Olof Emanuelsson
works in close collaboration with Gunnar von Heijne on developing methods for predicting subcellular localisation of proteins. He is at the Stockholm Bioinformatics Center (SBC), a joint effort between Stockholm University, the Royal Institute of Technology and Karolinska Institute, located in Stockholm, Sweden.

Keywords: *protein sorting, subcellular localisation, prediction, sequence analysis*

Predicting protein subcellular localisation from amino acid sequence information

Olof Emanuelsson

Date received (in revised form): 10th September 2002

Abstract

Predicting the subcellular localisation of proteins is an important part of the elucidation of their functions and interactions. Here, the amino acid sequence motifs that direct proteins to their proper subcellular compartment are surveyed, different methods for localisation prediction are discussed, and some benchmarks for the more commonly used predictors are presented.

INTRODUCTION

Computationally based characterisation of the features of the proteins found or predicted in completely sequenced genomes is an important task in the search for knowledge of protein function. A central issue is to predict the subcellular localisation, which has implications both for the function of the protein and its possibility of interacting with other proteins.[1,2]

In a prokaryotic cell, basically only three locations are possible: inside or outside the plasma membrane, or inserted into the membrane. In Gram-negative bacteria, the presence of an outer membrane adds two possible locations, the periplasmic space and the outer membrane. A eukaryotic cell, on the other hand, is full of various membrane-surrounded compartments: the mitochondrion, the microbodies (peroxisomes/glyoxysomes/glycosomes), and the nucleus to mention just a few. Additionally, plants and algae (and some parasites) also contain plastids, such as the chloroplast where the photosynthetic reaction takes place.

The prevailing principal mechanism of protein sorting is that an amino acid signal in the protein is recognised by some kind of import machinery on the surface of the compartment into which the protein is to be transferred. This is often carried out with the help of chaperones, soluble proteins in the cytoplasm that guide the protein in question to the surface of its final compartment. Protein translocation across membranes often demands an ATP- or GTP-dependent active transport or at least a membrane potential.[3] However, exceptions to this exist, such as the nuclear pore complexes, through which small proteins (<60 kDa) can diffuse more or less freely into (and out from) the nucleus.[4] The recognised signal is in most cases present and detectable on the primary sequence level (and predominantly located at the N-terminus), but, obviously, structural considerations are important for a full understanding of protein localisation mechanisms.[5,6]

A wide variety of methods have been tried throughout the years in order to predict the subcellular localisation of proteins. The methods differ in terms of what input data they demand and what technique is employed to make the decision (prediction) about location. Once the input data type is fixed, the methods for prediction-making are basically of two types: construct manually explicit rules for localisation prediction from current knowledge of sorting signals, or apply data-driven machine learning techniques (eg neural networks or hidden Markov models, HMMs) that automatically extract decision rules from sets of proteins with known location, but

Olof Emanuelsson,
Stockholm Bioinformatics Center,
Stockholm Center for Physics
Astronomy and Biotechnology,
Stockholm University,
S-106 91 Stockholm,
Sweden

Tel: +46 (0) 8 5537 8574
Fax: +46 (0) 8 5537 8214
E-mail: olof@sbc.su.se

without the need for making any detailed prior assumptions as to what features it is interesting to look for. (However, it is always advisable to incorporate as much biological knowledge as possible even into an 'automatic scheme'.)

Different prediction performance for different locations

The degree of performance differs markedly between different predictors and, perhaps biologically more intriguing, also between different compartments. For some compartments this is partly due to lack of sufficient data for rule construction (be it manual or data-driven), but there is also a difference in performance that is due to the complexity of the signal. Some protein locations are simply easier to predict than others.

Signal peptides (SP) for secretion consist of three regions

This paper is an attempt to review the amino acid sequence motifs that direct proteins to their proper subcellular compartment, to discuss the available methods for localisation prediction, and to present some benchmarks for the more commonly used multi-category predictors.

THE COMPARTMENTS AND THEIR SORTING SIGNALS

The major eukaryotic compartments will be presented along with their sorting signals; see also Table 1 and Figure 1. For a more extensive description of protein sorting there is an excellent review by Nakai.[7]

Secreted proteins

The basic function of trans-locating proteins from inside the cytoplasmic membrane to the outside seems to be vital to most organisms.[8] Accordingly, protein secretion is one of the most studied protein translocation processes, both for its great biological importance and for its commercial potential in, for example, drug manufacturing. The signal peptide (SP) for secretion is N-terminal and approximately 20–25 residues long. It is cleaved off by the signal peptidase (SPase) during the export process.[9,10] Small and apolar residues (preferably alanine) are found at positions -1 and -3 relative to the cleavage site. The -3, -1 rule is a motif that was observed many years ago.[11] Another feature is the regional structure of the SP, with an N-terminal domain that is positively charged, a hydrophobic central region (*h*-region; leucines are most common), and a cleavage site region with mostly small residues,[12,13] Figures 1 and 2.

Table 1: Summary of common eukaryotic protein sorting signals

Destination	Name of signal	Typical length
Extracellular (secreted)	Signal peptide, SP	20–30
Mitochondrion (matrix)	Mitochondrial transfer peptide, mTP	25–45
Chloroplast	Chloroplast transit peptide, cTP	40–70
Thylakoid	Lumenal transfer peptide, lTP	20–30
Nucleus	Nuclear localisation signal (mono-partite), NLS	4–6
Nucleus	Nuclear localisation signal (bi-partite), NLS	15–20
Peroxisome	Peroxisomal targeting signal 1, PTS1	3
Peroxisome	Peroxisomal targeting signal 2, PTS2	9

There are several different pathways for secretion present in both eukaryotes and prokaryotes. The two most common are the SRP-dependent (signal recognition particle) pathway[15,16] and the SRP-independent pathway, which is usually called the Sec-dependent pathway in prokaryotes.[17] In prokaryotes, the delta-pH or TAT (twin-arginine translocation) pathway is yet another option for secretion.[18,19] The SRP-dependent pathway relies on the recognition of nascent polypeptide chains by the SRP, which halts translation and brings the translation complex to the SRP receptor, where translocation (through the Sec machinery) subsequently occurs cotranslationally. The Sec-dependent (or SRP-independent) pathway uses only the Sec machinery, which involves many proteins and the hydrolysis of ATP, for identification of the signal peptide and translocation, which is post-translational. The delta-pH pathway translocation consumes no ATP but, as the name suggests, requires a pH-gradient over the membrane. Proteins transported via this route contain a twin-arginine motif in the N-terminal part of the signal peptide, and the signal peptide is in general somewhat longer.

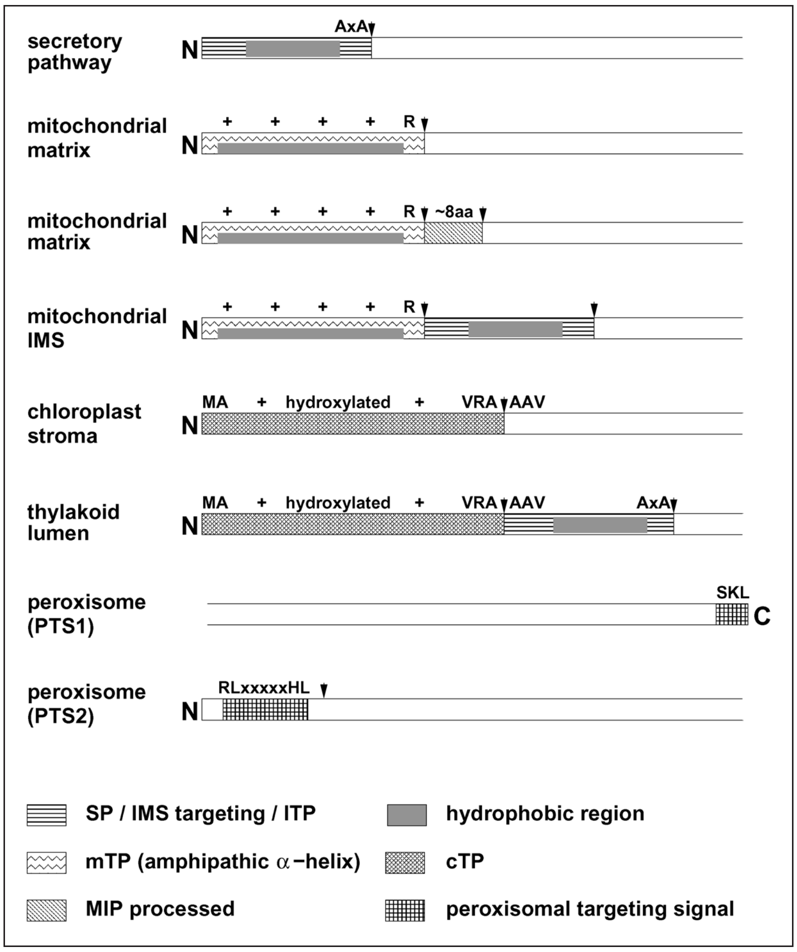

Figure 1: Schematic view of sorting signals, the corresponding final compartments, and reported sequence features. *Arrowhead*, cleavage site; *SP*, signal peptide; *cTP*, chloroplast transit peptide; *mTP*, mitochondrial targeting peptide; *IMS*, intermembrane space (in mitochondria); *MIP*, mitochondrial intermediate peptidase; *PTS*, peroxisomal targeting signal; *aa*, amino acids. A = Alanine; x = any amino acid; R = Arginine; M = Methionine; V = Valine; S = Serine; K = Lysine; L = Lucine; H = Histidine.

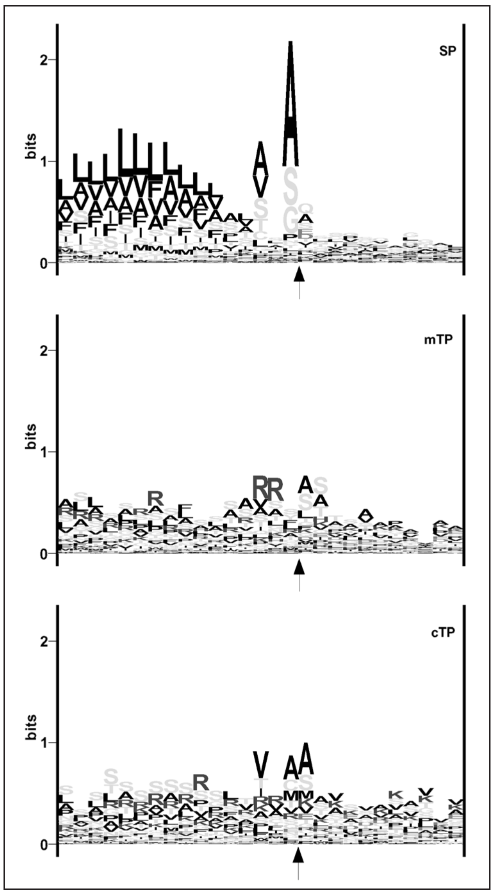

Figure 2: Sequence logos[14] of 269 secreted (SP, *upper panel*), 368 mitochondrial (mTP, *middle panel*), and 141 chloroplast (cTP, *lower panel*) proteins, aligned around their annotated cleavage site (*arrow*). A sequence logo is a way to visualise a multiple sequence alignment and, specifically, the degree of amino acid conservation at the positions in the alignment. The scale on the *y*-axis is measured in bits, and for protein sequences it has a maximum value of 4.3 which would correspond to a totally conserved position

In eukaryotes, proteins that are to be secreted are translated either on ribosomes attached to the endoplasmic reticulum (ER) membrane, or on free ribosomes in the cytosol. In both cases, it is the signal peptide that directs the protein into the ER, where the signal is cleaved off and degraded, and from which the protein is further directed to the outside of the cell.[20] Furthermore, there is no TAT pathway for protein secretion in eukaryotes, except for the export of proteins from the chloroplast stroma into the thylakoid lumen (see thylakoid section).[19]

Many other variations on the signal peptide theme exist. There is a particular class of uncleaved signal peptides, termed signal anchors (SAs), which are stuck in the membrane during the translocation.[21] The result is that the entire protein is anchored in the membrane, hence the name. SAs differ from SPs not only in the cleavage site region: in general they also have a longer *h*-region, typically of the same length as a transmembrane α-helix. Another important type of membrane association is mediated via glycosylphosphatidylinositol (GPI). A

The mitochondrial targeting peptide (mTP) is N-terminal

protein is first targeted to the ER by its (cleaved) SP, where the C-terminus of the protein is cleaved off, while a GPI anchor is covalently bound to the so-called π-site.[22,23]

Membranes

There is no known generic signal for localisation into the different membranes in the cell. However, most transmembrane regions are present in the form of an α-helix, with the notable exception of β-barrel proteins.[24] For helical transmembrane proteins, the mere fact that the transmembrane regions are significantly more hydrophobic than an average piece of sequence has been used as an indication of transmembrane location.[25] Transmembrane regions range between 14 and 36 residues in length,[26] depending on the angle between the helix and the membrane (tilting) and the kind of membrane the protein resides in, since different membranes have a different thickness. A striking feature of cell membrane proteins is their skewed distribution of charges between inner and outer loops, the positive-inside rule, and this has also been used for topology determination in predictive methods.[27]

Two types of membrane proteins: α-helical and β-barrel

The 'positive-reside' rule is a topology determinant

The majority of eukaryotic membrane proteins are inserted into the ER membrane employing the translocon complex used for protein secretion. It has been proposed that the insertion is a result of a stop-transfer process, where the translocation of the protein is halted when a stop-transfer signal is encountered in the sequence (cf. the description of signal anchors in previous section). Various models have been proposed for the exact molecular mechanisms for the subsequent insertion of transmembrane regions of multi-spanning proteins.[28–30]

The mitochondrion is the result of an endosymbiotic event

In outer membranes, such as bacterial (Gram-negative) outer membranes and chloroplast and mitochondrial outer membranes, the predominant protein class is the β-barrel proteins. They consist of an even number of β-strands (from 8 to 22) and their functions include, for example, both passive nutrient import (of molecules < 6 kDa) and active ion transport.[31]

Another interesting group of membrane-associated proteins are the ones that demand both N-terminal myristoylation (PROSITE[32] motif is G-{EDRKHPFYW}-x(2)-[STAGCN]-{P}) and palmitoylation (which occurs at a Cys residue) for their proper localisation to the membrane,[33,34] and it has been suggested that the reversible nature of palmitoylation may function as a regulator of subcellular localisation in this case.

Mitochondrion

It is believed that the mitochondrion is the result of an endosymbiotic uptake of an ancestor to what today is alpha-proteobacteria into an early eukaryotic host.[35] The evolutionary timing of the endosymbiotic event is still being discussed,[36] but in any case, most of the prokaryotic genes have been transferred to the host nuclear genome during the course of evolution.

Two membranes, the inner and the outer mitochondrial membranes, surround the mitochondrial matrix. Proteins destined to the mitochondrion usually contain an N-terminal mitochondrial transfer peptide (mTP), on average 35 amino acids long, which is recognised by the protein import machinery, the TOM proteins (translocation machinery of the outer membrane), on the mitochondrial surface. The mTP initially interacts with the TOM20 receptor[37] and the protein is transported, via the GIP complex (general import pore), in an ATP-requiring process through the outer mitochondrial membrane, and further through the inner membrane by a complex of TIM proteins (translocation machinery of the inner membrane) requiring a membrane potential.[38,39]

After having entered the mitochondrial matrix, the mTP is cleaved off by the mitochondrial processing peptidase, MPP.[40,41] Some mitochondrial proteins are then cleaved a second time by the mitochondrial intermediate peptidase,

The chloroplast is the result of an endosymbiotic event

The chloroplast transit peptide (cTP) is N-terminal

Some proteins have a dual location

Thylakoid proteins have a 'bi-partite' localisation signal

MIP (Figure 1), which cuts off an additional eight or nine residues from the N-terminus.[42,43] For yet other proteins, a second adjacent targeting signal that resembles the signal peptide for secretion is exposed after MPP cleavage. These proteins are re-exported from the matrix to the intermembrane space (IMS), or inserted into the inner membrane, in a process very similar to bacterial protein secretion (it has accordingly been termed conservative sorting). Alternatively, the translocation over either of the membranes is halted by a stop-transfer signal, which is specifically recognised by a TOM or TIM component,[44,45] and the protein is subsequently inserted into the outer or inner membrane, respectively. There are also a few mitochondrial proteins (mostly inner membrane metabolite carriers) that have been shown to contain internal (ie not N-terminal) localisation signals.[46] Thus, the mitochondrial protein import machinery seems to be extremely versatile.[39]

In mTP sequences there is an over-representation of Arg, Ala and Ser, while negatively charged residues (Asp, Glu) are scarce.[47] Otherwise, it is hard to find any obvious features that distinguish the mTP from other N-terminal sequences and the degree of sequence conservation around the cleavage site is rather poor. It has been reported that many mTPs have an arginine in position -2 or -3 relative to the MPP cleavage site;[48,49] Figures 1 and 2. Additionally, it was recently confirmed in a nuclear magnetic resonance (NMR) structure that, as expected by theoretical studies, the mTP forms an amphipathic α-helix when bound to the receptor protein, TOM20.[5] On the other hand, when processed by the MPP, it adopts an extended structure.[50]

Chloroplast

The chloroplast is an organelle present in photosynthetic plants and algae and, like the mitochondrion, it is believed to be of bacterial origin.[51] Thus, it has a small genome of its own, a reminiscence from its pre-endosymbiotic days. The majority of chloroplast proteins are encoded in the nuclear genome and post-translationally imported into the organelle. Virtually all chloroplast proteins encoded in the nucleus have an N-terminal chloroplast transit peptide (cTP) recognised by cytosolic chaperones, whereupon the complex docks to the Toc (translocon at the outer membrane) machinery and the protein is then further transported via the Tic (translocon at the inner membrane) complex into the chloroplast stroma in an ATP- and GTP-dependent manner.[52,53] Upon entry, the cTP is cleaved off by the stromal processing peptidase (SPP).[54]

cTPs from different proteins show a wide variation in length (20–120 residues, average is 55) and sequence, but they tend to be rich in hydroxylated residues, especially serines, and to have a low content of acidic residues.[47] cTPs from algae are in general shorter than those from higher plants.[55] At the N-terminus there is a conserved alanine next to the initial methionine. A semi-conserved motif, V-R-A-(↓)-A-A-V, around the SPP cleavage site (arrow) has also been identified;[56] Figures 1 and 2. The signal is not very strong and there are actually several examples of proteins that are located to both mitochondria and chloroplasts using identical sorting signals.[57,58]

Thylakoid

Proteins designated for the lumen of the intra-chloroplastic thylakoid compartment generally have a bi-partite targeting sequence composed of an N-terminal stroma-targeting cTP followed by a thylakoid lumen transfer peptide (LTP) that shares important features with signal sequences required for protein secretion in bacteria (Figure 1).[10,59] This fact is due to the endosymbiotic origin of the chloroplast, where a photosynthetic cyanobacteria-like prokaryote was engulfed by a eukaryote, and consequently, the stroma to lumen transport is topologically equivalent to bacterial protein secretion.[51,60]

There are at least two different

pathways from the chloroplast stroma into the thylakoid lumen, the Sec-dependent pathway and the delta-pH or twin-arginine translocation (TAT) pathway.[61] The signals for the two pathways are very similar, the only significant difference being that the TAT pathway proteins contain a twin-arginine (RR) motif in the LTP (KR and RK may also be accepted[18]). The -3,-1 motif found at the SP cleavage site in secreted proteins is present also in LTPs, here even more strongly conserved.[19,62]

Peroxisomal targeting signals (PTS) are either C- or N-terminal

Proteins are transported into the nucleus in a folded state

Nucleus

Proteins are transported into the nucleus in a folded state. Various nuclear import pathways have been detected which all involve carrier proteins (eg importins) that form a complex with the nuclear-to-be protein, a complex that is subsequently translocated through the nuclear pore, where it is dissociated and the importin is shuttled back to the cytoplasm and reused.

The nuclear localisation signals (NLSs), which facilitate the nuclear import of a protein, can be present anywhere in the protein sequence. NLSs do not generally show any particular consensus sequence and it is thus rather hard to discriminate an NLS from a non-NLS region. The two classical types of NLS are the mono-partite, which consists of four basic and one helix-breaking residues, and the bi-partite, which consists of two clusters of basic residues with a spacer of 9–12 amino acids in between.[63,64] Unfortunately, these patterns are not at all unique to nuclear proteins but may well be observed in many other proteins.[65]

Many other signals mediating nuclear import have been found, the 38 amino acid long M9 sequence[66] and the repeated G-R motif[67] to just mention two.[68] However, these signals are in general significantly less frequent than the mono- and bipartite NLS. There are also signals for nuclear protein export and retention, but these will not be considered in this review.

Peroxisome

Proteins destined for the peroxisome contain either of two peroxisomal targeting signals (PTSs): one in the N-terminal region (PTS2), and another one that comprises the three most C-terminal residues (PTS1), and which by far is the predominant signal of the two (Figure 1). The PTS1 consensus sequence is -Ser-Lys-Leu,[69] but in a recent survey of peroxisomal proteins in SWISS-PROT, 35 different variations of the C-terminal peroxisome-targeting tripeptide motif were found (Emanuelsson, Elofsson, von Heijne, Cristobal, manuscript in preparation).

PTS1-containing proteins are recognised by the soluble Pex5 receptor. The Pex5-PTS1-protein complex is then docked to the translocation machinery on the peroxisomal surface.[70] A two-hybrid system was used to show that also part of the adjacent upstream region may play a role in the binding of PTS1-containing proteins to the Pex5 receptor,[71] and this was recently confirmed by a crystallographic study of Pex5 with a bound PTS1.[6]

As hinted, PTS2-containing proteins are scarce. PTS2 is a bipartite signal with consensus sequence [R/K]-[L/V/I]-x-x-x-x-x-[H/Q]-[L/A], usually but not necessarily located in the N-terminal part.[72,73] In mammals and plants the PTS2 is located within a cleavable part of the sequence, much like, for example, the secretory case, except that import and cleavage do not occur in a coordinated manner.

METHODS FOR PREDICTING PROTEIN LOCALISATION

Subcellular localisation predictors can be classified according to (i) the input data they demand and (ii) how the prediction rules are constructed. The input data may be the amino acid composition of the entire protein, some features derived from the sequence, eg hydrophobicity in certain regions, the presence of certain motifs or the actual amino acid sequence itself (usually only part of the entire sequences), or, of course, a mixture. Expression patterns have also been used for predicting the subcellular localisation of sets of proteins, together with sequence motifs, based on the observation that expression data and subcellular localisation are correlated.[2] A slightly different idea is to use the phylogenetic profile of a protein (ie a list obtained by checking in several fully sequenced genomes the presence or absence of homologues to the query protein), with the assumption that similar phylogenetic profiles imply co-localisation, since the endosymbiotic origin of the organelles will be reflected in the profiles.[74] The two lastly mentioned principles have not been taken all the way to independent automated predictors where users can submit their own data, and will thus not be further covered in this review.

Localisation rules can be manually or automatically derived

Methods for the construction of prediction rules span from completely manual collections of rules from the literature to entirely automatic pattern recognition techniques such as neural networks. However, even when using automated feature extraction methods, it is wise to incorporate as much biological knowledge as possible in the model. For instance, when the input data comprise an amino acid sequence one obvious operation is to restrict the automated feature extraction to sequence regions where the signal is known to be located (eg the N-terminal when looking for signal peptides), or when a signal is known to have a regional structure, prediction will benefit from modelling the various regions differently.[75]

Binary predictors often produce many false positives

Many predictors are constructed to make a binary decision whether the protein belongs to one particular compartment of interest, while multi-category predictors aim at sorting the query protein to one out of several possible locations. A common problem with binary predictors is that they tend to produce many false positives by overestimating the number of proteins predicted to the compartment of their special interest, or in other words their sensitivity is good while their specificity is poor.[48]

A related issue is the differences in prediction accuracy for various compartments, which may be due to scarce data or the complexity (or lack of complexity) of the signal. Generally, the more data available for a particular signal, the better prediction rules may be constructed (manually or automatically). There are two aspects of the signal complexity. A very complex signal means that a lot of data will be needed to ensure (at some degree of confidence) that we have covered all variants of the signal in the data set used for predictor construction. On the other hand, a less complex signal means that there will probably be many proteins that are not located to the particular compartment but still contain a signal-like motif by chance. For instance, for peroxisomal proteins with the C-terminal PTS1 signal, there are in SWISS-PROT (release 40)[76] approximately twice as many non-peroxisomal proteins containing a valid PTS1 signal at their C-terminus as there are truly peroxisome-located proteins with PTS1-signal (Emanuelsson, Elofsson, von Heijne, Cristobal, manuscript in preparation). Finally, it should also be noted that some predictors are specialised on a certain group (or groups) of organisms, eg plant proteins only.

In the following, some of the most widely used subcellular localisation predictors will be presented (see Table 2 for web references to the predictors). The benchmark presented in Table 3 was done using the TargetP training and test set.[77] Accordingly, only cross-validated test set performance figures (ie measured on sequences that were not used in the training) are presented for TargetP, in order not to favour this particular predictor over the others.

PSORT

PSORT[78,79] is a knowledge-based, multi-category subcellular localisation program. It represents an impressive work and is the

Table 2: Web addresses of subcellular localisation predictors

Predictor	Web address
PSORT	http://psort.nibb.ac.jp/
iPSORT	http://hypothesiscreator.net/iPSORT/
TargetP	http://www.cbs.dtu.dk/services/TargetP/
SignalP	http://www.cbs.dtu.dk/services/SignalP/
ChloroP	http://www.cbs.dtu.dk/services/ChloroP/
Predotar	http://www.inra.fr/predotar/
NNPSL	http://www.doe-mbi.ucla.edu/~astrid/astrid.html
SubLoc	http://www.bioinfo.tsinghua.edu.cn/SubLoc/
MitoProt	http://www.mips.biochem.mpg.de/cgi-bin/proj/medgen/mitofilter/
predictNLS	http://cubic.bioc.columbia.edu/predictNLS/
TMHMM	http://www.cbs.dtu.dk/services/TMHMM/

PSORT deals with 17 different subcellular locations

gold standard in the field of localisation prediction, both because of its early appearance and its hitherto unchallenged comprehensiveness. Predictions are made by calculating a set of sequence-derived parameters and comparing them with a representation of a number of localisation rules that have been collected from the literature. Many of these rules concern the presence of various sequence motifs that enable proteins to be localised to a certain compartment. PSORT is actually two different programs. PSORT I deals with 17 subcellular compartment (discriminating, for example, between different types of membranes), and was trained on a set of 295 proteins from various species. PSORT II comprises 10 localisations, and the web version was trained on 1,080 yeast proteins. Thus, PSORT II does not deal with plant sequences. Besides this, the most apparent difference is between the reasoning algorithms of the two versions. While PSORT I sequentially applies a set of if-then rules in a tree-like manner (Figure 3) and calculates a certainty factor for each localisation, PSORT II stores all the sequence-derived features of the entire training set, and predicts a new sequence by choosing the majority localisation among a certain number of nearest training examples (*k* nearest-neighbours classification). Apart from resulting in better performance, this approach also facilitates the incorporation of new training data. PSORT manages to

Table 3: Performance of multi-category predictors on a set of 940 plant proteins or 2,738 non-plant proteins.[77] All predictors were tested using the version that appeared at their respective web site (Table 2), using default parameter settings. Note that performances usually differ when measuring on different test sets. Consequently, the figures in this table should be interpreted with care and not automatically be taken as proof that one predictor is better than another

(a) Plant data set

Predictor	% correct	Mitochondrion		Secreted		Chloroplast		'Other'	
		sens	spec	sens	spec	sens	spec	sens	spec
iPSORT	83.4	0.84	0.86	0.91	0.98	0.68	0.71	0.83	0.70
PSORT	69.8	0.66	0.87	0.82	0.74	0.47	0.69	0.78	0.47
TargetP	85.3	0.82	0.90	0.91	0.95	0.85	0.69	0.85	0.78
Predotar	84.8	0.86	0.87	–	–	0.82	0.77	0.85*	0.86*

(b) Non-plant data set

Predictor	% correct	Mitochondrion		Secreted		Nuclear		Cytosolic	
		sens	spec	sens	spec	sens	spec	sens	spec
iPSORT	88.5	0.74	0.68	0.92	0.92		'other': 0.90 0.92		
PSORT	83.2	0.81	0.60	0.64	0.93	0.84	0.75	0.44	0.46
TargetP	90.0	0.89	0.67	0.96	0.92		'other': 0.88 0.97		
NNPSL	73.1	0.74	0.46	0.62	0.77	0.72	0.80	0.45	0.42
SubLoc	77.4	0.67	0.61	0.50	0.74	0.84	0.79	0.64	0.46

sens (sensitivity): the fraction of proteins known to belong to the compartment that actually are predicted to be there;
spec (specificity): the fraction of proteins predicted to the compartment that are known to belong there.
*Secreted and 'other' sets pooled.

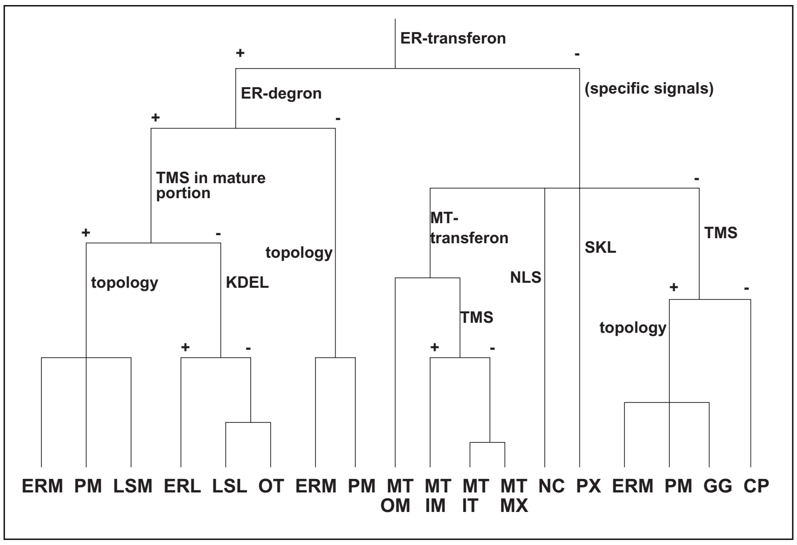

Figure 3: Reasoning tree of PSORT I (simplified). Adapted from the original PSORT paper.[79] At each node, a decision is made based on the result of the corresponding calculation. (+), yes; (−), no; *ER*, endoplasmic reticulum; *TMS*, transmembrane segment; *KDEL*, ER retention signal; *NLS*, nuclear localisation signal; *SKL*, peroxisomal location signal; *PM*, integral plasma membrane; *LSM*, lysosome membrane; *ERL*, endoplasmic reticulum lumen; *LSL*, lysosome lumen; *OT*, extracellular; *MT*, mitochondrion (*OM*, outer membrane; *IM*, inner membrane; *IT*, intermembrane space; *MX*, matrix); *NC*, nuclear; *PX*, peroxisomal; *GG*, Golgi complex; *CP*, cytoplasmic

TargetP uses amino acid sequence as input

correctly predict from slightly below 70 per cent of plant proteins and up to above 80 per cent for non-plant proteins, for several different test sets, and taking into account various numbers of possible locations[48,77] (Table 3).

iPSORT[80] is a development of PSORT restricted to deal with secreted, mitochondrial, chloroplast and other localisations. Compared with the original PSORT, the rules are updated and the prediction algorithm is constructed in a greedy way, according to the degree of discriminative performance. Thus, in the first step rules that concern whether a protein is secreted are applied. If they are not fulfilled, rules for organelle (mitochondrial/chloroplast) localisation are tried, and if these also fail, the protein is assigned to the other group. iPSORT was trained on the TargetP data sets (see below), and is available in plant and non-plant versions. See Table 3 for performance comparisons.

TargetP/ChloroP/SignalP

TargetP[77] is conceptually based on the earlier binary predictors SignalP[81] and ChloroP[56] which deal with SPs and cTPs, respectively. These predictors both employ neural networks and are able to predict not only whether there is an N-terminal sorting signal, but also, with some degree of success, where it is cleaved. SignalP is also available in a hidden Markov model version,[75] which in addition to the SP/non-SP prediction is able to discriminate uncleaved signal anchors from cleaved signal peptides.

When developing SignalP and ChloroP it soon became evident that the mix-up between different locations due to overprediction was significant. Specifically, many mitochondrial proteins were wrongly predicted as secretory or chloroplast. The four-category predictor TargetP was then constructed with the intention to reduce this problem.

TargetP comes in two versions: one for plant proteins (trained to recognise cTP, SP and mTP) and one for other eukaryotic proteins (recognising SP and mTP). Both versions are otherwise virtually identical, with two layers of neural networks, where the first layer consists of two (three for the plant version) parallel networks, each of which assigns a value of signal peptide-ness or mTP-ness (or cTP-ness for the plant version), respectively, to each residue in the query protein. The top layer network then integrates these values for the 100 N-terminal residues in the sequence and outputs one score per potential subcellular location, ie three scores for the non-plant version (mTP, SP, other) and four scores for the plant version (cTP, mTP, SP, other); Figure 4. TargetP was trained (using five-fold cross-validation) on a set of 940 plant and 2,738 non-plant SWISS-PROT sequences[76] that were redundancy-reduced by aligning all against all (within the categories), and removing too similar sequences using the Hobohm algorithm.[82] On a SWISS-PROT test set with by and large equal amounts of chloroplast, mitochondrial,

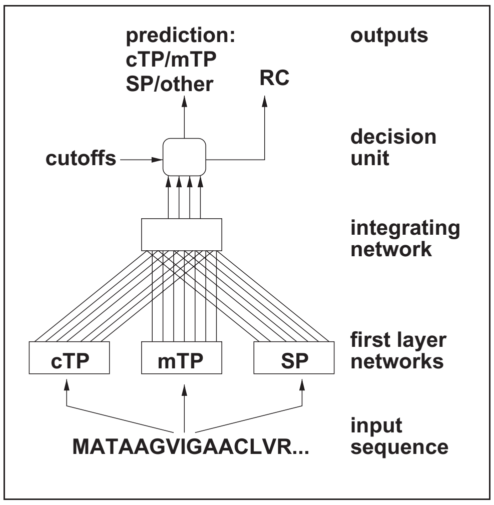

Figure 4: TargetP architecture (plant version).[78] TargetP was constructed with two layers of neural networks (first layer and integrating networks) and a decision unit that outputs a prediction and a reliability class (RC), which is a measure of the certainty of each prediction

secreted and other proteins, TargetP was able to predict 90 per cent of the non-plant sequences and 85 per cent of the plant sequences correctly;[77] Table 3.

Predotar

Predotar is able to predict dual location

Predotar[83] is primarily aimed at distinguishing between chloroplast and mitochondrial proteins, and is thus exclusively constructed on, and aimed at, plant sequences. It can also predict dual location, ie proteins that are found in both chloroplasts and mitochondria, which has been experimentally detected and reported for several proteins.[58,84] Predotar was constructed using neural networks, but with a slightly different architecture than TargetP. Even though the construction of data sets differs, the performance of Predotar is very similar to that of TargetP with the exception that Predotar does not deal with predicting secreted proteins (Table 3).

NNPSL

NNPSL and SubLoc use amino acid composition as input

NNPSL[85] is yet another neural network-based predictor. Unlike TargetP and Predotar it uses the overall amino acid contents of the protein to assign one of four subcellular localisations (cytosolic, extracellular, mitochondrial and nuclear) to a query sequence. The advantage of using the amino acid frequencies instead of particular motifs is that the prediction is less sensitive to errors and omissions in the query protein sequence. The idea of using amino acid composition for localisation prediction was pioneered by Nakashima and Nishikawa, who used it in combination with residue pair frequencies to discriminate between extra- and intracellular proteins.[86]

In the training of NNPSL, some 2,400 eukaryotic SWISS-PROT sequences were included, redundancy-reduced so that the mutual identity was at most 90 per cent for any pair of proteins in the training set. In the original paper, 66.1 per cent of test set proteins were reported to be correctly located in a multi-category prediction. See also Table 3 for performance comparisons.

SubLoc

SubLoc[87] predicts subcellular localisation using support vector machines (SVM), a machine-learning technique that finds the optimal separating hyperplane between classes of data. Like the above-mentioned NNPSL, SubLoc uses the amino acid composition as input data, and was in fact trained on the same data sets. Performance mounted to 79.4 per cent using the SVM instead of the NN technique. The large difference in performance between the support vector approach (SubLoc) and the neural network approach (NNPSL) is remarkable, considering that the data used for training are identical. However, the benchmark study in this review using another test set indicates a smaller difference in performance (Table 3).

MitoProt

MitoProt II[88,89] predicts mitochondrial localisation of a sequence by calculating several physicochemical parameters from the sequence and then computing a linear discriminant function (LDF). The parameters considered are obtained from the literature, and the ones with greatest influence on the prediction turned out to be the amino acid content in certain regions (quite unsurprisingly, the N-terminus is one of the considered parts), and the hydrophobicity of regions. When tested on a set of human proteins from SWISS-PROT (3,419 sequences), MitoProt II was able to predict 86.7 per cent correctly in a binary prediction (mitochondrial/non-mitochondrial).[48] MitoProt II can also be used to predict chloroplast location, but this property has not been documented in any paper.

predictNLS

predictNLS[65] represents an interesting approach to motif finding. It predicts whether a protein is nuclear (ie if it contains a nuclear localisation signal, NLS) or not by scanning the query protein for the presence of any of a set of more than two hundred known or constructed NLS motifs. An initial set of 91 experimentally verified motifs (NLSs) were found by searching the literature. The number of motifs associated with nuclear localisation was then increased by '*in silico* mutatogenesis' of the known NLSs in such a way that only motifs that exclusively matched known nuclear proteins but no other proteins were accepted. In this way, the specificity (when applied to known sequences) was 1.00 while sensitivity was reported to be 0.43 when testing on the eukaryotic proteins in SWISS-PROT.[76]

TMHMM

TMHMM predicts transmembrane proteins and their topology

TMHMM[90] is a general predictor of α-helical transmembrane regions in proteins. It was constructed using HMMs, which are particularly suitable for modelling the different (and sequentially appearing) regions that a transmembrane protein consists of: cytoplasmic and extracellular loops, hydrophobic transmembrane helices and border regions (helical caps). The HMM approach also permits TMHMM to predict not only the actual transmembrane regions but also which non-membrane loops/regions that are on the inside (cytosolic side) of the membrane and which are on the outside, ie the topology. On a protein level, approximately 80 per cent of their cross-validated data set were reported to be correctly predicted, and an independent study has coined TMHMM as the currently best-performing transmembrane prediction program.[91]

Very recently, a similar HMM-based approach aimed at recognising β-barrel regions in proteins was presented.[92] The most interesting feature of this predictor is perhaps the inclusion of evolutionary information in the training set, by using sequence profiles derived from PSI-BLAST searches.[93,94] On a residue level, 83 per cent of the test set was reported to be correctly assigned.

DISCUSSION

Predictions – of many various kinds – have indeed speeded up the pace of molecular biology and have become indispensable tools for experimentalists. It appears as if localisation prediction is possible to carry out with decent performance, and there is a range of programs available, all with their own strengths and weaknesses. A benchmark for some of the more widely used predictors is presented in Table 3, but the figures should be considered as rough estimates rather than absolute performance standards. It is hard to assess how well the predictors are doing when encountering large amounts of unknown data (such as complete genomes) since all of them – for very natural reasons – in one way or the other are based upon the current knowledge about sorting mechanisms and upon the databases that have been collected with this knowledge, and generally this knowledge *only*, in mind. However, extrapolating (conservatively) from the figures presented in Table 3 and counting on some general prediction performance improvement when incorporating even more data from the ever-growing databases (or database in singular, rather, since SWISS-PROT is the prime choice for virtually all players on the subcellular localisation prediction field) and the literature, it does not seem unlikely that some 90 per cent of all proteins in an organism could soon be correctly sorted *in silico* among the compartments mentioned in this review. Careful manual examination of prediction results could raise this figure additionally. The issue is, however, corrupted by phenomena such as dual location and erroneous annotations, which effectively prevent the construction of an even near-perfect predictor.

Many molecular mechanisms behind protein sorting are still unknown

Another aspect, already mentioned in the Introduction, is that some compartments are harder to predict than others, which partly is a reflection of the fact that we do not know all the underlying mechanisms in enough detail. For instance, the role of the three-dimensional structures of many types of signals (and their receptors) remains to be investigated. Structures of complexes of signal sequence and receptor have been reported for mTP[5] and PTS1,[6] but are still lacking for other signals. For chloroplast transit peptides, it has been speculated that cTPs have evolved to provide maximum random coil potential,[95] while NMR studies of both algal and plant cTPs in the presence of a membrane-mimetic solvent system, suggest a mixture of α-helix and random coil.[96–99]

The figures in Table 3 also reveal that the methods looking for the actual localisation motifs are advantageous to those relying solely on amino acid composition. While it is still obvious that overall residue composition is correlated with subcellular localisation, this correlation may be a secondary effect stemming from the fact that various compartments perform different tasks and thus are inhabited by different protein families with different compositional biases.

Use localisation predictors in coordination with other bioinformatic tools to gain maximum insight

The use of dedicated localisation predictors should of course always be coordinated with more general bioinformatic tools such as alignments and phylogenetic analyses to ensure maximum insight. Another observation is that, if possible, it is wise to try more than one localisation predictor since that tends to increase the reliability of the prediction (if they agree, that is). In addition if you're not satisfied with the verdict of the first doctor, why not get a second opinion? In homology modelling of 3D-structures, consensus predictors have been in use for a while[99] but within the field of subcellular localisation, the idea of consensus predictions has not as of yet been very extensively tried. When applied to whole-genome data, a related approach is to serially apply predictors to narrow down the set of predicted proteins. Peltier *et al.*[100] used a combined approach of several serially connected localisation predictors (TargetP, SignalP, TMHMM) and cutoff parameters derived from sets of experimentally verified proteins to predict the entire protein complement ('proteome') of the *A. thaliana* thylakoid. The study highlights another important consideration in the localisation prediction business, namely that the goal may not necessarily be perfect predictions from a single predictor but instead a set of comparably reliable predictions potentially giving new insights into the functions carried out in a particular compartment and providing a starting point for further experimental studies.

Acknowledgments

I would like to thank Gunnar von Heijne for critically reading the manuscript, and Kenta Nakai for the permission to use Figure 3. The author is supported by a grant from the Swedish Foundation for Strategic Research (SFF).

References

1. Bork, P. and Eisenhaber, F. (1998), 'Wanted: subcellular localization of proteins based on sequence', *Trends Cell Biol.*, Vol. 8, pp. 169–170.
2. Drawid, A. and Gerstein, M. (2000), 'A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome', *J. Mol. Biol.*, Vol. 301, pp. 1059–1075.
3. Schatz, G. and Dobberstein, B. (1996),

'Common principles of protein translocation across membranes', *Science*, Vol. 271, pp. 1519–1526.

4. Izaurralde, E., Kann, M., Panté, N. *et al.* (1999), 'Viruses, microorganisms and scientists meet the nuclear pore', *EMBO J.*, Vol. 18, pp. 289–296.

5. Abe, Y., Shodai, T., Muto, T. *et al.* (2000), 'Structural basis of presequence recognition by the mitochondrial protein import receptor Tom20', *Cell*, Vol. 100, pp. 551–560.

6. Gatto Jr, G. J., Geisbrecht, B. V., Gould, S. J. and Berg, J. M. (2000), 'Peroxisomal targeting signal–1 recognition by the TPR domains of human PEX5', *Nat. Struct. Biol.*, Vol. 7, pp. 1091–1095.

7. Nakai, K. (2000), 'Protein sorting signals and prediction of subcellular localization', *Adv. Protein Chem.*, Vol. 54, pp. 277–344.

8. Koonin, E. V. (2000), 'How many genes can make a cell: the minimal-gene-set concept', *Annu. Rev. Genomics Hum. Genet.*, Vol. 1, pp. 99–116.

9. Rapoport, T. A. (1992), 'Transport of proteins across the endoplasmic reticulum membrane', *Science*, Vol. 258, pp. 931–936.

10. von Heijne, G. (1990), 'The signal peptide', *J. Membr. Biol.*, Vol. 115, pp. 195–201.

11. von Heijne, G. (1983), 'Patterns of amino acids near signal sequence cleavage sites', *Eur. J. Biochem.*, Vol. 133, pp. 17–21.

12. von Heijne, G. (1985), 'Signal sequences. The limits of variation', *J. Mol. Biol.*, Vol. 184, pp. 99–105.

13. von Heijne, G. and Abrahmsén, L. (1989), 'Species-specific variation in signal peptide design. Implications for protein secretion in foreign hosts', *FEBS Lett.*, Vol. 244, pp. 439–446.

14. Schneider, T. D. and Stephens, R. M. (1990), 'Sequence logos: a new way to display consensus sequences', *Nucleic Acids Res.*, Vol. 18, pp. 6097–6100.

15. de Gier, J. W., Scotti, P. A., Sääf, A. *et al.* (1998), 'Differential use of the signal recognition particle translocase targeting pathway for inner membrane protein assembly in *Escherichia coli*', *Proc. Natl Acad. Sci.*, Vol. 95, pp. 14646–14651.

16. Keenan, R. J., Freymann, D. M., Stroud, R. M. and Walter, P. (2001), 'The signal recognition particle', *Annu. Rev. Biochem.*, Vol. 70, pp. 755–775.

17. Driessen, A. J., Fekkes, P. and van der Wolk, J. P. (1998), 'The Sec system', *Curr. Opin. Microbiol.*, Vol. 1, pp. 216–222.

18. Ize, B., Gerard, F., Zhang, M., Chanal, A. *et al.* (2002), '*In vivo* dissection of the Tat translocation pathway in *Escherichia coli*', *J. Mol. Biol.*, Vol. 317, pp. 327–335.

19. Robinson, C. and Bolhuis, A. (2001), 'Protein targeting by the twin-arginine translocation pathway', *Nat. Rev. Mol. Cell. Biol.*, Vol. 2, pp. 350–356.

20. Sakaguchi, M. (1997), 'Eukaryotic protein secretion', *Curr. Opin. Biotechnol.*, Vol. 8, pp. 595–601.

21. von Heijne, G. (1998), 'Transcending the impenetrable: how proteins come to terms with membranes', *Biochim. Biophys. Acta*, Vol. 947, pp. 307–333.

22. Thompson, G. A. and Okuyama, H. (2000), 'Lipid-linked proteins of plants', *Progr. Lipid Res.*, Vol. 39, pp. 19–39.

23. Ikezawa, H. (2002), 'Glycosylphosphatidylinositol (GPI)-anchored proteins', *Biol Pharm. Bull.*, Vol. 25, pp. 409–417.

24. Delcour, A. H. (2002), 'Structure and function of pore-forming beta-barrels from bacteria', *J. Mol. Microbiol. Biotechnol.*, Vol. 4, pp. 1–10.

25. Claros, M. G. and von Heijne, G. (1994), 'TopPred II: an improved software for membrane protein structure predictions', *Comput. Appl. Biosci.*, Vol. 10, pp. 685–686.

26. Bowie, J. U. (1997), 'Helix packing in membrane proteins', *J. Mol. Biol.*, Vol. 272, pp. 780–789.

27. von Heijne, G. and Gavel, Y. (1988), 'Topogenic signals in integral membrane proteins', *Eur. J. Biochem.*, Vol. 174, pp. 671–678.

28. Ota, K., Sakaguchi, M., von Heijne, G. *et al.* (1998), 'Forced transmembrane orientation of hydrophylic polypeptide segments in multispanning membrane proteins', *Mol. Cell*, Vol. 2, pp. 495–503.

29. Hegde, R. S. and Lingappa, V. R. (1997), 'Membrane protein biogenesis: regulated complexity at the endoplasmic reticulum', *Cell*, Vol. 91, pp. 575–582.

30. Matlack, K. E., Mothes, W. and Rapoport, T. A. (1998), 'Protein translocation: tunnel vision', *Cell*, Vol. 92, pp. 381–390.

31. Schulz, G. E. (2000), 'Beta-barrel membrane proteins', *Curr. Opin. Struct. Biol.*, Vol. 10, pp. 443–447.

32. Bucher, P. and Bairoch, A. (1994), 'A generalized profile syntax for biomolecular sequences motifs and its function in automatic sequence interpretation', in 'Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology', AAAI Press, Menlo Park, CA, pp. 53–61.

33. Denny, P. W., Gokool, S., Russell, D. G. *et al.* (2000), 'Acylation-dependent protein

export in *Leishmania*', *J. Biol. Chem.*, Vol. 275, pp. 11017–11025.

34. Martin, M. L. and Busconi, L. (2000), 'Membrane localization of a rice calcium-dependent protein kinase (cdpk) is mediated by myristoylation and palmitoylation', *Plant J.*, Vol. 24, pp. 429–435.

35. Andersson, S. G., Zomorodipour, A., Andersson, J. O. *et al.* (1998), 'The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria', *Nature*, Vol. 396, pp. 133–140.

36. Dacks, J. B. and Doolittle, W. F. (2001), 'Reconstructing/deconstructing the earliest eukaryotes: how comparative genomics can help', *Cell*, Vol. 16, pp. 419–425.

37. Shleiff, E., Heard, T. S. and Weiner, H. (1999), 'Positively charged residues, the helical conformation and the structural flexibility of the leader sequence of pALDH are important for recognition by hTom20', *FEBS Lett.*, Vol. 461, pp. 9–12.

38. Neupert, W. (1997), 'Protein import into mitochondria', *Annu. Rev. Biochem.*, Vol. 66, pp. 863–917.

39. Pfanner, N. and Geissler, A. (2001), 'Versatility of the mitochondrial protein import machinery', *Nat. Rev. Mol. Cell. Biol.*, Vol. 2, pp. 339–349.

40. Arretz, M., Schneider, H., Wienhues, U. and Neupert, W. (1991), 'Processing of mitochondrial precursor proteins', *Biomed. Biochim. Acta*, Vol. 50, pp. 403–412.

41. Brunner, M., Klaus, C. and Neupert, W. (1994), 'The mitochondrial processing peptidase', in von Heijne, G., Ed., 'Signal Peptidases', R. G. Landes Company, Austin, pp. 73–86.

42. Kalousek, F., Hendrick, J. P. and Rosenberg, L. E. (1988), 'Two mitochondrial matrix proteases act sequentially in the processing of mammalian matrix enzymes', *Proc. Natl Acad. Sci. USA*, Vol. 85, pp. 7536–7540.

43. Isaya, G. and Kalousek, F. (1994), 'Mitochondrial intermediate peptidase', in von Heijne, G., Ed., 'Signal Peptidases', R. G. Landes Company, Austin, pp. 87–103.

44. Glick, B. S., Brandt, A., Cunningham, K. *et al.* (1992), 'Cytochromes c1 and b2 are sorted to the intermembrane space of yeast mitochondria by a stop-transfer mechanism', *Cell*, Vol. 69, pp. 809–822.

45. Gärtner, F., Bömer, U., Guiard, B. and Pfanner, N. (1995), 'The sorting signal of cytochrome b2 promotes early divergence from the general mitochondrial import pathway and restricts the unfoldase activity of matrix Hsp70', *EMBO J.*, Vol. 14, pp. 6043–6057.

46. Brix, J., Bukau, S. R., Schneider-Mergener, J. and Pfanner, N. (1999), 'Distribution of binding sequences for the mitochondrial import receptors Tom20, Tom22, and Tom70 in a presequence-carrying preprotein and a non-cleavable preprotein', *J. Biol. Chem.*, Vol. 274, pp. 16522–16530.

47. von Heijne, G., Steppuhn, J. and Hermann, S. G. (1989), 'Domain structure of mitochondrial and chloroplast targeting peptides', *Eur. J. Biochem.*, Vol. 180, pp. 535–545.

48. Emanuelsson, O., von Heijne, G. and Schneider, G. (2001), 'Analysis and prediction of mitochondrial targeting peptides', *Methods Cell Biol.*, Vol. 65, pp. 175–187.

49. Schneider, G., Sjöling, S., Wallin, E. *et al.* (1998), 'Feature-extraction from endopeptidase cleavage sites in mitochondrial targeting peptides', *Proteins*, Vol. 30, pp. 49–60.

50. Taylor, A. B., Smith, B. S., Kitada, S. *et al.* (2001), 'Crystal structures of mitochondrial processing peptidase reveal the mode for specific cleavage of import signal sequences', *Structure*, Vol. 9, pp. 615–625.

51. Moreira, D., Guyader, H. L. and Philippe, H. (2000), 'The origin of red algae and the evolution of chloroplasts', *Nature*, Vol. 405, pp. 69–72.

52. Soll, J. and Tien, R. (1998), 'Protein translocation into and across the chloroplastic envelope membranes', *Plant Mol. Biol.*, Vol. 38, pp. 191–207.

53. Jarvis, P. and Soll, J. (2001), 'Toc, Tic, and chloroplast protein import', *Biochim. Biophys. Acta*, Vol. 1541, 64–79.

54. Robinson, C. and Ellis, R. J. (1984), 'Transport of proteins into chloroplasts, partial purification of a chloroplast protease involved in the processing of important precursor polypeptides', *Eur. J. Biochem.*, Vol. 142, pp. 337–342.

55. Bruce, B. D. (2001), 'The paradox of plastid transit peptides: conservation of function despite divergence in primary structure', *Biochim. Biophys. Acta*, Vol. 1541, pp. 2–21.

56. Emanuelsson, O., Nielsen, H. and von Heijne, G. (1999), 'ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites', *Prot. Sci.*, Vol. 8(5), pp. 978–984.

57. Chabregas, S. M., Luche, D. D., Farias, L. P. *et al.* (2001), 'Dual targeting properties of the N-terminal signal sequence of *Arabidopsis thaliana* THI1 protein to mitochondria and chloroplasts', *Plant Mol. Biol.*, Vol. 46, pp. 639–650.

58. Small, I., Wintz, H., Akashi, K. and Mireau, H. (1998), 'Two birds with one stone: genes that encode products targeted to two or more

compartments', *Plant Mol. Biol.*, Vol. 38, pp. 265–277.

59. Robinson, C., Hynds, P. J., Robinson, D. and Mant, A. (1998), 'Multiple pathways for the targeting of thylakoid proteins in chloroplasts', *Plant Mol. Biol.*, Vol. 38, pp. 209-221.

60. Bonen, L. and Doolittle, W. F. (1975), 'On the prokaryotic nature of red algal chloroplasts', *Proc. Natl Acad. Sci.*, Vol. 72, pp. 2310–2314.

61. Mori, G. and Cline, K. (2001), 'Post-translational protein translocation into thylakoids by the Sec and DeltapH-dependent pathways', *Biochim. Biophys. Acta*, Vol. 1541, pp. 80–90.

62. Shackleton, J. B. and Robinson, C. (1991), 'Transport of proteins into chloroplasts. The thylakoidal processing peptidase is a signal-type peptidase with stringent substrate requirements at the -3 and -1 positions', *J. Biol. Chem.*, Vol. 266, pp. 12152–12156.

63. Kalderon, D., Roberts, B. L., Richardson, W. D. and Smith, A. E. (1984), 'A short amino acid sequence able to specify nuclear location', *Cell*, Vol. 39, pp. 499–509.

64. Robbins, J., Dilworth, S. M., Laskey, R. A. and Dingwall, C. (1991), 'Two interdependent basic domains in nucleoplasmin nuclear targeting sequence: identification of a class of bipartite nuclear targeting sequence', *Cell*, Vol. 64, pp. 615–623.

65. Cokol, M., Nair, R. and Rost, B. (2000), 'Finding nuclear localization signals', *EMBO Rep.*, Vol. 1, pp. 411–415.

66. Pollard, V. W., Michael, W. M., Nakielny, S. *et al.* (2000), 'A novel receptor-mediated nuclear protein import pathway', *Cell*, Vol. 86, pp. 985–994.

67. Dono, R., James, D. and Zeller, R. (1998), 'A GR-motif functions in nuclear accumulation of the large FGF-2 isoforms and interferes with mitogenic signalling', *Oncogene*, Vol. 16, pp. 2151–2158.

68. Christophe, D., Christophe-Hobertus, C. and Pichon, B. (2000), 'Nuclear targeting of proteins: how many different signals?', *Cell Signal*, Vol. 12, pp. 337–341.

69. Gould, S. J., Keller, G. A., Hosken, N. *et al.* (1989), 'A conserved tripeptide sorts proteins to peroxisomes', *J. Cell Biol.*, Vol. 108, pp. 1657–1664.

70. Hettema, E. H., Distel, B. and Tabak, H. F. (1999), 'Import of proteins into peroxisomes', *Biochim. Biophys. Acta*, Vol. 1451, pp. 17–34.

71. Lametschwandtner, G., Brocard, C., Fransen, M. *et al.* (1998), 'The difference in recognition of terminal tripeptides as peroxisomal targeting signal 1 between yeast and human is due to different affinities of their receptor Pex5p to the cognate signal and to residues adjacent to it', *J. Biol. Chem.*, Vol. 273, pp. 33635–33643.

72. Swinkels, B. W., Gould, S. J., Bodnar, A. G. *et al.* (1991), 'A novel, cleavable peroxisomal targeting signal at the amino-terminus of the rat 3-ketoacyl-CoA thiolase', *EMBO J.*, Vol. 10, pp. 3255–3262.

73. Chudzik, D. M., Michels, P. A., de Walque, S. and Hol, W. G. (2000), 'Structures of type 2 peroxisomal targeting signals in two trypanosomatid aldolases', *J. Mol. Biol.*, Vol. 300, pp. 697–707.

74. Marcotte, E. M., Xenarios, I., van Der Bliek, A. M. and Eisenberg, D. (2000), 'Localizing proteins in the cell from their phylogenetic profiles', *Proc. Natl Acad. Sci.*, Vol. 97, pp. 12115–12120.

75. Nielsen, H. and Krogh, A. (1998), 'Prediction of signal peptides and signal anchors by a hidden Markov model', in 'Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology', AAAI Press, Menlo Park, CA, pp. 122–130.

76. Bairoch, A. and Apweiler, R. (2000), 'The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000', *Nucleic Acids Res.*, Vol. 28, pp. 45–48.

77. 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.*, Vol. 300, pp. 1005–1016.

78. Nakai, K. and Kanehisa, M. (1992), 'A knowledge base for predicting protein localization sites in eukaryotic cells', *Genomics*, Vol. 14, pp. 897–911.

79. Horton, P. and Nakai, K. (1997), 'Better prediction of protein cellular localization sites with the *k* nearest neighbors classifier', in 'Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology', AAAI Press, Menlo Park, CA, pp. 147–152.

80. Bannai, H., Tamada, Y., Maruyama, O. *et al.* (2002), 'Extensive feature detection of n-terminal protein sorting signals', *Bioinformatics*, Vol. 18, pp. 298–305.

81. Nielsen, H., Engelbrecht, J., Brunak, S. and von Heijne, G. (1997), 'Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites', *Prot. Eng.*, Vol. 10, pp. 1–6.

82. Hobohm. U., Scharf, M., Schneider, R. and Sander, C. (1992), 'Selection of representative protein data sets', *Prot. Sci.*, Vol. 1, pp. 409–417.

83. URL: http://www.inra.fr/predotar/

84. Zhang, X. P. and Glaser, E. (2002), 'Interaction of plant mitochondrial and chloroplast signal peptides with the hsp70 molecular chaperone', *Trends Plant Sci.*, Vol. 7, pp. 14–21.

85. Reinhardt, A. and Hubbard, T. (1998), 'Using neural networks for prediction of the subcellular location of proteins', *Nucleic Acids Res.*, Vol. 26, pp. 2230–2236.

86. Nakashima, H. and Nishikawa, K. (1994), 'Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies', *J. Mol. Biol.*, Vol. 238, pp. 54–61.

87. Hua, S. and Sun, Z. (2001), 'Support vector machine approach for protein subcellular localization prediction', *Bioinformatics*, Vol. 17, pp. 721–728.

88. Claros, M. G. (1995), 'MitoProt, a Macintosh application for studying mitochondrial proteins', *Comput. Appl. Biosci.*, Vol. 11, pp. 441–447.

89. Claros, M. G. and Vincens, P. (1996), 'Computational method to predict mitochondrially imported proteins and their targeting sequences', *Eur. J. Biochem.*, Vol. 241, pp. 779–786.

90. Krogh, A., Larsson, B., von Hejne, G. and Sonnhammer, E. (2001), 'Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes', *J. Mol. Biol.*, Vol. 305, pp. 567–580.

91. Möller, S., Croning, M. and Apweiler, R. (2001), 'Evaluation of methods for the prediction of membrane spanning regions', *Bioinformatics*, Vol. 17, pp. 646–653.

92. Martelli, P. L., Fariselli, P., Krogh, A. and Casadio, R. (2002), 'A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins', *Bioinformatics*, Vol. 18(Suppl. 1), pp. 46–53.

93. Altschul, S. F., Madden, T. L., Schaffer, A. A. *et al.* (1997), 'Gapped BLAST and PSI-BLAST: a new generation of protein database search programs', *Nucleic Acids Res.*, Vol. 25, pp. 3389–3402.

94. Schaffer, A. A., Aravind, L., Madden, T. L. *et al.* (2001), 'Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements', *Nucleic Acids Res.*, Vol. 29, pp. 2994–3005.

95. von Heijne, G. and Nishikawa, K. (1991), 'Chloroplast transit peptides. The perfect random coil?', *FEBS Lett.*, Vol. 278, pp. 1–3.

96. Lancelin, J. M., Bally, I., Arlaud, G. J. *et al.* (1994), 'NMR structures of ferredoxin chloroplastic transit peptide from *Chlamydomonas reinhardtii* promoted by trifluoroethanol in aqueous solution', *FEBS Lett.*, Vol. 343, pp. 261–266.

97. Krimm, I., Gans, P., Hernandez, J. F. *et al.* (1999), 'A coil-helix instead of a helix-coil motif can be induced in a chloroplast transit peptide from *Chlamydomonas reinhardtii*', *Eur. J. Biochem.*, Vol. 265, pp. 171–180.

98. Wienk, H. L. J., Czisch, M. and de Kruijff, B. (1999), 'The structural flexibility of the preferredoxin transit peptide', *FEBS Lett.*, Vol. 453, pp. 318–326.

99. Lundström, J., Rychlewski, L., Bujnicki, J. and Elofsson, A. (2001), 'Pcons: A neural-network-based consensus predictor that improves fold recognition', *Protein Sci.*, Vol. 10, pp. 2354–2362.

100. Peltier, J.-B., Emanuelsson. O., Kalume, D. E. *et al.* (2002), 'Central functions of the lumenal and peripheral thykaloid proteome of *Arabidopsis* determined by experimentation and genome-wide prediction', *Plant Cell*, Vol. 14, pp. 211–236.