, has been recently updated (Bernhofer et al., 2018). In this update, the potential set of novel NLSs and NESs has been generated by an *in silico* mutagenesis protocol. Then, the potential NLSs and NESs match at least one nuclear protein but do not match any non-nuclear proteins. The updated NLSdb contains 2,253 NLSs (1,614 are potential NLSs) and 398 NESs (192 are potential NESs). The data would be useful to further improve the NLS and NES prediction performances.

Prediction of Subcellular Localization Site of Protein in a Cell

Existing methods for predicting subcellular localization sites can be grouped into four categories. The first category of prediction methods uses only sequence-based features. Some sequence-based features are used in localization site prediction because their differences are empirically known to be correlated with the differences between localization sites. Such empirical features include the frequency of dipeptides, *n*-grams, and *k*-mers as well as the pseudo amino acid composition of the entire amino acid sequence (or that of predicted mature sequence). Pseudo amino acid composition is more informative in terms of incorporating sequence-order information of a protein sequence (Chou, 2001). These empirical sequence-based features have also been popular in various amino acid sequence-based predictions. Besides these systematically defined features, sequence features of various known targeting signals are more or less useful, as mentioned above. Functional motifs are also used in the prediction because sequence motifs associated with the function of a protein are closely related to its localization site (for example, a protein containing a DNA-binding motif is likely to be localized in the nucleus). The first sequence-based method was PSORT (I) (Nakai and Kanehisa, 1992), which was developed about 30 years ago, and later many other methods, such as WoLF PSORT (Horton et al., 2007), CELLO2.5 (Yu et al., 2006), and DeepLoc (Almagro Armenteros et al., 2017), have been developed. WoLF PSORT is an update of PSORT II (Horton and Nakai, 1997), which converts the input amino acid sequences into a numerical vector consisting of amino acid composition and PSORT/iPSORT (Nakai and Kanehisa, 1992; Bannai et al., 2002)

localization features, and then classifies proteins into subcellular locations with a weighted *k*-NN classifier. CELLO2.5 is a two-level SVM classifier system: the first level comprises a number of SVM classifiers, each based on distinctive sets of feature vectors generated from amino acid sequence data, and the second level SVM classifier functions as the jury machine to generate the probability distribution of decisions for possible localizations. Recently, several deep learning-based predictors are developed. DeepLoc is their representative. DeepLoc uses recurrent neural networks (RNNs) with long short-term memory (LSTM) cells that process the entire amino acid sequence and an attention mechanism identifying sequence regions important for the subcellular localization.

The second category of predictors uses annotation-based features obtained from experimental evidence. GO terms, localization annotation in UniProt, functional domain, protein-protein interaction, and literature information from PubMed abstracts are categorized into this type of features. mGOASVM (Wan et al., 2012) is a predictor for the subcellular localization of multi-location proteins based on GO-terms. In mGOASVM, multi-label GO vectors, which are the occurrences of GO terms of homologous proteins, are constructed, and then GO vectors are recognized by SVM classifiers equipped with a decision strategy that can produce multiple-class labels for a query protein. pLoc-mEuk (Cheng et al., 2018) is recently developed by extracting the key GO information into "Chou's general Pseudo Amino Acid Composition." pLoc-mEuk can also deal with proteins with multiple locations. Generally speaking, however, compared with those features, the transfer of localization annotation from homologous protein seems to be simpler and more useful. We previously pointed out that a simple homology-based inference outperforms methods based on machine learning if a homologous protein with localization annotation is available (Imai and Nakai, 2010).

The third category is the predictors combining sequence-based and annotation-based features, such as MultiLoc2 (Blum et al., 2009), SherLoc2 (Briesemeister et al., 2009), YLoc (Briesemeister et al., 2010), and LocTree3 (Goldberg et al., 2014). MultiLoc2 utilizes an SVM predictor, MultiLoc (Höglund et al., 2006), which is based on overall amino acids and the presence of known sorting signals, combined with phylogenetic profiles and GO terms. SherLoc2 combines MultiLoc2 and EpiLoc (Brady and Shatkay, 2008), a prediction system based on features derived from PubMed abstracts. YLoc is based on a simple naive Bayes classifier, which combines various features ranging from simple amino acid composition to annotation information, like PROSITE domains, and GO terms from close homologs. LocTree3 improves over a machine learning-based predictor, LocTree2 (Goldberg et al., 2012), by the combination of the machine learning-based method with a homology-based inference transfer through PSI-BLAST.

The fourth category is the ensemble of several prediction methods (meta-servers), which collects prediction scores of several predictors, and then they are trained by a machine learning technique, such as the Random Forest classifier and SVM. SubCons (Salvatore et al., 2017) is a recent ensemble method, which combines four predictors (CELLO2.5, LocTree2, MultiLoc2, and SherLoc2) using a Random Forest classifier. BUSCA also integrates different prediction methods. Prediction pipeline of BUSCA consists of predictors for targeting signals [DeepSig (Savojardo et al., 2018a) and TPpred3 (Savojardo et al., 2015)], for GPI-anchors [PredGPI (Pierleoni et al., 2008)], for transmembrane domains [ENSEMBLE3.0 (Martelli et al., 2003) and BetAware (Savojardo et al., 2013)], and for discriminators of subcellular localization of both globular and membrane proteins [BaCelLo (Pierleoni et al., 2006), MemLoci (Pierleoni et al., 2011), and SChloro (Savojardo et al., 2017)].

Recent Benchmarks for Subcellular Localization Prediction

Evaluation of prediction performance of subcellular localization prediction is often difficult due to the following reasons: (i) There are often overlaps between their own training data and the test data of different methods. In those cases, the performances could be overestimated. (ii) Comparison of sequence-based methods with annotation-based methods or methods combining sequence- and annotation-based methods tends to be unfair. For example, the measured accuracy of annotation-based methods would become apparently higher if the majority of test data used for sequence-based methods are included in the databases used for the prediction by annotation-based methods.

To evaluate the prediction performance with less bias, Salvatore et al. recently made a benchmark dataset which consists of proteins containing identical subcellular annotations in at least two out of the three resources (Salvatore et al., 2017): two large-scale study data on subcellular localization of human proteins (Uhlen et al., 2010; Fagerberg et al., 2011; Breckels et al., 2013; Christoforou et al., 2014) and proteins with "manually curated" annotation of subcellular localization in UniProt (UniProt Consortium, 2019). Then, they examined the performance of six state-of-the-art methods [CELLO2.5 (Yu et al., 2006), LocTree2 (Goldberg et al., 2012), MultiLoc2 (Blum et al., 2009), SherLoc2 (Briesemeister et al., 2009), WoLF PSORT (Horton et al., 2007), and YLoc (Briesemeister et al., 2010)] as well as SubCons (Salvatore et al., 2017) for eight localization sites (nucleus, mitochondria, ER, Golgi apparatus, lysosome, peroxisome, plasma membrane, and cytoplasm). They used the Generalized Squared Correlation (GC^2; Baldi et al., 2000) for performance evaluation. GC^2 is a subtype of Gorodkin measure (Gorodkin, 2004), which can be seen as a generalization of MCC that applies to *K*-categories. The Gorodkin measure is more informative than the accuracy measure when there is an imbalance of classes. For $K = 2$, the Gorodkin measure squared is GC^2. In this assessment, SubCons showed the best overall prediction performance, $GC^2 = 0.32$, and the second best was SherLoc2 ($GC^2 = 0.27$). On the other hand, during the development of DeepLoc (Almagro Armenteros et al., 2017), the authors made an independent test set by performing a stringent homology partitioning against experimentally annotated protein data in UniProt. Homologous proteins that fulfill a certain threshold of similarity were clustered, and then each

cluster of homologous proteins was assigned to one of the five folds, ensuring that similar proteins were not mixed between the different folds. Four were used for the training and validation while the remaining one for testing. Using the test set, they compared the prediction performance of DeepLoc with the above six methods (CELLO2.5, LocTree2, MultiLoc2, SherLoc2, WoLF PSORT, and YLoc) and iLoc-Euk (Chou et al., 2011) in 10 localization sites (extracellular and plastid are added into the above eight localization sites). DeepLoc showed the best Gorodkin measure of 0.735, and the second and third best were achieved by iLoc-Euk at 0.682 and YLoc at 0.533, respectively.

Although efforts to evaluate the prediction performance with less bias have been made, more efforts seem to be necessary. According to recent benchmarking reports based on human data sets and membrane proteins (Orioli and Vihinen, 2019; Shen et al., 2020), sequence-based methods tend to show lower performance than annotation-based methods, including meta methods. However, a certain number of proteins (or their highly homologous ones) in the benchmark test data seem to be included in the database used in annotation-based methods. In addition, methods trained and tested with newly constructed data tend to show better performance because older data tend to include more mislabeled or questionable examples. Indeed, Almagro Armenteros et al. (2017) pointed out a considerable decrease of experimentally confirmed proteins in UniProt after a major change in the annotation standards on release 2014_09. The prediction performances of machine learning algorithms significantly depend on the datasets used. Some of the previously developed methods may outperform newer methods when they are trained and tested with the latest datasets. For fair assessments, performance comparison should therefore be done in each category with standardized benchmark data sets, ensuring independence between training and test data sets. Unfortunately, to the best of our knowledge, such standardized benchmark data sets have not been constructed so far. The data sets used in previous studies are often used in the development of novel methods. The standardization of prediction performance comparison is a big challenge but this is essential and important in this field. Recent progress in proteome-wide subcellular protein mapping (see below) would provide substantial information on the subcellular localization of unverified or unseen proteins as well as the information for correcting mislabeled proteins, which should be helpful in constructing standardized benchmark data sets, obviously.

PROTEIN LOCALIZATION RESOURCES OBTAINED FROM RECENT SPATIAL PROTEOMICS APPROACHES

Proteomics data for capturing the spatial distribution of proteins at the subcellular level (subcellular protein mapping) are useful resources for their predictive studies. Recent advances in high-throughput microscopy, quantitative mass spectrometry (MS), interactome mapping, and machine learning applications for data analysis have enabled proteome-wide subcellular protein mapping (Lundberg and Borner, 2019; Borner, 2020). Three experimental approaches are generally used for spatial proteomics: proteome-wide imaging of protein localization, protein–protein interaction network analysis, and MS-based organelle profiling. All of these approaches have produced numerous available data of human protein subcellular localization. The Human Cell Atlas provides an invaluable resource of imaging data at a single-cell level (localization of 12,003 proteins; Thul et al., 2017). The global organellar map based on biotin identification (BioID) data is now available as a resource of protein–protein interaction network analysis (4,145 proteins; Go et al., 2019). Several organelle profiling resources are obtained from fibroblasts (2,533 proteins; Jean Beltran et al., 2016) and cell lines: HeLa (8,710 proteins; Itzhak et al., 2016), five different cancer cell lines (12,418 proteins; Orre et al., 2019), and U-2 OS (2,412 proteins; Geladaki et al., 2019). In addition, organelle profiling resources of mouse primary neurons (Itzhak et al., 2017), mouse liver (Krahmer et al., 2018), mouse pluripotent stem cell (Christoforou et al., 2016), rat liver (Jadot et al., 2017), and *Saccharomyces cerevisiae* (Nightingale et al., 2019) are also available.

Each of these approaches has its own merits for the identification of protein localization: the imaging approach provides multiple localizations and has a single-cell resolution while the MS-based approach can provide peptide-level resolution and reveal the differential localization of splicing isoforms, proteolytically processed forms, and the isoforms *via* differential post-translational modifications. A recent imaging-based large-scale study reports that about a half of all proteins are localized at multiple compartments, suggesting that there is a shared pool of proteins even among functionally unrelated organelles (Thul et al., 2017). Prediction of proteins that exist in two or more subcellular location sites is an important issue for understanding the biological process in a cell. A recent review summarizes the prediction methods that can deal with proteins with multiple locations (Chou, 2019).

A number of differentially localized isoform pairs were found by MS-based approaches (Christoforou et al., 2016; Geladaki et al., 2019). Such localization change at the isoform level is an interesting issue in terms of targeting signal usage. Protein isoforms seem to be generated by a stress response or in a tissue-specific manner. Thus, a number of localization changes at the isoform level may have been unidentified still. For mitochondrial proteins, we previously applied MitoFates to search for differentially-localized candidates of isoforms and obtained 517 genes, which were 44% of the predicted mitochondrial genes (Fukasawa et al., 2015), suggesting that the major localization changes of mitochondrial protein isoforms are regulated by the changes in their N-terminal targeting signal. Recently, relative protein levels of more than 12,000 genes across 32 normal human tissues were quantified and tissue-specific or tissue-enriched proteins were identified (Jiang et al., 2020). Also, they identified a total of 2,436 tissue-enriched protein isoforms. Those isoforms may be useful for the investigation of tissue-specific localization changes at the isoform level.

Multiple localization proteins and localization changes among isoforms imply potential "moonlighting" activity. Comprehensive analyses of these proteins should boost our further understanding in cell biology.

CONCLUSION

A number of computational tools for the analyses of protein subcellular localization are introduced in this review. Although many of the localization sites of a given protein would be able to be predicted through a mere homology transfer nowadays, we would like to emphasize that the subcellular localization prediction problem is not a pedantic one at all. The authors believe that the *in silico* accumulation of various knowledge on protein sorting/targeting processes is important. Prediction methods can be used for assessing how much we understand these processes quantitatively. The future methods should be useful for various purposes, such as for the evaluation of artificial proteins, for understanding why some proteins are localized at multiple positions and for inferring how tissue-specific and/or condition-specific isoforms can change their localization sites. Therefore, in our opinion, the knowledge-based approach would be most important in the future of this field and such knowledge should be integrated into the wider knowledge on the *in vivo* fate of proteins since all of the processes are interrelated with each other (Nakai, 2001).

AUTHOR CONTRIBUTIONS

Both the authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

KI acknowledges support from JSPS KAKENHI (grant number 18K11543).

REFERENCES

Almagro Armenteros, J. J., Sønderby, C. K., Sønderby, S. K., Nielsen, H., and Winther, O. (2017). DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics* 33, 3387–3395. doi: 10.1093/bioinformatics/btx431

Almagro Armenteros, J. J., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., et al. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* 37, 420–423. doi: 10.1038/s41587-019-0036-z

Araiso, Y., Tsutsumi, A., Qiu, J., Imai, K., Shiota, T., Song, J., et al. (2019). Structure of the mitochondrial import gate reveals distinct preprotein paths. *Nature* 575, 395–401. doi: 10.1038/s41586-019-1680-7

Armenteros, J. J. A., Salvatore, M., Emanuelsson, O., Winther, O., von Heijne, G., Elofsson, A., et al. (2019). Detecting sequence signals in targeting peptides using deep learning. *Life Sci. Alliance* 2:e201900429. doi: 10.26508/lsa.201900429

Audagnotto, M., and Dal Peraro, M. (2017). Protein post-translational modifications: in silico prediction tools and molecular modeling. *Comput. Struct. Biotechnol. J.* 15, 307–319. doi: 10.1016/j.csbj.2017.03.004

Ba, A. N. N., Pogoutse, A., Provart, N., and Moses, A. M. (2009). NLStradamus: a simple Hidden Markov Model for nuclear localization signal prediction. *BMC Bioinformatics* 10:202. doi: 10.1186/1471-2105-10-202

Bakelar, J., Buchanan, S. K., and Noinaj, N. (2017). Structural snapshots of the β-barrel assembly machinery. *FEBS J.* 284, 1778–1786. doi: 10.1111/febs.13960

Baker, D. (2019). What has de novo protein design taught us about protein folding and biophysics? *Protein Sci.* 28, 678–683. doi: 10.1002/pro.3588

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F., and Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* 16, 412–424. doi: 10.1093/bioinformatics/16.5.412

Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., and Miyano, S. (2002). Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics* 18, 298–305. doi: 10.1093/bioinformatics/18.2.298

Bernhofer, M., Goldberg, T., Wolf, S., Ahmed, M., Zaugg, J., Boden, M., et al. (2018). NLSdb-major update for database of nuclear localization signals and nuclear export signals. *Nucleic Acids Res.* 46, D503–D508. doi: 10.1093/nar/gkx1021

Bhushan, S., Kuhn, C., Berglund, A. K., Roth, C., and Glaser, E. (2006). The role of the N-terminal domain of chloroplast targeting peptides in organellar protein import and miss-sorting. *FEBS Lett.* 580, 3966–3972. doi: 10.1016/j.febslet.2006.06.018

Blum, T., Briesemeister, S., and Kohlbacher, O. (2009). MultiLoc2: integrating phylogeny and gene ontology terms improves subcellular protein localization prediction. *BMC Bioinformatics* 10:274. doi: 10.1186/1471-2105-10-274

Borner, G. H. H. (2020). Organellar maps through proteomic profiling - a conceptual guide. *Mol. Cell. Proteomics* 19, 1076–1087. doi: 10.1074/mcp.R120.001971

Bourgeois, B., Hutten, S., Gottschalk, B., Hofweber, M., Richter, G., and Sternat, J. (2020). Nonclassical nuclear localization signals mediate nuclear import of CIRBP. *Proc. Natl. Acad. Sci. U. S. A.* 117, 8503–8514. doi: 10.1073/pnas.1918944117

Brady, S., and Shatkay, H. (2008). EPILOC: a (working) text-based system for predicting protein subcellular location. *Pac. Symp. Biocomput.* 13, 604–615. doi: 10.1142/9789812776136_0058

Brameier, M., Krings, A., and MacCallum, R. M. (2007). NucPred — predicting nuclear localization of proteins. *Bioinformatics* 23, 1159–1160. doi: 10.1093/bioinformatics/btm066

Breckels, L. M., Gatto, L., Christoforou, A., Groen, A. J., Lilley, K. S., and Trotter, M. W. B. (2013). The effect of organelle discovery upon sub-cellular protein localisation. *J. Proteomics* 88, 129–140. doi: 10.1016/j.jprot.2013.02.019

Briesemeister, S., Blum, T., Brady, S., Lam, Y., Kohlbacher, O., and Shatkay, H. (2009). SherLoc2: a high-accuracy hybrid method for predicting subcellular localization of proteins. *J. Proteome Res.* 8, 5363–5366. doi: 10.1021/pr900665y

Briesemeister, S., Rahnenführer, J., and Kohlbacher, O. (2010). Going from where to why-interpretable prediction of protein subcellular localization. *Bioinformatics* 26, 1232–1238. doi: 10.1093/bioinformatics/btq115

Bruce, B. D. (2001). The paradox of plastid transit peptides: conservation of function despite divergence in primary structure. *Biochim. Biophys. Acta* 1541, 2–21. doi: 10.1016/s0167-4889(01)00149-5

Calvo, S. E., Julien, O., Clauser, K. R., Shen, H., Kamer, K. J., Wells, J. A., et al. (2017). Comparative analysis of mitochondrial N-termini from mouse, human, and yeast. *Mol. Cell. Proteomics* 16, 512–523. doi: 10.1074/mcp.M116.063818

Chacinska, A., Koehler, C. M., Milenkovic, D., Lithgow, T., and Pfanner, N. (2009). Importing mitochondrial proteins: machineries and mechanisms. *Cell* 138, 628–644. doi: 10.1016/j.cell.2009.08.005

Cheng, X., Xiao, X., and Chou, K. C. (2018). pLoc-mEuk: predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC. *Genomics* 110, 50–58. doi: 10.1016/j.ygeno.2017.08.005

Choo, K. H., Tan, T. W., and Ranganathan, S. (2009). A comprehensive assessment of N-terminal signal peptides prediction methods. *BMC Bioinformatics* 10:S2. doi: 10.1186/1471-2105-10-S15-S2

Chou, K. C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins Struct. Funct. Genet.* 43, 246–255. doi: 10.1002/prot.1035

Chou, K. C. (2019). Advances in predicting subcellular localization of multi-label proteins and its implication for developing multi-target drugs. *Curr. Med. Chem.* 26, 4918–4943. doi: 10.2174/0929867326666190507082559

Chou, K. C., Wu, Z. C., and Xiao, X. (2011). iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. *PLoS One* 6:e18258. doi: 10.1371/journal.pone.0018258

Christoforou, A., Arias, A. M., and Lilley, K. S. (2014). Determining protein subcellular localization in mammalian cell culture with biochemical fractionation and iTRAQ 8-plex quantification. *Methods Mol. Biol.* 1156, 157–174. doi: 10.1007/978-1-4939-0685-7_10

Christoforou, A., Mulvey, C. M., Breckels, L. M., Geladaki, A., Hurrell, T., Hayward, P. C., et al. (2016). A draft map of the mouse pluripotent stem cell spatial proteome. *Nat. Commun.* 7:8992. doi: 10.1038/ncomms9992

Claros, M. G. (1995). MitoProt, a macintosh application for studying mitochondrial proteins. *Bioinformatics* 11, 441–447. doi: 10.1093/bioinformatics/11.4.441

Du, P., and Xu, C. (2013). Predicting multisite protein subcellular locations: progress and challenges. *Expert Rev. Proteomics* 10, 227–237. doi: 10.1586/epr.13.16

El Arnaout, T., and Soulimane, T. (2019). Targeting lipoprotein biogenesis: considerations towards antimicrobials. *Trends Biochem. Sci.* 44, 701–715. doi: 10.1016/j.tibs.2019.03.007

Eldeeb, M. A., Siva-Piragasam, R., Ragheb, M. A., Esmaili, M., Salla, M., and Fahlman, R. P. (2019). A molecular toolbox for studying protein degradation in mammalian cells. *J. Neurochem.* 151, 520–533. doi: 10.1111/jnc.14838

Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* 300, 1005–1016. doi: 10.1006/jmbi.2000.3903

Fagerberg, L., Stadler, C., Skogs, M., Hjelmare, M., Jonasson, K., Wiking, M., et al. (2011). Mapping the subcellular protein distribution in three human cell lines. *J. Proteome Res.* 10, 3766–3777. doi: 10.1021/pr200379a

Fariselli, P., Finocchiaro, G., and Casadio, R. (2003). SPEPlip: the detection of signal peptide and lipoprotein cleavage sites. *Bioinformatics* 19, 2498–2499. doi: 10.1093/bioinformatics/btg360

Fu, S. C., Imai, K., and Horton, P. (2011). Prediction of leucine-rich nuclear export signal containing proteins with NESsential. *Nucleic Acids Res.* 39:e111. doi: 10.1093/nar/gkr493

Fukasawa, Y., Tsuji, J., Fu, S. C., Tomii, K., Horton, P., and Imai, K. (2015). MitoFates: improved prediction of mitochondrial targeting sequences and their cleavage sites. *Mol. Cell. Proteomics* 14, 1113–1126. doi: 10.1074/mcp.M114.043083

Fung, H. Y. J., Fu, S. C., Brautigam, C. A., and Chook, Y. M. (2015). Structural determinants of nuclear export signal orientation in binding to exportin CRM1. *eLife* 4:e10034. doi: 10.7554/eLife.10034

Fung, H. Y. J., Fu, S. C., and Chook, Y. M. (2017). Nuclear export receptor CRM1 recognizes diverse conformations in nuclear export signals. *eLife* 6:e23961. doi: 10.7554/eLife.23961

Gardy, J. L., and Brinkman, F. S. L. (2006). Methods for predicting bacterial protein subcellular localization. *Nat. Rev. Microbiol.* 4, 741–751. doi: 10.1038/nrmicro1494

Gardy, J. L., Spencer, C., Wang, K., Ester, M., Tusnády, G. E., Simon, I., et al. (2003). PSORT-B: improving protein subcellular localization prediction for gram-negative bacteria. *Nucleic Acids Res.* 31, 3613–3617. doi: 10.1093/nar/gkg602

Gavel, Y., and von Heijne, G. (1990). A conserved cleavage-site motif in chloroplast transit peptides. *FEBS Lett.* 261, 455–458. doi: 10.1016/0014-5793(90)80614-O

Ge, C., Spånning, E., Glaser, E., and Wieslander, Å. (2014). Import determinants of organelle-specific and dual targeting peptides of mitochondria and chloroplasts in Arabidopsis thaliana. *Mol. Plant* 7, 121–136. doi: 10.1093/mp/sst148

Geladaki, A., Kočevar Britovšek, N., Breckels, L. M., Smith, T. S., Vennard, O. L., Mulvey, C. M., et al. (2019). Combining LOPIT with differential ultracentrifugation for high-resolution spatial proteomics. *Nat. Commun.* 10:331. doi: 10.1038/s41467-018-08191-w

Go, C., Knight, J., Rajasekharan, A., Rathod, B., Hesketh, G., Abe, K., et al. (2019). A proximity biotinylation map of a human cell. bioRxiv [Preprint]. doi: 10.1101/796391

Goldberg, T., Hamp, T., and Rost, B. (2012). LocTree2 predicts localization for all domains of life. *Bioinformatics* 28, i458–i465. doi: 10.1093/bioinformatics/bts390

Goldberg, T., Hecht, M., Hamp, T., Karl, T., Yachdav, G., Ahmed, N., et al. (2014). LocTree3 prediction of localization. *Nucleic Acids Res.* 42, W350–W355. doi: 10.1093/nar/gku396

Gorodkin, J. (2004). Comparing two K-category assignments by a K-category correlation coefficient. *Comput. Biol. Chem.* 28, 367–374. doi: 10.1016/j.compbiolchem.2004.09.006

Guo, Y., Yang, Y., Huang, Y., and Shen, H. B. (2020). Discovering nuclear targeting signal sequence through protein language learning and multivariate analysis. *Anal. Biochem.* 591:113565. doi: 10.1016/j.ab.2019.113565

Harvey Millar, A., and Taylor, N. L. (2014). Subcellular proteomics-where cell biology meets protein chemistry. *Front. Plant Sci.* 5:55. doi: 10.3389/fpls.2014.00055

Höglund, A., Dönnes, P., Blum, T., Adolph, H. W., and Kohlbacher, O. (2006). MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinformatics* 22, 1158–1165. doi: 10.1093/bioinformatics/btl002

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

Horton, P., Park, K. J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C. J., et al. (2007). WoLF PSORT: protein localization predictor. *Nucleic Acids Res.* 35, W585–W587. doi: 10.1093/nar/gkm259

Hutten, S., and Kehlenbach, R. H. (2007). CRM1-mediated nuclear export: to the pore and beyond. *Trends Cell Biol.* 17, 193–201. doi: 10.1016/j.tcb.2007.02.003

Imai, K., and Nakai, K. (2010). Prediction of subcellular locations of proteins: where to proceed? *Proteomics* 10, 3970–3983. doi: 10.1002/pmic.201000274

Imai, K., and Nakai, K. (2019). "Prediction of protein localization" in *Encyclopedia of Bioinformatics and Computational Biology*, *Vol. 2*. Elsevier, 53–59.

Itzhak, D. N., Davies, C., Tyanova, S., Mishra, A., Williamson, J., Antrobus, R., et al. (2017). A mass spectrometry-based approach for mapping protein subcellular localization reveals the spatial proteome of mouse primary neurons. *Cell Rep.* 20, 2706–2718. doi: 10.1016/j.celrep.2017.08.063

Itzhak, D. N., Tyanova, S., Cox, J., and Borner, G. H. H. (2016). Global, quantitative and dynamic mapping of protein subcellular localization. *eLife* 5:e16950. doi: 10.7554/eLife.16950

Ivankov, D. N., Payne, S. H., Galperin, M. Y., Bonissone, S., Pevzner, P. A., and Frishman, D. (2013). How many signal peptides are there in bacteria? *Environ. Microbiol.* 15, 983–990. doi: 10.1111/1462-2920.12105

Jadot, M., Boonen, M., Thirion, J., Wang, N., Xing, J., Zhao, C., et al. (2017). Accounting for protein subcellular localization: a compartmental map of the rat liver proteome. *Mol. Cell. Proteomics* 16, 194–212. doi: 10.1074/mcp.M116.064527

Jäkel, S., and Görlich, D. (1998). Importin β, transportin, RanBP5 and RanBP7 mediate nuclear import of ribosomal proteins in mammalian cells. *EMBO J.* 17, 4491–4502. doi: 10.1093/emboj/17.15.4491

Jarvis, P. (2008). Targeting of nucleus-encoded proteins to chloroplasts in plants. *New Phytol.* 179, 257–285. doi: 10.1111/j.1469-8137.2008.02452.x

Jean Beltran, P. M., Mathias, R. A., and Cristea, I. M. (2016). A portrait of the human organelle proteome in space and time during cytomegalovirus infection. *Cell Syst.* 3, 361–373.e6. doi: 10.1016/j.cels.2016.08.012

Jiang, L., Wang, M., Lin, S., Jian, R., Li, X., Chan, J., et al. (2020). A quantitative proteome map of the human body. *Cell* 183, 269–283.e19. doi: 10.1016/j.cell.2020.08.036

Kanapin, A., Batalov, S., Davis, M. J., Gough, J., Grimmond, S., Kawaji, H., et al. (2003). Mouse proteome analysis. *Genome Res.* 13, 1335–1344. doi: 10.1101/gr.978703

Kimura, M., and Imamoto, N. (2014). Biological significance of the importin-β family-dependent nucleocytoplasmic transport. *Traffic* 15, 727–748. doi: 10.1111/tra.12174

Kimura, M., Morinaka, Y., Imai, K., and Kose, S. (2017). Extensive cargo identification reveals distinct biological roles of the 12 importin pathways. *eLife* 6:e21184. doi: 10.7554/eLife.21184

Kosugi, S., Hasebe, M., Entani, T., Takayama, S., Tomita, M., and Yanagawa, H. (2008a). Article design of peptide inhibitors for the importin α/β nuclear import pathway by activity-based profiling. *Chem. Biol.* 15, 940–949. doi: 10.1016/j.chembiol.2008.07.019

Kosugi, S., Hasebe, M., Matsumura, N., Takashima, H., Miyamoto-sato, E., Tomita, M., et al. (2009). Six classes of nuclear localization signals specific

to different binding grooves of importin α. *J. Biol. Chem.* 284, 478–485. doi: 10.1074/jbc.M807017200

Kosugi, S., Hasebe, M., Tomita, M., and Yanagawa, H. (2008b). Nuclear export signal consensus sequences defined using a localization-based yeast selection system. *Traffic* 9, 2053–2062. doi: 10.1111/j.1600-0854.2008.00825.x

Kosugi, S., Yanagawa, H., Terauchi, R., and Tabata, S. (2014). NESmapper: accurate prediction of leucine-rich nuclear export signals using activity-based profiles. *PLoS Comput. Biol.* 10:e1003841. doi: 10.1371/journal.pcbi.1003841

Krahmer, N., Najafi, B., Schueder, F., Quagliarini, F., Steger, M., Seitz, S., et al. (2018). Organellar proteomics and Phospho-proteomics reveal subcellular reorganization in diet-induced hepatic steatosis. *Dev. Cell* 47, 205–221.e7. doi: 10.1016/j.devcel.2018.09.017

Krogh, A., Sonnhammer, E. L. L., and Ka, L. (2007). Advantages of combined transmembrane topology and signal peptide prediction — the Phobius web server. *Nucleic Acids Res.* 35, W429–W432. doi: 10.1093/nar/gkm256

La Cour, T., Kiemer, L., Mølgaard, A., Gupta, R., Skriver, K., and Brunak, S. (2004). Analysis and prediction of leucine-rich nuclear export signals. *Protein Eng. Des. Sel.* 17, 527–536. doi: 10.1093/protein/gzh062

Lange, A., Mills, R. E., Lange, C. J., Stewart, M., Devine, S. E., and Corbett, A. H. (2007). Classical nuclear localization signals: definition, function, and interaction with importin α. *J. Biol. Chem.* 282, 5101–5105. doi: 10.1074/jbc.R600026200

Lee, B. J., Cansizoglu, A. E., Su, K. E., Louis, T. H., Zhang, Z., and Chook, Y. M. (2006). Rules for nuclear localization sequence recognition by karyopherin β 2. *Cell* 126, 543–558. doi: 10.1016/j.cell.2006.05.049

Lee, D. W., Lee, S., Lee, J., Woo, S., Razzak, M. A., Vitale, A., et al. (2019). Molecular mechanism of the specificity of protein import into chloroplasts and mitochondria in plant cells. *Mol. Plant* 12, 951–966. doi: 10.1016/j.molp.2019.03.003

Lertampaiporn, S., Nuannimnoi, S., Vorapreeda, T., Chokesajjawatee, N., Visessanguan, W., and Thammarongtham, C. (2019). PSO-LocBact: a consensus method for optimizing multiple classifier results for predicting the subcellular localization of bacterial proteins. *Biomed. Res. Int.* 2019:5617153. doi: 10.1155/2019/5617153

Li, H. M., and Chiu, C. C. (2010). Protein transport into chloroplasts. *Annu. Rev. Plant Biol.* 61, 157–180. doi: 10.1146/annurev-arplant-042809-112222

Liku, M. E., Legere, E. A., and Moses, A. M. (2018). NoLogo: a new statistical model highlights the diversity and suggests new classes of Crm1-dependent nuclear export signals. *BMC Bioinformatics* 19:65. doi: 10.1186/s12859-018-2076-7

Lin, J., and Hu, J. (2013). SeqNLS: nuclear localization signal prediction based on frequent pattern mining and linear motif scoring. *PLoS One* 8:e76864. doi: 10.1371/journal.pone.0076864

Lisitsyna, O. M., Seplyarskiy, V. B., and Sheval, E. V. (2017). Comparative analysis of nuclear localization signal (NLS) prediction methods. *Biopolym. Cell* 33, 147–154. doi: 10.7124/bc.00094C

Lundberg, E., and Borner, G. H. H. (2019). Spatial proteomics: a powerful discovery tool for cell biology. *Nat. Rev. Mol. Cell Biol.* 20, 285–302. doi: 10.1038/s41580-018-0094-y

Maertens, G. N., Cook, N. J., Wang, W., Hare, S., Shree, S., and Öztop, I. (2014). Structural basis for nuclear import of splicing factors by human transportin 3. *Proc. Natl. Acad. Sci. U. S. A.* 111, 2728–2733. doi: 10.1073/pnas.1320755111

Martelli, P. L., Fariselli, P., and Casadio, R. (2003). An ENSEMBLE machine learning approach for the prediction of all-alpha membrane proteins. *Bioinformatics* 19, i205–i211. doi: 10.1093/bioinformatics/btg1027

Mathur, D., Singh, S., Mehta, A., Agrawal, P., and Raghava, G. P. S. (2018). In silico approaches for predicting the half-life of natural and modified peptides in blood. *PLoS One* 13:e0196829. doi: 10.1371/journal.pone.0196829

Mehdi, A. M., Sehgal, M. S. B., Kobe, B., Bailey, T. L., and Bodén, M. (2011). A probabilistic model of nuclear import of proteins. *Bioinformatics* 27, 1239–1246. doi: 10.1093/bioinformatics/btr121

Mossmann, D., Meisinger, C., and Vögtle, F. N. (2012). Processing of mitochondrial presequences. *Biochim. Biophys. Acta Gene Regul. Mech.* 1819, 1098–1106. doi: 10.1016/j.bbagrm.2011.11.007

Nakai, K. (2001). Review: prediction of *in vivo* fates of proteins in the era of genomics and proteomics. *J. Struct. Biol.* 134, 103–116. doi: 10.1006/jsbi.2001.4378

Nakai, K., and Kanehisa, M. (1991). Expert system for predicting protein localization sites in gram-negative bacteria. *Proteins Struct. Funct. Bioinforma.* 11, 95–110. doi: 10.1002/prot.340110203

Nakai, K., and Kanehisa, M. (1992). A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics* 14, 897–911. doi: 10.1016/S0888-7543(05)80111-9

Nielsen, H. (2017). Protein sorting prediction. *Methods Mol. Biol.* 1615, 23–57. doi: 10.1007/978-1-4939-7033-9_2

Nielsen, H., Tsirigos, K. D., Brunak, S., and von Heijne, G. (2019). A brief history of protein sorting prediction. *Protein J.* 38, 200–216. doi: 10.1007/s10930-019-09838-3

Nightingale, D. J. H., Oliver, S. G., and Lilley, K. S. (2019). Mapping the *Saccharomyces cerevisiae* spatial proteome with high resolution using hyperLOPIT. *Methods Mol. Biol.* 2049, 165–190. doi: 10.1007/978-1-4939-9736-7_10

Nilsson, I., Lara, P., Hessa, T., Johnson, A. E., von Heijne, G.V., and Karamyshev, A. L. (2015). The code for directing proteins for translocation across ER membrane: SRP cotranslationally recognizes specific features of a signal sequence. *J. Mol. Biol.* 427, 1191–1201. doi:10.1016/j.jmb.2014.06.014

Orioli, T., and Vihinen, M. (2019). Benchmarking subcellular localization and variant tolerance predictors on membrane proteins. *BMC Genomics* 20:547. doi: 10.1186/s12864-019-5865-0

Orre, L. M., Vesterlund, M., Pan, Y., Arslan, T., Zhu, Y., Fernandez Woodbridge, A., et al. (2019). SubCellBarCode: proteome-wide mapping of protein localization and relocalization. *Mol. Cell* 73, 166–182.e7. doi: 10.1016/j.molcel.2018.11.035

Paila, Y. D., Richardson, L. G. L., and Schnell, D. J. (2015). New insights into the mechanism of chloroplast protein import and its integration with protein quality control, organelle biogenesis and development. *J. Mol. Biol.* 427, 1038–1060. doi: 10.1016/j.jmb.2014.08.016

Palmer, T., and Stansfeld, P. J. (2020). Targeting of proteins to the twin-arginine translocation pathway. *Mol. Microbiol.* 113, 861–871. doi: 10.1111/mmi.14461

Paramasivam, N., and Linke, D. (2011). Clubsub-P: cluster-based subcellular localization prediction for gram-negative bacteria and archaea. *Front. Microbiol.* 2:218. doi: 10.3389/fmicb.2011.00218

Peabody, M. A., Laird, M. R., Vlasschaert, C., Lo, R., and Brinkman, F. S. L. (2016). PSORTdb: expanding the bacteria and archaea protein subcellular localization database to better reflect diversity in cell envelope structures. *Nucleic Acids Res.* 44, D663–D668. doi: 10.1093/nar/gkv1271

Peabody, M. A., Lau, W. Y. V., Hoad, G. R., Jia, B., Maguire, F., Gray, K. L., et al. (2020). PSORTm: a bacterial and archaeal protein subcellular localization prediction tool for metagenomics data. *Bioinformatics* 36, 3043–3048. doi: 10.1093/bioinformatics/btaa136

Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* 8, 785–786. doi: 10.1038/nmeth.1701

Pfanner, N., Warscheid, B., and Wiedemann, N. (2019). Mitochondrial proteins: from biogenesis to functional networks. *Nat. Rev. Mol. Cell Biol.* 20, 267–284. doi: 10.1038/s41580-018-0092-0

Pierleoni, A., Martelli, P., and Casadio, R. (2008). PredGPI: a GPI-anchor predictor. *BMC Bioinformatics* 9:392. doi: 10.1186/1471-2105-9-392

Pierleoni, A., Martelli, P. L., and Casadio, R. (2011). MemLoci: predicting subcellular localization of membrane proteins in eukaryotes. *Bioinformatics* 27, 1224–1230. doi: 10.1093/bioinformatics/btr108

Pierleoni, A., Martelli, P. L., Fariselli, P., and Casadio, R. (2006). BaCelLo: a balanced subcellular localization predictor. *Bioinformatics* 22, e408–e416. doi: 10.1093/bioinformatics/btl222

Prieto, G., Fullaondo, A., and Rodriguez, J. A. (2014). Prediction of nuclear export signals using weighted regular expressions (Wregex). *Bioinformatics* 30, 1220–1227. doi: 10.1093/bioinformatics/btu016

Salvatore, M., Warholm, P., Shu, N., Basile, W., and Elofsson, A. (2017). SubCons: a new ensemble method for improved human subcellular localization predictions. *Bioinformatics* 33, 2464–2470. doi: 10.1093/bioinformatics/btx219

Savojardo, C., Fariselli, P., and Casadio, R. (2013). BETAWARE: a machine-learning tool to detect and predict transmembrane beta-barrel proteins in prokaryotes. *Bioinformatics* 29, 504–505. doi: 10.1093/bioinformatics/bts728

Savojardo, C., Martelli, P. L., Fariselli, P., and Casadio, R. (2015). TPpred3 detects and discriminates mitochondrial and chloroplastic targeting peptides in eukaryotic proteins. *Bioinformatics* 31, 3269–3275. doi: 10.1093/bioinformatics/btv367

Savojardo, C., Martelli, P. L., Fariselli, P., and Casadio, R. (2017). SChloro: directing Viridiplantae proteins to six chloroplastic sub-compartments. *Bioinformatics* 33, 347–353. doi: 10.1093/bioinformatics/btw656

Savojardo, C., Martelli, P. L., Fariselli, P., and Casadio, R. (2018a). DeepSig: deep learning improves signal peptide detection in proteins. *Bioinformatics* 34, 1690–1696. doi: 10.1093/bioinformatics/btx818

Savojardo, C., Martelli, P. L., Fariselli, P., Profiti, G., and Casadio, R. (2018b). BUSCA: an integrative web server to predict subcellular localization of proteins. *Nucleic Acids Res.* 46, W459–W466. doi: 10.1093/nar/gky320

Schneider, G., Sjöling, S., Wallin, E., Wrede, P., Glaser, E., and von Heijne, G. (1998). Feature-extraction from endopeptidase cleavage sites in mitochondrial targeting peptides. *Proteins Struct. Funct. Genet.* 30, 49–60. doi: 10.1002/(SICI)1097-0134(19980101)30:1<49::AID-PROT5>3.0.CO;2-F

Shen, Y., Ding, Y., Tang, J., Zou, Q., and Guo, F. (2020). Critical evaluation of web-based prediction tools for human protein subcellular localization. *Brief. Bioinform.* 21, 1628–1640. doi: 10.1093/bib/bbz106

Siegel, S. D., Reardon, M. E., and Ton-That, H. (2017). Anchoring of LPXTG-like proteins to the gram-positive cell wall envelope. *Curr. Top. Microbiol. Immunol.* 404, 159–175. doi: 10.1007/82_2016_8

Small, I., Peeters, N., Legeai, F., and Lurin, C. (2004). Predotar: a tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics* 4, 1581–1590. doi: 10.1002/pmic.200300776

Stekhoven, D. J., Omasits, U., Quebatte, M., Dehio, C., and Ahrens, C. H. (2014). Proteome-wide identification of predominant subcellular protein localizations in a bacterial model organism. *J. Proteomics* 99, 123–137. doi: 10.1016/j.jprot.2014.01.015

Thul, P. J., Akesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., et al. (2017). A subcellular map of the human proteome. *Science* 356:eaal3321. doi: 10.1126/science.aal3321

Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., et al. (2010). Towards a knowledge-based human protein atlas. *Nat. Biotechnol.* 28, 1248–1250. doi: 10.1038/nbt1210-1248

UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515. doi: 10.1093/nar/gky1049

Vakser, I. A. (2020). Challenges in protein docking. *Curr. Opin. Struct. Biol.* 64, 160–165. doi: 10.1016/j.sbi.2020.07.001

Vögtle, F. N., Wortelkamp, S., Zahedi, R. P., Becker, D., Leidhold, C., Gevaert, K., et al. (2009). Global analysis of the mitochondrial N-proteome identifies a processing peptidase critical for protein stability. *Cell* 139, 428–439. doi: 10.1016/j.cell.2009.07.045

von Heijne, G. (1986). Mitochondrial targeting sequences may form amphiphilic helices. *EMBO J.* 5, 1335–1342. doi: 10.1002/j.1460-2075.1986.tb04364.x

von Heijne, G. (1990). The signal peptide. *J. Membr. Biol.* 115, 195–201. doi: 10.1007/BF01868635

Wan, S., Mak, M. W., and Kung, S. Y. (2012). mGOASVM: multi-label protein subcellular localization based on gene ontology and support vector machines. *BMC Bioinformatics* 13:290. doi: 10.1186/1471-2105-13-290

Wang, X., Zhang, J., and Li, G. Z. (2015). Multi-location gram-positive and gram-negative bacterial protein subcellular localization using gene ontology and multi-label classifier ensemble. *BMC Bioinformatics* 16:S1. doi: 10.1186/1471-2105-16-S12-S1

Xu, D., Marquis, K., Pei, J., Fu, S. C., Caıatay, T., Grishin, N. V., et al. (2015). LocNES: a computational tool for locating classical NESs in CRM1 cargo proteins. *Bioinformatics* 31, 1357–1365. doi: 10.1093/bioinformatics/btu826

Yu, C. S., Chen, Y. C., Lu, C. H., and Hwang, J. K. (2006). Prediction of protein subcellular localization. *Proteins Struct. Funct. Genet.* 64, 643–651. doi: 10.1002/prot.21018

Yu, N. Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., et al. (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* 26, 1608–1615. doi: 10.1093/bioinformatics/btq249

Zhang, S., Xia, X., Shen, J., Zhou, Y., and Sun, Z. (2008). DBMLoc: a database of proteins with multiple subcellular localizations. *BMC Bioinformatics* 9:127. doi: 10.1186/1471-2105-9-127

Zybailov, B., Rutschow, H., Friso, G., Rudella, A., Emanuelsson, O., Sun, Q., et al. (2008). Sorting signals, N-terminal modifications and abundance of the chloroplast proteome. *PLoS One* 3:e1994. doi: 10.1371/journal.pone.0001994

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.