DOI 10.1002/pmic.201000274

Review

Prediction of subcellular locations of proteins: Where to proceed?

Kenichiro Imai[1,2] *and Kenta Nakai*[3]

[1] Computational Biology Research Center, AIST, Tokyo, Japan
[2] Japan Society for the Promotion of Science, Japan
[3] Human Genome Center, Institute of Medical Science, University of Tokyo, Japan

Since the proposal of the signal hypothesis on protein subcellular sorting, a number of computational analyses have been performed in this field. A typical example is the development of prediction algorithms for the subcellular localization sites of input protein sequences. In this review, we mainly focus on the biological grounds of the prediction methods rather than the algorithmic issues because we believe the former will be more fruitful for future development. Recent advances on the study of protein sorting signals will hopefully be incorporated into future prediction methods. Unfortunately, many of the state-of-the-art methods are published without sufficient objective tests. In fact, a simple test employed in this article shows that the performance of specifically developed predictors is not significantly better than that of a homology search. We suspect that this is a general problem associated with the interpretation of genome sequences, which have evolved through gene duplication and speciation.

Received: April 23, 2010
Revised: July 30, 2010
Accepted: August 3, 2010

Keywords:
Bioinformatics / Pattern recognition / Prediction / Sequence analysis / Sorting signal

1 Introduction

In living cells, proteins are sorted and transported into appropriate subcellular locations to play their proper cellular roles. This is also true in bacterial cells, which do not have intracellular membrane structure (organelles). Thus, knowing the subcellular localization site of a given protein could be a great help to elucidate its cellular function. More importantly, the information on the final subcellular localization site of a protein is basically encoded as a part of its amino acid sequence and such a sequence is thought to be recognized by a specific receptor protein as a protein sorting signal. Thus, it would be possible, at least in principle, for us to recognize/predict the subcellular localization site of proteins from their amino acid sequences. So far, a number of prediction methods have been developed and the prediction of subcellular localization sites is now regarded as an established field in the sequence analysis. However, there seem to be many studies where the prediction problem is just a convenient test bed for novel machine-learning algorithms. Such studies would not give us useful biological insights; rather the studies could be problematic because using the data sets in this way for the prediction of subcellular localization suffers from an inherent problem that prevents fair benchmarking purposes, as discussed below. In this review, we would like to emphasize topics that we believe are important biologically as well as important for the future of this field in this era of proteomics. That is, we (i) briefly summarize the current approaches based on the features that are fed to the classifier, (ii) review the recent advances on our understanding of protein sorting signals, in particular, and (iii) discuss the difficulties in objective

Correspondence: Professor Kenta Nakai, Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokandedai, Minato-ku, Tokyo 108-8839, Japan
E-mail: knakai@ims.u-tokyo.ac.jp
Fax: +81-3-5449-5133

Abbreviations: ER, endoplasmic reticulum; **IMS,** intermembrane space; **LR-NES,** leucine-rich NES; **MCC,** Matthews correlation coefficient; **MPP,** mitochondrial processing peptidase; **NES,** nuclear export signal; **NLS,** nuclear localization signal; **PTS,** peroxisomal targeting signal; **SVM,** support vector machine; **Tat,** twin-arginine translocation assessment. Inevitably, the scope of this review is not fully comprehensive. For a wider view of this field, please consult other reviews [1–6].

2 Brief history and classification of methodology

2.1 Brief history

As an example of earlier studies on the prediction of subcellular localization, von Heijne developed a simple prediction method in 1986 based on (the cleavage site of) signal peptides [7]. A few years earlier (in 1982), Nishikawa and Ooi discovered that the amino acid composition of intracellular and extracellular proteins is markedly different [8] and Kyte and Doolittle's hydropathy indices, which were also published in 1982, have been widely used for the recognition of membrane proteins and their transmembrane domains [9]. Nakai and Kanehisa tried to integrate all these information into one system for the prediction of several possible subcellular localization sites using the input information of an amino acid sequence and its sequence origin [10, 11]. The program was named PSORT and was applied to the analysis of the newly determined yeast chromosome III sequence. Since then the PSORT program has been available through the Internet [12, 13]. Then, the age of genomics came and the need for prediction programs continued to increase. Since the prediction accuracy of general predictors such as PSORT was not enough for the automatic annotation of the genome sequence, programs for the detection of specific sorting signals became more popular. For example, Nielsen *et al.*'s SignalP and Emanuelsson *et al.*'s TargetP were widely used [14, 15]. On the other hand, surprisingly many papers reporting the use of amino acid composition and variants of this theme have also been published. For example, Nakashima and Nishikawa and Cedano *et al.* are some of the earlier attempts [16, 17]. Since these methods do not use prior knowledge on protein sorting, they are more suited for the application of various machine-learning algorithms.

After the initial completion of human genome sequencing in 2001, it seems that comparative genomics rather than *de novo* sequence-based prediction has attracted more attention for annotating genome sequences. However, the development of subcellular localization predictors continued and there have been several notable progresses. For example, Nair and Rost developed a predictor mimicking the cellular decision tree for protein sorting [18]. Although this idea itself was already seen in PSORT, their algorithm was tested much more objectively. Pierleoni *et al.* tried to improve this approach by creating a balanced prediction tree [19]. Another recent issue is how to treat proteins that are localized at multiple sites. Although several prediction algorithms have been developed [20–22], their practical values are questionable because the selection/annotation of such proteins is still quite immature. More recently, a potentially useful prediction method, YLoc, was published [23]; this method is unique in that the predictions are shown as texts, which are usually preferred by experimental researchers. In addition, a recently published study by Lin *et al.* [24], described a hierarchical prediction method; at each step, they used discriminative motifs (motifs that are present in a set of sequences but are absent in the other set), which were automatically identified beforehand using their discriminative hidden Markov model approach. Such an approach would be more effective if we know that a subset of sequences is specifically localized at a certain place through a common pathway from some experimental evidence. What makes their approach attractive is that it shows a new possibility of *in silico* methods to be practical for cell biological studies.

At the end of this brief overview, we list the names of three typical prediction programs, which are currently publicly available now (they will be used for the discussion in Section 4). WoLF PSORT is based on the *k*-nearest neighbor algorithm and uses both the knowledge of protein sorting signals and amino acid composition [20]. CELLO is a hybrid method of an support vector machine (SVM)-based predictor using the *n*-peptide composition and a homology search [25]. MultiLoc2 is the successor to MultiLoc, which is an SVM-based method incorporating N-terminal targeting sequences, sequence motifs, amino acid composition, phylogenetic profiles, and Gene Ontology terms [26, 27].

2.2 Classification of input features

Since the prediction of subcellular localization is a kind of pattern recognition problem, the wealth of pattern recognition algorithms can in principle be applicable. Therefore, we provide an overview of the prediction methods from the nature of their input information rather than the nature of the pattern recognition algorithms used. Note that by 'input information', we mean the input vector that is fed to the pattern recognition algorithms rather than the primary input information, which is the amino acid sequence itself in most cases. Note also that the classifications are not mutually exclusive, as will be pointed out below. There are four categories of features which we outline below: features based on (i) protein sorting signals, (ii) empirically correlated characteristics, (iii) sequence homology with known answer sets and (iv) other sources.

2.2.1 Features based on protein sorting signals

The first category of the input features is protein sorting signals. In fact, knowledge of sorting signals is the foundation for making the prediction of subcellular localization feasible. However, since not all proteins in an organelle share a common signal, attempts to develop proteome-wide predictions always have some inherent limitation. In addition, the fact that the translational start site of many proteins is sometimes predicted erroneously from their cDNA sequences limits the performance since many of the typical sorting signals reside at the *N*-terminus.

2.2.2 Features based on empirically correlated characteristics

The second category of the input features is the global/local sequence features that may not be directly related to the sorting mechanism but their existence/absence is correlated to (certain) localization sites. A typical example of the global features is the amino acid composition of the entire (or mature) sequence. There are many variants on this theme, such as the dipeptide (or the *n*-gram, *k*-mer) composition. One drawback of these features is that there are no biological grounds for why such a correlation exists (although there is the classical work by Andrade *et al.* [28], where they concluded that surface residues of proteins have adapted to their external conditions as characterized by their subcellular locations). Another drawback is that the similarity of generalized composition based on *k*-mers would be closer to the global homology information with larger *k*-values: such methods might not be so effective for proteins without homology. On the other hand, they are simple, general, and easier to test objectively (with cross-validation, for example). In signal-based predictions, it is necessary to recognize each signal with its specifically tailored sub-predictors first and the need for reoptimizing each of these sub-predictors from a relatively small set of known examples in cross-validation would not be suitable (see Section 4 also for homology-based methods).

Some local features, such as functional motifs or conserved domains, which are stored in public databases, *e.g.* PROSITE/Pfam [29, 30], also belong to this category. The content of these databases includes the conserved sequence patterns, many of which are characteristic of the functional sites of proteins, such as the helix-loop-helix DNA binding motif (of course, motifs of sorting signals, *e.g.* 'KDEL', are also found in PROSITE). These functional motifs are useful for prediction because it is very likely that DNA-binding proteins are localized in the nucleus irrespective of their sorting mechanisms. In fact, the use of such motifs for prediction is similar to the approaches that use overrepresented *k*-mers in proteins localized to a certain organelle.

2.2.3 Features based on sequence homology with known answer sets

The third category of input features is the use of sequence homology: typically, the localization site of the most (or the majority of) similar protein(s) in the training data set is also regarded as the site of the query protein itself. As we will discuss later, this feature is the most practical in terms of achieving good effectiveness because most of the localization sites have been determined in yeast through proteomic experiments. However, the performance of the prediction methods relying on this feature is hard to evaluate and does not have importance to cell biology except in one situation: there exist several pairs of proteins of almost identical sequences (*i.e.* isoforms) but are known to be localized at different sites. If sufficiently annotated databases are provided, finding such differentially localized pairs from the homology search may be biologically fruitful. In Section 4, we will discuss such a possibility. There are several algorithms that use the annotation information of query proteins, such as the Gene Ontology information, for their prediction [31–33]. Although these features can be classified into the next (fourth) category, they can also be regarded as belonging to this category because usually a homology search method is used to find the corresponding protein in the database.

2.2.4 Features from other sources

The fourth category of input features is the use of external information in addition to the amino acid sequence information. A typical example is the use of protein–protein interaction data [34]. Other examples include the use of gene expression data as well as evolutionary information [35]. As noted above, some prediction algorithms use the annotation information of public databases such as Swiss-Prot. It is not surprising that such algorithms are benefited by the wealth of text-mining techniques. Prediction methods that rely on such external information have limited practical value since completely unknown proteins do not have any corresponding information in these databases.

3 Newly obtained knowledge on sorting signals

In this section, we review the recent advances in our understanding of various protein sorting signals, hoping that they will be incorporated into future prediction algorithms. We begin by covering seven different types of signals followed by signals for the secretory pathway, membrane protein topology, and signals for bacterial protein transport.

3.1 Signal peptides

Signal peptides, which are the N-terminal extensions of secretory and membrane proteins, target these proteins from the cytosol to the cytoplasmic membrane of prokaryotes or to the endoplasmic reticulum (ER) membrane of eukaryotes. They are typically 15–25 residues in length, consisting of three regions; a (typically) positively charged region (*n*-region), a hydrophobic region (*h*-region), which is functionally essential, and a polar *c*-region, which contains a weak consensus sequences used as the signal peptidase cleavage site [36, 37].

One recent topic on signal peptides is the proposal of a unique set of peptide motifs (*h*-motifs) in the *h*-region of signal peptides in several eukaryotes (human, *Saccharomyces cerevisiae* and *Trypanosoma brucei*) as well as prokaryote(s) (*Escherichia coli*) [38]. This proposal challenges our belief that there is no consensus sequence in signal peptides as exemplified by the functional signal peptide with a simple polyleucine *h*-region (see [39–41], for example). It is reported that human *h*-motifs have four identity components, whereas those from the other species are characterized by three identity components. For example, human Hs1 is L[A/G/L/V]XL$X_{0,1}$L, *S. cerevisiae* Sc1 is L$X_{0,2}$S$X_{0,3}$A, *T. brucei* Tb1 is LLX[A/I/L/V], and *E. coli* Ec1 is A$X_{0,2}$L$X_{0,3}$ (where $X_{1,2}$ indicate one or two unspecified residues and $X_{1–3}$ represents a stretch of unspecified residues of length one to three residues). These *h*-motifs are conserved within each species and are detectable in 75–90% of all signal sequences. Moreover, they experimentally confirmed that a variant, whose *h*-motif was disrupted by scrambling the *h*-region sequence while keeping its hydrophobicity, was not functional. Thus, it would be of interest to see if the use of this motif will improve the prediction accuracy of signal peptides.

As another interesting topic, Hiss and Schneider evaluated the prediction performance of the currently available prediction methods for a set of 136 annotated eukaryotic long signal peptides (*i.e.* longer than 40 residues) [42]. Such long signal peptides usually have an additional post-targeting function, such as nuclear localization [43]. The best performance was achieved by SignalP [44] and Signal-CF [45], but their detection abilities were surprisingly lower for long signal peptides compared with the usual values for typical signal peptides: only 61 and 63%, respectively, suggesting the need for prediction methods specific for longer signal peptides.

3.2 Nuclear localization signals

Although the nuclear pore complexes allow passive diffusion of small proteins, the transport of many nuclear proteins that have the nuclear localization signals (NLSs), seem to be mediated by import receptors termed karyopherin or importins [46]. In particular, the importin α/β pathway mediates the import of proteins containing classical NLSs. The classical NLSs are characterized by one (monopartite) or two (bipartite) stretches of basic amino acids, often appearing in the middle of the amino acid sequences. A large number of monopartite and bipartite classical NLSs have been experimentally identified and their consensus sequences characterized as K[K/R]X[K/R] and KR$X_{10–12}$KRXK, respectively [46]. However, this information is still quite incomplete.

Lange *et al.* estimated the prevalence of classical NLSs for data sets including 1515 nuclear proteins identified in a global yeast GFP-fusion library as well as 224 proteins that interact with importin α according to the BioGRID database [46]. In the nuclear protein data set, 30.9 and 25.8% of proteins contained monopartite and bipartite classical NLSs that were predicted by PSORT II [12], respectively. In the importin α data set, the proportion was 27.2 and 21.9%, respectively. Thus, they estimate that about 43% of nuclear proteins may use other transport mechanisms and that about half of the proteins that interact with importin α do not possess the predicted classical NLSs. Kosugi *et al.* therefore classified classical NLSs into six classes (three of which are novel) by screening peptides that bind to distinct binding pockets of importin α [47]. They suggested that the consensus patterns for the classical NLS classes are KR[K/R]R or K[K/R]RK for class 1, [P/R]XXKR[^DE][K/R] for class 2, KRX[W/F/Y]XXAF for class 3, [R/P]XXKR[K/R][^DE] for class 4, LGKR[K/R][W/F/Y] for class 5, and KR$X_{10–12}$K[K/R][K/R] and KR$X_{10–12}$K[K/R]X[K/R] for the bipartite class, where [^DE] represents any amino acids except for aspartates and glutamates. However, later on Lange *et al.* experimentally confirmed that Rrp4 contains a bipartite classical NLS with a much longer linker length (25 residues) than the conventional consensus sequence [48]. They pointed out that the linker length in the definition of bipartite classical NLS consensus sequence should be expanded and that this revision leads to an increase in the number of potential proteins containing the bipartite classical NLS. In addition, Lee *et al.* proposed a new NLS, 'PY-NLS', which is recognized by karyopherinβ2 from their structural analysis of the Karyopherinβ2-M9NLS complex and biochemical studies [49]. PY-NLS is defined to be structurally disordered in free substrates, have a net basic charge, and possess a hydrophobic or basic region followed by a C-terminal [R/H/K]$X_{2–5}$PY consensus sequence. These rules are identified in seven known Karyopherinβ2 substrates and have led to the discovery of 81 new candidates.

In spite of the above studies on NLSs, it is still evident that their existence cannot cover all the known nuclear proteins. Thus, some NLS predictors, such as PredictNLS [50] and NucPred [51] use enlarged sets of NLS patterns that were obtained from either an *in silico* mutagenesis from known NLSs or a *de novo* discovery using a genetic programming algorithm.

3.3 Nuclear export signals

In contrast to NLSs, only one class of nuclear export signals (NESs) is known: the leucine-rich NES (LR-NES). The LR-NESs, which are exemplified by 'LPPLERLTL' in HIV Rev protein, are recognized by CRM1 (also known as exportin1) [52].

At first, the LR-NES was characterized by a consensus sequence; L$X_{2,3}$[L/I/V/F/M] $X_{2,3}$LX[L/I] [53], but only 36% of experimentally defined NESs matched this consensus. la Cour *et al.* therefore analyzed the amino acid composition of 67 high-confidence NESs and proposed a new consensus sequence: [L/I/V/F/M]$X_{2,3}$[L/IV/F/M]$X_{2,3}$[L/I/V/F/M]X[L/I/V/F/M] [54]. They also pointed out the overrepresentation of glutamates, aspartates, and serines in the NES region and its flanking regions. Furthermore, another set of extended consensus sequences of NESs was recently proposed: Kosugi *et al.* obtained 101 distinct CRM-1 dependent NESs experimentally and classified them into six classes [55]. The defined patterns covered 99% of the 101 NESs. Moreover, proline substitutions in non-conserved positions of NES significantly decreased the NES activity. Then, they proposed three consensus patterns for the NES classes, Φ$X_{1,2}$[^P] Φ [^P]$_{2,3}$Φ [^P] Φ, Φ [^P] Φ [^P]$_2$Φ [^P] Φ and ΦX[^P] Φ [^P]$_3$Φ [^P]$_2$Φ, where Φ represents L, I, V, M, F, C, W, A or T (C, T, A, and W are allowable only at one of the four positions), and [^P]$_3$ represents any three amino acids except for prolines.

A recent structural analysis of the CRM1-snurportin 1 complex [56] explains the preference for glutamates, aspartates, and serines in the NES region and its flanking region, as initially reported by la Cour *et al.* [54]. It was also speculated that many CRM1 cargos may have another basic surface NES enabling multipartite recognition like NLSs.

3.4 Mitochondrial targeting signals

Since mitochondria have two membranes, there are four subcompartments: the outer membrane, the intermembrane space (IMS), the inner membrane, and the matrix. Of the approximately 1500 mitochondrial proteins, 99% are synthesized in and imported from the cytosol. Those mitochondrial precursor proteins can be classified into two groups: those that have an *N*-terminal presequence and those that do not.

Most matrix proteins as well as several inner membrane and IMS proteins belong to the former group and are recognized by the translocators, which consist of the TOM and TIM complex in the outer and inner membranes, respectively. The presequence is in most cases cleaved off by the mitochondrial processing peptidase (MPP) in the matrix upon import [57]. Presequences consist of 10–90 amino acid residues and have the ability to form a positively charged amphiphilic helical structure. Like signal peptides, they have been considered to lack a consensus sequence motif. However, in 2003, Obita *et al.* proposed a consensus motif, σΦXβΦΦ (where σ, Φ, and β represent ahydrophilic, hydrophobic, and basic residue, respectively) for the Tom20 import receptor, based on a peptide library analysis [58]. Moreover, the existence of weak motifs around the cleavage site has been pointed out [59]. A large number of mutational studies indicate the importance of the so-called R-2 and R-3 motifs (XRX↓X[S/X] and XRX[Y/X]↓[S/A/X]X, where represents the cleavage site) for the cleavage by MPP [60], and [S/X] represents that serine residue is not dominant but loosely conserved at this position. In addition, a recent global analysis of the processing site of presequences by Vögtle *et al.* identified a new processing peptidase, Icp55 [61]. Icp55 removes a single residue after the MPP cleavage site, converting the R-3 motif into R-2. Interestingly, the residues removed by Icp55 correlate with destabilizing amino acids in the *N*-end rule for prokaryotes, which suggests that the intermediate peptidase converts destabilizing *N*-termini generated by MPP into the stable ones. The authors also tested the prediction performance of MitoProt II [62] and TargetP [15] for cleavage sites by using their data set of experimentally determined presequences. The prediction accuracies of MitoProtII and TargetP were only 21 and 33%, respectively, showing the difficulty of this problem.

The second group of mitochondrial proteins, including the polytopic inner membrane proteins, soluble IMS proteins and the outer membrane proteins are synthesized without a cleavable presequence but contain internal targeting signals within their mature sequences. Although these internal targeting signals are less understood, a new pathway for presequence-less mitochondrial precursor proteins, the SAM pathway, has been characterized in the last seven years [63, 64]. Reflecting its evolutionary origin from bacteria, the mitochondrial outer membrane contains β-barrel membrane proteins (MBOMPs), but only five families (VDAC, Tom40, Sam50, Mdm10, and Mmm2) are known in eukaryotes. The import and the integration of MBOMPs into the outer membrane are mediated by the TOM40 complex, small Tim proteins, and the SAM complex [57]. Until recently, the sorting signal to the SAM complex had been unclear but Kutik *et al.* experimentally identified the first specific sorting signal for MBOMPs: the β-signal $P_oX\mathrm{G}XXH_yXH_y$ (P_o = a polar residue, G = Gly, and H_y = a large hydrophobic residue) [65]. They also reported that Sam35 possesses receptor-like properties and binds to the β-signal. A recently resolved 3-D structure of VDAC1 revealed a β-signal located near the end of the C-terminal β-strand [66, 67]. In addition, a multiple sequence alignment between the orthologs of MBOMPs by Imai *et al.* showed that four of the five known MBOMPs have a conserved β-signal near their C-termini, while Mmm2 has a conserved β-signal far from the C-terminus [68]. According to a secondary structure prediction, the β-signal of Mmm2 is located at the last β-strand of the *N*-terminal β-strand-rich region. They further proposed a minor refinement of this signal from the multiple alignment between orthologs (the refined signal is $P_oX\mathrm{G}h_yXH_yXH_y$, where h_y is a hydrophobic residue with a large side chain as well as smaller ones, such as alanine and cysteine). In contrast to the previous bioinformatics sequence analysis that estimated the number of yeast MBOMPs to be about 100 [69], Imai *et al.*'s result suggests that there may be only a few MBOMPs remain undiscovered.

3.5 Chloroplast transit peptides

Chloroplasts are composed of six subcompartments: three membranes (the outer envelope membrane, the inner envelope membrane, and the thylakoid membrane) and three distinct aqueous regions (the IMS, the stroma, and the thylakoid lumen). Most chloroplast proteins are encoded in the nuclear genome and are synthesized in the cytosol as precursor proteins. Except for outer envelope membrane proteins, many chloroplast precursor proteins have a chloroplast transit peptide as a cleavable *N*-terminal extension. The signal works as a stromal targeting signal. The proteins with the signal are imported into the chloroplast interior by specific translocators (the TOC and TIC complex in the outer and inner envelope membranes, respectively) and then the transit peptide is recognized and cleaved off by the stromal processing peptidase [70]. The *N*-terminal transit peptides are unstructured in aqueous environment but have the ability to form a helical structure in a membrane mimetic environment [71]. They consist of 30–80 amino acid residues, have high preference for hydrophobic, hydroxylated, and positively charged amino acids, and are deficient in acidic amino acids [72]. The amino acid composition of plant mitochondrial presequences and chloroplast transit signals are significantly similar, but a recent sequence analysis showed that arginines are much more abundant in the first 16 amino acids of mitochondrial presequences, while serines and prolines are over-represented in chloroplast transit peptides [73]. Similar to mitochondrial presequences, any clear consensus sequence has not been identified in the chloroplast targeting signal. However, Lee *et al.* suggested that the transit sequences are classified into multiple subgroups by applying hierarchical clustering on 208 transit peptides of *Arabidopsis thaliana,* and that the representative proteins selected from the seven subgroups contained specific sequence motifs critical for protein import into chloroplasts [74]. They even claimed that the grouping of the transit peptides lead to an increase in prediction accuracy. As for the cleavage site motif, a loosely conserved motif VXA↓A has been proposed [72]. Zybailov *et al.* pointed out the difference between the experimentally obtained cleavage site motif and the one predicted by ChloroP [75, 76]. Their comparison revealed a different consensus motif with the cleavage site shifted upstream by one residue. This shift may indicate the involvement of an intermediate processing peptidase, such as Icp55 in mitochondria [61].

3.6 Dual targeting signals for mitochondria and chloroplasts

As described above, chloroplast transit peptides are similar to mitochondrial presequences. In fact, about 50 proteins are known to be targeted by both organelles [77]. Not surprisingly, the dual targeting signals are also similar to both targeting signals. Although there has been a report that phenylalanines and leucines are more abundant in the dual signals [78], another recent analysis did not confirm this [73]: The first 20 residues of the chloroplast protein *N*-termini contained more serines and less arginines, whereas mitochondrial protein *N*-termini have significantly more arginines. The dual targeted proteins have intermediate contents of serines and arginines. It is not clear, however, whether these observed differences are functionally important or not.

3.7 Peroxisomal targeting signals

There are three kinds of peroxisomal targeting signals (PTSs). Most matrix proteins possess PTS1 or PTS2, while membrane proteins contain mPTS.

PTS1 is basically a specific tripeptide (SKL or its conserved variants) at the C-terminus of proteins [79]. Neuberger *et al.* reported that at least 12 residues of the C-terminus, which consists of the C-terminal tripeptide, a region interacting with a soluble PTS1 receptor (PEX5), and an unstructured linker region, may be involved in the recognition of PTS1 [80]. Recently, new variants of tripeptide motifs (SSL, SSI, SLM, SKF, and ASL) were discovered in proteins involved in novel functions from a proteome analysis of Arabidopsis leaf peroxisomes [81, 82].

PTS2 is approximately characterized as a nine-residue motif, [R/K][L/V/I]X_5[H/Q][L/A], located at the N-terminus [83]. The N-terminal portion containing the motif is cleaved off after the import. Petriv *et al.* proposed two revised consensus sequences based on an analysis of functional and non-functional PTS2 motifs: The motif R[L/V/I/Q]*XX*[L/V/I/H][L/S/G/A]*X*[H/Q][L/A] covers most PTS2s, while the other [RK][L/V/I/Q]*XX*[L/V/I/H/Q][L/S/G/A/K]*X*[H/Q][L/A/F] covers almost all PTS2s [84].

The mPTS consists of one or more transmembrane domains with a flanking cluster of basic residues. Although several motifs, as exemplified by [R/K]*X*[R/K]*X*[R/K][L/I]$X_{9,10}$[F/Y] and F[F/L]*X*[R/Q/K]X_3[I/L][L/S][K/R][V/I/L]L[F/I/V]P, have been proposed, no attempts have been successful for identifying a clear and universal motif representing mPTS [85]. Many peroxisomal membrane proteins are found to interact with Pex19 and a consensus sequence of Pex19-binding sites, X_3[C/F/I/L/T/V/W]X_2[A/C/F/I/L/Q/V/W/Y][C/I/L/V]X_2[A/C/F/I/L/V/W/Y][I/L/Q/R/V]X_3 was also proposed [86, 87]. According to a recent motif search against PeroxisomeDB, containing 2706 identified peroxisomal proteins, only 944 proteins had putative PTSs, where the numbers of proteins containing PTS1, PTS2 and the Pex19 binding site motif were 571, 73, and 482, respectively [88]. Remaining peroxisomal matrix proteins were thought to be imported through a piggy-backing mechanism [89, 90].

3.8 Sorting signals for the secretory pathway

So far, several types of COPII-dependent export signals from the ER, such as di-acidic or di-hydrophobic motifs, have been identified for transmembrane and luminal proteins [91]. Fernández-Sánchez *et al.* recently identified another conserved ER exit signal, RLX$_8$D, in glycine transporter-1 [92]. In addition, Tsukumo *et al.* claimed that a proline at the +2-position from the signal peptide cleavage site is a determinant for the export of NUCB1 protein from the ER and for its subsequent transport to the Golgi body [93]. Duvernay *et al.* also reported that a conserved single leucine residue in the first intracellular loop of G protein-coupled receptors, α_{2B}-adrenergic receptors, regulates their export from the ER to the cell surface [94]. As for the ER-retention/retrieval signals, the C-terminal KDEL sequence of soluble proteins, the KK*XX* or K*X*K*XX* motifs in the C-terminus of type I membrane proteins, and the R*X*R motif in ion channels have been identified [95].

The sorting signals of lysosomal proteins have also been studied well (reviewed in [96]). In brief, most soluble acid hydrolases are transported through the mannose 6-phosphate-dependent pathway, whereas the transport of other soluble proteins is mediated through mannose 6-phosphate-independent receptors, such as the lysosomal integral membrane protein LIMP-2 or sortilin. The sorting signals of cargo receptors and lysosomal transmembrane proteins include the di-leucine motif, D*XX*LL or [D/E]*XXX*L[L/I], and the tyrosine-based motif, Y*XX*Φ (Φ represents bulky hydrophobic residues), all of which exist on their cytosolic domains [96].

3.9 Determinants of membrane protein topology

Determining whether an input protein is an integral membrane protein or a soluble protein from its amino acid is an important step for the prediction of its subcellular localization. In the case of α-helical membrane proteins, this discrimination is not thought to be difficult. However, for making their localization prediction more accurate, it is also essential to predict their membrane topology because the membrane topology could be critical for defining the context of their sorting signals: a signal that is recognized in the cytosolic domain would not work if it is exposed to the extracellular domain. And the prediction of membrane topology is far more difficult than the recognition of membrane proteins. There are several difficulties: recognizing less hydrophobic helices, distinguishing signal peptides with signal anchors, and recognizing re-entrant helices that enter and exit the membrane on the same side. Recently, Nugent and Jones developed a SVM-based topology predictor and achieved the accuracies of 93, 64, and 89% for the prediction of the topology of proteins containing the signal peptide, topology of proteins containing the re-entrant helices, and the overall topology, respectively [97]. While the topology of transmembrane domains follows the so-called positive-inside rule, Seppälä *et al.* recently pointed out that a single positively charged residue, placed in different locations within not only any loop regions but even the region adjacent to the C-terminal, can control the global topology [98]. Bogdanov *et al.* showed that the final topology can also be dependent on the membrane lipid composition, too [99].

3.10 Sorting signals for bacterial protein transport

Gram-positive and -negative bacteria have four (cytoplasm, cytoplasmic membrane, cell wall, and extracellular space) and five (cytoplasm, cytoplasmic membrane, periplasm, outer membrane, and extraclellular space) major localization sites, respectively.

In order to cross or become integrated into the cytoplasmic membrane, proteins are transported either into an unfolded state through the Sec pathway or into a folded state through the Tat (twin-arginine translocation) pathway [100, 101]. Like usual (Sec-dependent) signal peptides, Tat-dependent signal peptides adapt the tripartite structure (*n*-, *h*-, and *c*-regions). But these signals are characterized by the [S/T]RR*X*FLK motif on the boundary between the *n*- and *h*-regions. In addition, they are less hydrophobic and tend to be longer, with a proline often appearing at the −6 position from the cleavage site [101]. Bendtsen *et al.* recently developed a prediction method for Tat signal peptides, based on an artificial neural network, which could predict 91 and 84% of Tat signal peptides and their cleavage sites, respectively [102].

The secretory pathways of Gram-negative bacteria are classified into at least five types (Types I–V) [103], which are either Sec dependent (Types II and V) or Sec independent (Types I, III, and IV). Recently, a sec-independent type VI secretion pathway was also identified [104, 105]. Since the sorting signals of most pathways are not fully understood, prediction of extracellular proteins is still difficult [4]. Recent analyses revealed that the N-terminal signal region of type III proteins is structurally flexible and that serines, threonines, and prolines are enriched, while leucines are poor in the signal region [106, 107]. Furthermore, the recently developed SOSUI-gramN system [108], which is based on classification through the physicochemical properties of the N- and C-termini, achieved higher precision and recall for extracellular protein than the existing methods, *i.e.* PSORTb, CELLO, and PSLpred [25, 109–111].

C-terminal signature sequences of β-barrel proteins that integrate into the outer membrane have been identified [112]. This sequence consists of a phenylalanine (or tryptophan) at the C-terminal position, and contains alternate hydrophobic residues at the −3, −5, −7, and −9 positions from the C-terminus like mitochondrial β-signals. Although the insertion mechanisms of β-barrel proteins have been poorly understood in bacteria and mitochondria, they may share common mechanisms according to the recent transport experiments of bacterial β-barrel proteins in yeast [113] and mitochondrial VDAC in *E. coli* [114]. Finally, the existence of two α-helical outer membrane proteins was recently identified although it was believed that all bacterial outer membrane proteins are the β-barrel type [115, 116]. These proteins would be a repertoire for the future predictors.

To conclude this section, we have reviewed the recent studies on a variety of protein sorting signals. Most of the knowledge has not been incorporated into prediction methods yet. As pointed out above, some of their motif patterns, such as the *h*-motif in signal peptides and the RLX_8D motif of the ER exit signal, may be very weak to be effectively used for prediction alone. However, the combination of these motifs and other features might be useful in improving the prediction accuracy. Further investigation in this regard is needed.

4 Difficulty of assessing prediction accuracy

Until now, a number of prediction methods for protein subcellular localization have been developed. To publish these results, it is common to compare the performance of a new method with that of others. Naturally, the claimed accuracy has been improved year by year – to the point where now some methods claim quite high performance (*e.g.* over 90%). In our opinion, however, the procedure for the comparison is sometimes problematic due to the redundancy of the data used.

In principle, there are two types of redundancy: one is the redundancy within the training and/or the test set, while the other is the redundancy between the training and test sets. Redundancy in the training set would lead to the overfitting of a predictor so that is ineffective for novel examples. Redundancy in the test set would be susceptible to a biased estimation of prediction quality because the prediction results for similar proteins would be amplified. More important is the redundancy between the two sets. If a protein from the test set is almost identical (*e.g.* $>50\%$ sequence similarity) to a protein in the training data set, the prediction is obvious and does not tell how the method would behave for a completely novel protein. Very often, one of these facts is violated. A typical case is a leave-one-out cross-validation on a training data set with proteins that share up to 80% sequence similarity.

In this section, we show results from our simple experiment, illustrating the importance of objective testing. This is a comparison of performance between the existing predictors and a simple homology-based method. In this age of flourishing subcellular proteomics [117–120], subcellular localizations of many proteins have been determined experimentally. For example, in the 6967 yeast proteins, 3458 of them are annotated with their subcellular localization information (Swiss-Prot/UniProt release 57.12) [121]. This implies the possibility that the subcellular localization of a significant fraction of proteins in any eukaryotic species could be predicted by a mere homology search. Thus, it is of special interest to see if the results from a general-purpose homology-based method match those from specifically developed state-of-the-art predictors of subcellular localization.

It is obvious that the performance of homology-based method is directly dependent on how many similar proteins are contained in the target database against the query data set. In principle, one would expect that homology-based methods are useless to novel (non-homologous) proteins. However, as we shown below, the situation is not so simple.

We used a simple BLAST-based prediction as the homology-based prediction, where the 'blastp' program was run against a set of amino acid sequences with the known subcellular localization sites. Among the hits, the localization site of the one with the lowest *E*-value except the query sequence itself was used as the predicted site. Note that the target database is regarded as a training data set as well as a test (query) set; the redundancy between the training and the test sets would be directly dependent on the redundancy of the database itself. As for the target database, we used two: one is the 12 807 annotated animal protein sequences from Swiss-Prot (release. 57.12), which are redundant (*i.e.* sequence identities less than 90%), and the other is their 6124 non-redundant representatives (*i.e.* sequence identities less than 30%) (see Table 1 and Supporting Information). We chose three prediction methods for evaluation, WoLF PSORT [122], CELLO ver2.5 [25, 110], and MultiLoc2 [26] (see Section 2), which can all use their nine locations in Table 1 as the prediction repertoire. Note that our attempt is not meant to be an objective test of these programs; it is likely that data sets used for the training of these algorithms overlap significantly with our data. As a performance measure, Matthews correlation coefficient (MCC) was used (for the definition of MCC, see [25], for example). The MCC values for all localization sites were averaged.

Table 1. Redundant and non-redundant animal protein data sets used for the performance comparison between prediction algorithms

	Redundant data	Non-redundant data
Total	12 807	6124
Extracellular space (Ex)	3984	1333
Cytoplasm (Cy)	1747	1056
Nucleus (Nu)	2961	2032
Mitochondria (Mi)	1370	619
Endoplasmic reticulum (ER)	364	237
Golgi apparatus (Go)	183	619
Lysosome (Ly)	149	63
Perooxisome (Pe)	103	48
Plasma membrane (Pm)	1946	597

Figure 1A shows the MCC of the four prediction methods for each localization site of the redundant data set. Obviously, the homology method is superior to the others: the highest averaged MCC (0.80) was achieved by this method while the 2nd highest MCC was 0.59 (by Multiloc2); in all sites, the homology method was the best. Next, we repeated the same test for the non-redundant (or less redundant, at least) data set (Fig. 1B). Surprisingly, the homology method still maintained a comparable performance level with the others: the average MCC of WoLF PSORT, CELLO v.2.5, and MultiLoc2 was 0.38, 0.48, and 0.57, respectively, while that of the homology method was 0.50. As pointed out above, prediction accuracy of the three algorithms may still be overestimated. A similar test has been tried by Briesemeister *et al.* [123], who showed that BLAST is less effective than other methods when an independent data set is used. One possible reason for this apparent discrepancy is the way of homology reduction in preparing data sets: Bacello test set [2] was made based on the similarity of local alignments, while ours was based on global similarity measurement, which may still leave locally homologous regions that are detectable by BLAST. Note that the target database itself is also reduced in homology in our test.

What can we learn from the above results? First, the state-of-the-art predictors are still not so practical in our objective test. At this stage, it may not make much sense to compare the difference of only a few percents of prediction accuracy between prediction algorithms. We still have much room for further development. Second, the relatively good result of the homology-based method for the 'non-redundant' data means that this data set is still redundant. Considering evolutionary history where gene duplications and speciation have played major roles, this may be reasonable. The internal similarity within our data sets seems to be a general and inherent problem in genome sequence analyses.

Before ending this section, we would like to propose a way of using the homology method for the discovery of biological importance. It seems quite interesting to see what happened when the homology method failed although there was a significantly homologous counterpart. There were 80 such pairs (Table 2). Although most of those pairs could be due to mere annotation errors, it is also possible that they

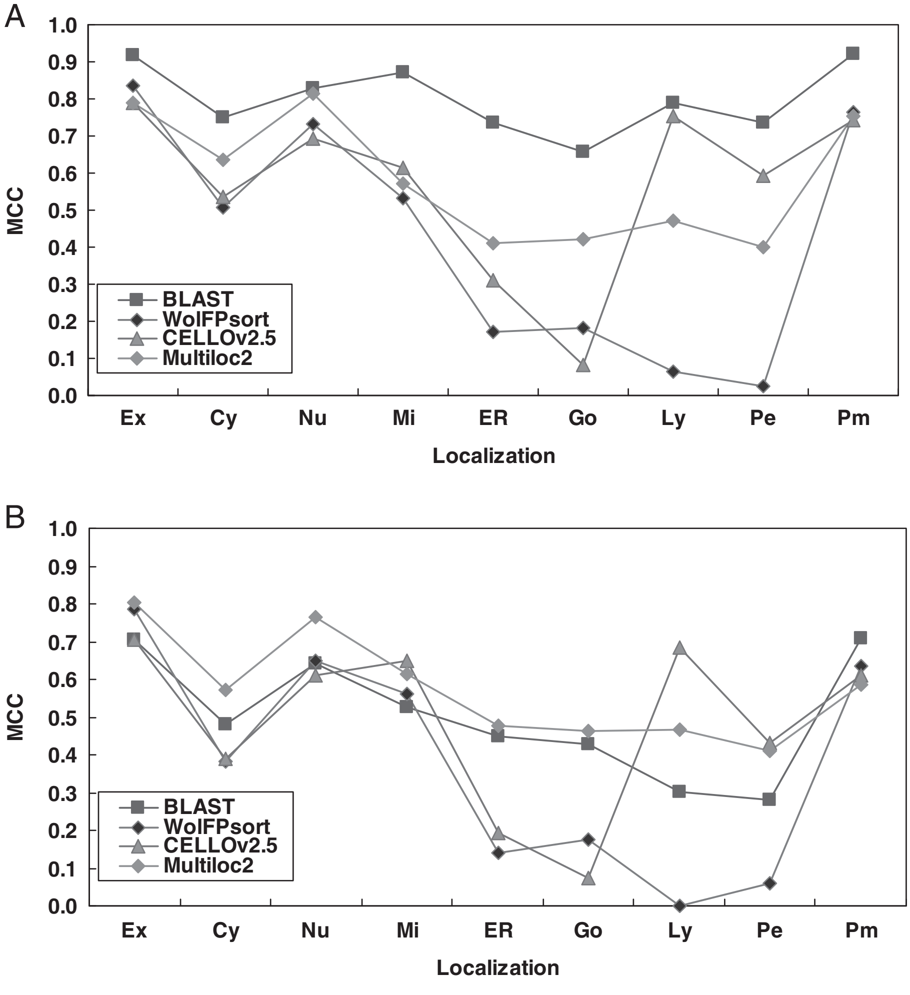

Figure 1. Comparison of performance between subcellular prediction methods on (A) redundant and (B) non-redundant data sets.

Table 2. Eighty protein pairs with significant homology but different annotations of subcellular location

Combination of subcellular sites of homolog pairs	Number of pairs	Combination of subcellular sites of homolog pairs	Number of pairs
Cy–Nu	17	Mi–ER	2
Mi–Cy	9	Mi–Nu	2
Ex–Ly	9	ER–Pm	2
Nu–Cy	7	Cy–Go	1
Cy–Mi	6	Nu–Ex	1
Cy–Pm	4	Nu–Pe	1
Cy–Pe	3	ER–Mi	1
Cy–Ex	3	Ly–Cy	1
Ex–Pm	3	Pe–Mi	1
ER–Go	3	Pm–Ex	1
Go–ER	3		

are isoforms that are localized at different sites. Therefore, it will be interesting to examine where their sequences are different and to see if these different regions could constitute the creation/deletion of sorting signals.

5 Concluding remarks

In this review, we have focused on subjects that may be useful for thinking about the future of this area. As described in Section 3, our knowledge on protein sorting signals is continuously expanding. Incorporation of such new knowledge into prediction algorithms should definitely be a future direction. However, such knowledge is still far from complete and insufficient to carry out comprehensive predictions against the entire proteomes. The main difficulty of clarifying all sorting signals seems to lie in their complexity, *i.e.* existence of multiple sorting pathways for one localization site. Therefore, we believe another promising direction for computational studies is the identification of novel sorting signals. In this sense, the study by Lin *et al.* [24], introduced in Section 2, is of special interest.

Another main topic of this review was a criticism of how to assess the performance of prediction methods. After a general discussion on the redundancy of the datasets, we introduced our simple test, where the performance of existing tools was not so good for novel proteins. To our surprise, although the performance of a simple homology-based method is highly dependent on the redundancy of the data sets, it performs at least at a comparable level to state-of-the-art predictors even if we used a data set with a 30% identity threshold. Thus, our conclusions are (i) a few percentages of performance difference between prediction methods do not have much significant meaning, and (ii) the problem of subcellular localization may not be suited for the benchmarking of machine-learning algorithms because the answer sets of amino acid sequences with known localization are inherently redundant. Therefore, developers of prediction methods in this field should pay more attention to their purposes. For example, development of programs specifically designed for the detection of potential annotation errors or for the finding of novel sorting signals would be more promising.

The authors thank Raymond Wan for critically reading the manuscript and Ziv Bar-Joseph for sending us his preprint. This work was partly supported by a Research Fellowship of the Japan Society for the Promotion of Science for Young Scientists.

The authors have declared no conflict of interest.

6 References

[1] Gaston, D., Tsaousis, A. D., Roger, A. J., Predicting proteomes of mitochondria and related organelles from genomic and expressed sequence tag data. *Methods Enzymol.* 2009, *457*, 21–47.

[2] Casadio, R., Martelli, P. L., Pierleoni, A., The prediction of protein subcellular localization from sequence: a shortcut to functional genome annotation. *Brief. Funct. Genomic. Proteomic.* 2008, *7*, 63–73.

[3] Chou, K. C., Shen, H. B., Recent progress in protein subcellular location prediction. *Anal. Biochem.* 2007, *370*, 1–16.

[4] Gardy, J. L., Brinkman, F. S., Methods for predicting bacterial protein subcellular localization. *Nat. Rev. Microbiol.* 2006, *4*, 741–751.

[5] Donnes, P., Hoglund, A., Predicting protein subcellular localization: past, present, and future. *Genomics Proteomics Bioinformatics* 2004, *2*, 209–215.

[6] Nakai, K., Horton, P., Computational prediction of subcellular localization, in: *Protein Targeting Protocols*, van der Giezen, M., (Ed.) 2007, Humana Press, Totowa, 429–466.

[7] von Heijne, G., A new method for predicting signal sequence cleavage sites. *Nucleic Acids Res.* 1986, *14*, 4683–4690.

[8] Nishikawa, K., Ooi, T., Correlation of the amino acid composition of a protein to its structural and biological characters. *J. Biochem.* 1982, *91*, 1821–1824.

[9] Kyte, J., Doolittle, R. F., A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 1982, *157*, 105–132.

[10] Nakai, K., Kanehisa, M., Expert system for predicting protein localization sites in Gram-negative bacteria. *Proteins* 1991, *11*, 95–110.

[11] Nakai, K., Kanehisa, M., A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics* 1992, *14*, 897–911.

[12] Nakai, K., Horton, P., PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci* 1999, *24*, 34–36.

[13] Goffeau, A., Nakai, K., Slonimski, P., Risler, J. L. *et al.*, The membrane proteins encoded by yeast chromosome III genes. *FEBS Lett.* 1993, *325*, 112–117.

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

[15] Emanuelsson, O., Nielsen, H., Brunak, S., von Heijne, G., Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* 2000, *300*, 1005–1016.

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

[17] Cedano, J., Aloy, P., Perez-Pons, J. A., Querol, E., Relation between amino acid composition and cellular location of proteins. *J. Mol. Biol.* 1997, *266*, 594–600.

[18] Nair, R., Rost, B., Mimicking cellular sorting improves prediction of subcellular localization. *J. Mol. Biol.* 2005, *348*, 85–100.

[19] Pierleoni, A., Martelli, P. L., Fariselli, P., Casadio, R., BaCelLo: a balanced subcellular localization predictor. *Bioinformatics.* 2006, *22*, e408–e416.

[20] Horton, P., Park, K.-J., Obayashi, T., Nakai, K., Protein subcellular localization prediction with WoLF PSORT. *Proc. 4th Asia-Pacific BIOINFORMATICS Conference (APBC2006)*, Jiang, T. *et al.* (Eds), Imperial College Press, 2006, pp. 39-48

[21] Chou, K. C., Shen, H. B., EuK-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. *J. Proteome Res.* 2007, *6*, 1728–1734.

[22] Lin, H. N., Chen, C. T., Sung, T. Y., Ho, S. Y. *et al.*, Protein subcellular localization prediction of eukaryotes using a knowledge-based approach. *BMC Bioinformatics* 2009, *10*, S8.

[23] Briesemeister, S., Rahnenfuhrer, J., Kohlbacher, O., Going from where to why – interpretable prediction of protein subcellular localization. *Bioinformatics* 2010, *26*, 1232–1238.

[24] Lin, T., Murphy, R. F., Bar-Joseph, Z., Discriminative motif finding for predicting protein subcellular localization. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 2010, in press

[25] Yu, C. S., Chen, Y. C., Lu, C. H., Hwang, J. K., Prediction of protein subcellular localization. *Proteins* 2006, *64*, 643–651.

[26] Blum, T., Briesemeister, S., Kohlbacher, O., MultiLoc2: integrating phylogeny and gene ontology terms improves subcellular protein localization prediction. *BMC Bioinformatics* 2009, *10*, 274.

[27] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D. *et al.*, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 2000, *25*, 25–29.

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

[29] Sigrist, C. J., Cerutti, L., de Castro, E., Langendijk-Genevaux, P. S. *et al.*, PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res.* 2010, *38* (Database issue), D161–D166.

[30] Finn, R. D., Mistry, J., Tate, J., Coggill, P. *et al.*, The Pfam protein families database. *Nucleic Acids Res.* 2010, *38* (Database issue), D211–D222.

[31] Lu, Z., Szafron, D., Greiner, R., Lu, P. *et al.*, Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics* 2004, *20*, 547–556.

[32] Brady, S., Shatkay, H., EpiLoc: a (working) text-based system for predicting protein subcellular location. *Pac. Symp. Biocomput.* 2008, 604–615.

[33] Fyshe, A., Liu, Y., Szafron, D., Greiner, R. *et al.*, Improving subcellular localization prediction using text classification and the gene ontology. *Bioinformatics* 2008, *24*, 2512–2517.

[34] Lee, K., Chuang, H. Y., Beyer, A., Sung, M. K. *et al.*, Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species. *Nucleic Acids Res* 2008, *36*, e136.

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

[36] von Heijne, G., Signal sequences. The limits of variation. *J. Mol. Biol.* 1985, *184*, 99–105.

[37] Nakai, K., Signal peptides, in *Cell-penetrating Peptides: Processes and Applications*, U. Langel (Ed) 2002, CRC Press, Boca Raton, pp. 295–324.

[38] Duffy, J., Patham, B., Mensa-Wilmot, K., Discovery of functional motifs in *h*-regions of trypanosome signal sequences. *Biochem. J.* 2010, *426*, 135–145.

[39] Kaiser, C. A., Preuss, D., Grisafi, P., Botstein, D., Many random sequences functionally replace the secretion signal sequence of yeast invertase. *Science* 1987, *235*, 312–317.

[40] Nilsson, I., Whitley, P., von Heijne, G., The COOH-terminal ends of internal signal and signal-anchor sequences are positioned differently in the ER translocase. *J. Cell Biol.* 1994, *126*, 1127–1132.

[41] Chou, M. M., Kendall, D. A., Polymeric sequences reveal a functional interrelationship between hydrophobicity and length of signal peptides. *J. Biol. Chem.* 1990, *265*, 2873–2880.

[42] Hiss, J. A., Schneider, G., Architecture, function and prediction of long signal peptides. *Brief. Bioinform.* 2009, *10*, 569–578.

[43] Dultz, E., Hildenbeutel, M., Martoglio, B., Hochman, J. *et al.*, The signal peptide of the mouse mammary tumor virus rem protein is released from the endoplasmic reticulum membrane and accumulates in nucleoli. *J. Biol. Chem.* 2008, *283*, 9966–9976.

[44] Bendtsen, J. D., Nielsen, H., von Heijne, G., Brunak, S., Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* 2004, *340*, 783–795.

[45] Chou, K. C., Shen, H. B., Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. *Biochem. Biophys. Res. Commun.* 2007, *357*, 633–640.

[46] Lange, A., Mills, R. E., Lange, C. J., Stewart, M. *et al.*, Classical nuclear localization signals: definition, function,

and interaction with importin alpha. *J. Biol. Chem.* 2007, *282*, 5101–5105.

[47] Kosugi, S., Hasebe, M., Matsumura, N., Takashima, H. *et al.*, Six classes of nuclear localization signals specific to different binding grooves of importin alpha. *J. Biol. Chem.* 2009, *284*, 478–485.

[48] Lange, A., McLane, L. M., Mills, R. E., Devine, S. E. *et al.*, Expanding the definition of the classical bipartite nuclear localization signal. *Traffic* 2010, *11*, 311–323.

[49] Lee, B. J., Cansizoglu, A. E., Suel, K. E., Louis, T. H. *et al.*, Rules for nuclear localization sequence recognition by karyopherin beta 2. *Cell* 2006, *126*, 543–558.

[50] Cokol, M., Nair, R., Rost, B., Finding nuclear localization signals. *EMBO Rep.* 2000, *1*, 411–415.

[51] Brameier, M., Krings, A., MacCallum, R. M., Nucpred – predicting nuclear localization of proteins. *Bioinformatics*, 2007, *23*, 1159–1160.

[52] Fischer, U., Huber, J., Boelens, W. C., Mattaj, I. W. *et al.*, The HIV-1 Rev activation domain is a nuclear export signal that accesses an export pathway used by specific cellular RNAs. *Cell* 1995, *82*, 475–483.

[53] Bogerd, H. P., Fridell, R. A., Benson, R. E., Hua, J. *et al.*, Protein sequence requirements for function of the human T-cell leukemia virus type 1 Rex nuclear export signal delineated by a novel *in vivo* randomization-selection assay. *Mol. Cell. Biol.* 1996, *16*, 4207–4214.

[54] la Cour, T., Kiemer, L., Molgaard, A., Gupta, R. *et al.*, Analysis and prediction of leucine-rich nuclear export signals. *Protein Eng. Des. Sel.* 2004, *17*, 527–536.

[55] Kosugi, S., Hasebe, M., Tomita, M., Yanagawa, H., Nuclear export signal consensus sequences defined using a localization-based yeast selection system. *Traffic* 2008, *9*, 2053–2062.

[56] Dong, X., Biswas, A., Suel, K. E., Jackson, L. K. *et al.*, Structural basis for leucine-rich nuclear export signal recognition by crm1. *Nature* 2009, *458*, 1136–1141.

[57] Neupert, W., Herrmann, J. M., Translocation of proteins into mitochondria. *Annu. Rev. Biochem.* 2007, *76*, 723–749.

[58] Obita, T., Muto, T., Endo, T., Kohda, D., Peptide library approach with a disulfide tether to refine the Tom20 recognition motif in mitochondrial presequences. *J. Mol. Biol.* 2003, *328*, 495–504.

[59] Habib, S. J., Neupert, W., Rapaport, D., Analysis and prediction of mitochondrial targeting signals. *Methods Cell Biol.* 2007, *80*, 761–781.

[60] Gakh, O., Cavadini, P., Isaya, G., Mitochondrial processing peptidases. *Biochim. Biophys. Acta* 2002, *1592*, 63–77.

[61] ogtle, F. N., Wortelkamp, S., Zahedi, R. P., Becker, D. *et al.*, Global analysis of the mitochondrial n-proteome identifies a processing peptidase critical for protein stability. *Cell* 2009, *139*, 428–439.

[62] Claros, M. G., Vincens, P., Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur. J. Biochem.* 1996, *241*, 779–786.

[63] Wiedemann, N., Kozjak, V., Chacinska, A., Schonfisch, B. *et al.*, Machinery for protein sorting and assembly in the mitochondrial outer membrane. *Nature* 2003, *424*, 565–571.

[64] Paschen, S. A., Waizenegger, T., Stan, T., Preuss, M. *et al.*, Evolutionary conservation of biogenesis of beta-barrel membrane proteins. *Nature* 2003, *426*, 862–866.

[65] Kutik, S., Stojanovski, D., Becker, L., Becker, T. *et al.*, Dissecting membrane insertion of mitochondrial beta-barrel proteins. *Cell* 2008, *132*, 1011–1024.

[66] Hiller, S., Garces, R. G., Malia, T. J., Orekhov, V. Y. *et al.*, Solution structure of the integral human membrane protein VDAC-1 in detergent micelles. *Science* 2008, *321*, 1206–1210.

[67] Bayrhuber, M., Meins, T., Habeck, M., Becker, S. *et al.*, Structure of the human voltage-dependent anion channel. *Proc. Natl. Acad. Sci. USA* 2008, *105*, 15370–15375.

[68] Imai, K., Gromiha, M. M., Horton, P., Mitochondrial beta-barrel proteins, an exclusive club? *Cell* 2008, *135*, 1158–1159; author reply 1159–1160.

[69] Wimley, W. C., The versatile beta-barrel membrane protein. *Curr. Opin. Struct. Biol.* 2003, *13*, 404–411.

[70] Jarvis, P., argeting of nucleus-encoded proteins to chloroplasts in plants. *New Phytol.* 2008, *179*, 257–285.

[71] Bruce, B. D., Chloroplast transit peptides: structure, function and evolution. *Trends Cell Biol.* 2000, *10*, 440–447.

[72] Zhang, X. P., Glaser, E., Interaction of plant mitochondrial and chloroplast signal peptides with the Hsp70 molecular chaperone. *Trends Plant Sci.* 2002, *7*, 14–21.

[73] Pujol, C., Marechal-Drouard, L., Duchene, A. M., How can organellar protein N-terminal sequences be dual targeting signals? *In silico* analysis and mutagenesis approach. *J. Mol. Biol.* 2007, *369*, 356–367.

[74] Lee, D. W., Kim, J. K., Lee, S., Choi, S. *et al.*, Arabidopsis nuclear-encoded plastid transit peptides contain multiple sequence subgroups with distinctive chloroplast-targeting sequence motifs. *Plant Cell* 2008, *20*, 1603–1622.

[75] Zybailov, B., Rutschow, H., Friso, G., Rudella, A. *et al.*, Sorting signals, N-terminal modifications and abundance of the chloroplast proteome. *PLoS One* 2008, *3*, e1994.

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

[77] Carrie, C., Giraud, E., Whelan, J., Protein transport in organelles: dual targeting of proteins to mitochondria and chloroplasts. *FEBS J.* 2009, *276*, 1187–1195.

[78] Peeters, N., Small, I., Dual targeting to mitochondria and chloroplasts. *Biochim. Biophys. Acta* 2001, *1541*, 54–63.

[79] Brocard, C., Hartig, A., Peroxisome targeting signal 1: is it really a simple tripeptide? *Biochim. Biophys. Acta* 2006, *1763*, 1565–1573.

[80] Neuberger, G., Maurer-Stroh, S., Eisenhaber, B., Hartig, A. *et al.*, Motif refinement of the peroxisomal targeting signal 1 and evaluation of taxon-specific differences. *J. Mol. Biol.* 2003, *328*, 567–579.

[81] Reumann, S., Babujee, L., Ma, C., Wienkoop, S. *et al.*, Proteome analysis of arabidopsis leaf peroxisomes reveals

novel targeting peptides, metabolic pathways, and defense mechanisms. *Plant Cell* 2007, *19*, 3170–3193.

[82] Reumann, S., Quan, S., Aung, K., Yang, P. *et al.*, In-depth proteome analysis of arabidopsis leaf peroxisomes combined with *in vivo* subcellular targeting verification indicates novel metabolic and regulatory functions of peroxisomes. *Plant Physiol.* 2009, *150*, 125–143.

[83] Rachubinski, R. A., Subramani, S., How proteins penetrate peroxisomes. *Cell* 1995, *83*, 525–528.

[84] Petriv, O. I., Tang, L., Titorenko, V. I., Rachubinski, R. A., A new definition for the consensus sequence of the peroxisome targeting signal type 2. *J. Mol. Biol.* 2004, *341*, 119–134.

[85] Van Ael, E., Fransen, M., Targeting signals in peroxisomal membrane proteins. *Biochim. Biophys. Acta* 2006, *1763*, 1629–1638.

[86] Rottensteiner, H., Kramer, A., Lorenzen, S., Stein, K. *et al.*, Peroxisomal membrane proteins contain common Pex19p-binding sites that are an integral part of their targeting signals. *Mol. Biol. Cell* 2004, *15*, 3406–3417.

[87] Halbach, A., Lorenzen, S., Landgraf, C., Volkmer-Engert, R. *et al.*, Function of the Pex19-binding site of human adrenoleukodystrophy protein as targeting motif in man and yeast. Pmp targeting is evolutionarily conserved. *J. Biol. Chem.* 2005, *280*, 21176–21182.

[88] Schluter, A., Real-Chicharro, A., Gabaldon, T., Sanchez-Jimenez, F. *et al.*, PeroxisomeDB 2.0: an integrative view of the global peroxisomal metabolome. *Nucleic Acids Res.* 2010, *38*(Database issue), D800–D805.

[89] Thoms, S., Debelyy, M. O., Nau, K., Meyer, H. E. *et al.*, Lpx1p is a peroxisomal lipase required for normal peroxisome morphology. *FEBS J.* 2008, *275*, 504–514.

[90] Islinger, M., Li, K. W., Seitz, J., Volkl, A. *et al.*, Hitchhiking of Cu/Zn superoxide dismutase to peroxisomes – evidence for a natural piggyback import mechanism in mammals. *Traffic* 2009, *10*, 1711–1721.

[91] Barlowe, C., Signals for COPII-dependent export from the ER: what's the ticket out? *Trends Cell Biol.* 2003, *13*, 295–300.

[92] Fernandez-Sanchez, E., Diez-Guerra, F. J., Cubelos, B., Gimenez, C. *et al.*, Mechanisms of endoplasmic-reticulum export of glycine transporter-1 (glyt1). *Biochem. J.* 2008, *409*, 669–681.

[93] Tsukumo, Y., Tsukahara, S., Saito, S., Tsuruo, T. *et al.*, A novel endoplasmic reticulum export signal: Proline at the+2-position from the signal peptide cleavage site. *J. Biol. Chem.* 2009, *284*, 27500–27510.

[94] Duvernay, M. T., Dong, C., Zhang, X., Robitaille, M. *et al.*, A single conserved leucine residue on the first intracellular loop regulates er export of g protein-coupled receptors. *Traffic* 2009.

[95] Ellgaard, L., Helenius, A., Quality control in the endoplasmic reticulum. *Nat. Rev. Mol. Cell. Biol.* 2003, *4*, 181–191.

[96] Braulke, T., Bonifacino, J. S., Sorting of lysosomal proteins. *Biochim. Biophys. Acta.* 2009, *1793*, 605–614.

[97] Nugent, T., Jones, D. T., Transmembrane protein topology prediction using support vector machines. *BMC Bioinformatics* 2009, *10*, 159.

[98] Seppala, S., Slusky, J. S., Lloris-Garcera, P., Rapp, M. *et al.*, Control of membrane protein topology by a single C-terminal residue. *Science* 2010, *328*, 1698–1700.

[99] Bogdanov, M., Xie, J., Heacock, P., Dowhan, W., To flip or not to flip: lipid-protein charge interactions are a determinant of final membrane protein topology. *J. Cell Biol.* 2008, *182*, 925–935.

[100] Driessen, A. J., Nouwen, N., Protein translocation across the bacterial cytoplasmic membrane. *Annu. Rev. Biochem.* 2008, *77*, 643–667.

[101] De Buck, E., Lammertyn, E., Anne, J., The importance of the twin-arginine translocation pathway for bacterial virulence. *Trends Microbiol.* 2008, *16*, 442–453.

[102] Bendtsen, J. D., Nielsen, H., Widdick, D., Palmer, T. *et al.*, Prediction of twin-arginine signal peptides. *BMC Bioinformatics* 2005, *6*, 167.

[103] Kostakioti, M., Newman, C. L., Thanassi, D. G., Stathopoulos, C., Mechanisms of protein export across the bacterial outer membrane. *J. Bacteriol.* 2005, *187*, 4306–4314.

[104] Cascales, E., The type VI secretion toolkit. *EMBO Rep.* 2008, *9*, 735–741.

[105] Pukatzki, S., McAuley, S. B., Miyata, S. T., The type VI secretion system: translocation of effectors and effector-domains. *Curr. Opin. Microbiol.* 2009, *12*, 11–17.

[106] Arnold, R., Brandmaier, S., Kleine, F., Tischler, P. *et al.*, Sequence-based prediction of type III secreted proteins. *PLoS Pathog.* 2009, *5*, e1000376.

[107] Samudrala, R., Heffron, F., McDermott, J. E., Accurate prediction of secreted substrates and identification of a conserved putative secretion signal for type III secretion systems. *PLoS Pathog.* 2009, *5*, e1000375.

[108] Imai, K., Asakawa, N., Tsuji, T., Akazawa, F. *et al.*, SOSUI-GramN performance prediction for sub-cellular localization of proteins in Gram-negative bacteria. *Bioinformation* 2008, *2*, 417–421.

[109] Gardy, J. L., Laird, M. R., Chen, F., Rey, S. *et al.*, PSORTb v.2.0: prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics* 2005, *21*, 617–623.

[110] Yu, C. S., Lin, C. J., Hwang, J. K., Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. *Protein Sci.* 2004, *13*, 1402–1406.

[111] Bhasin, M., Garg, A., Raghava, G. P., Pslpred: prediction of subcellular localization of bacterial proteins. *Bioinformatics* 2005, *21*, 2522–2524.

[112] Robert, V., Volokhina, E. B., Senf, F., Bos, M. P. *et al.*, Assembly factor Omp85 recognizes its outer membrane protein substrates by a species-specific c-terminal motif. *PLoS Biol.* 2006, *4*, e377.

[113] Walther, D. M., Papic, D., Bos, M. P., Tommassen, J. *et al.*, Signals in bacterial beta-barrel proteins are functional in

eukaryotic cells for targeting to and assembly in mitochondria. *Proc. Natl. Acad. Sci. USA* 2009, *106*, 2531–2536.

[114] Walther, D. M., Bos, M. P., Rapaport, D., Tommassen, J., The mitochondrial porin, VDAC, has retained the ability to be assembled in the bacterial outer membrane. *Mol. Biol. Evol.* 2010, *27*, 887–895.

[115] Dong, C., Beis, K., Nesper, J., Brunkan-Lamontagne, A. L. *et al.*, Wza the translocon for *E. coli* capsular polysaccharides defines a new class of membrane protein. *Nature* 2006, *444*, 226–229.

[116] Chandran, V., Fronzes, R., Duquerroy, S., Cronin, N. *et al.*, Structure of the outer membrane complex of a type IV secretion system. *Nature* 2009, *462*, 1011–1015.

[117] Kumar, A., Agarwal, S., Heyman, J. A., Matson, S. *et al.*, Subcellular localization of the yeast proteome. *Genes Dev.* 2002, *16*, 707–719.

[118] Huh, W. K., Falvo, J. V., Gerke, L. C., Carroll, A. S. *et al.*, Global analysis of protein localization in budding yeast. *Nature* 2003, *425*, 686–691.

[119] Matsuyama, A., Arai, R., Yashiroda, Y., Shirai, A. *et al.*, Orfeome cloning and global analysis of protein localization in the fission yeast *Schizosaccharomyces pombe*. *Nat. Biotechnol.* 2006, *24*, 841–847.

[120] Anderen, J. S., Wilkinson, C. J., Mayor, T., Mortensen, P. *et al.*, Proteomic characterization of the human centrosome by protein correlation profiling. *Nature* 2003, *426*, 570–574.

[121] UniProt Consortium. The universal protein resource (UniProt) in 2010*. Nucleic Acids Res.* 2010, *38* (Database issue), D142–D148.

[122] Horton, P., Park, K. J., Obayashi, T., Fujita, N. *et al.*, WoLF PSORT: Protein localization predictor. *Nucleic Acids Res.* 2007, *35* (Web Server issue), W585–W587.

[123] Briesemeister, S., Blum, T., Brady, S., Lam, Y. *et al.*, SherLoc2: a high-accuracy hybrid method for predicting subcellular localization of proteins. *J. Proteome Res.* 2009, *8*, 5363–5366.