Locating proteins in the cell using TargetP, SignalP and related tools

Olof Emanuelsson[1], Søren Brunak[2], Gunnar von Heijne[3] & Henrik Nielsen[2]

[1]Stockholm Bioinformatics Center, Albanova, Stockholm University, SE-10691 Stockholm, Sweden. [2]Center for Biological Sequence Analysis, Technical University of Denmark, Building 208, DK-2800 Lyngby, Denmark. [3]Department of Biochemistry and Biophysics, Center for Biomembrane Research, Stockholm University, SE-10691 Stockholm, Sweden. Correspondence should be addressed to H.N. (hnielsen@cbs.dtu.dk).

Published online 19 April 2007; doi:10.1038/nprot.2007.131

Determining the subcellular localization of a protein is an important first step toward understanding its function. Here, we describe the properties of three well-known N-terminal sequence motifs directing proteins to the secretory pathway, mitochondria and chloroplasts, and sketch a brief history of methods to predict subcellular localization based on these sorting signals and other sequence properties. We then outline how to use a number of internet-accessible tools to arrive at a reliable subcellular localization prediction for eukaryotic and prokaryotic proteins. In particular, we provide detailed step-by-step instructions for the coupled use of the amino-acid sequence-based predictors TargetP, SignalP, ChloroP and TMHMM, which are all hosted at the Center for Biological Sequence Analysis, Technical University of Denmark. In addition, we describe and provide web references to other useful subcellular localization predictors. Finally, we discuss predictive performance measures in general and the performance of TargetP and SignalP in particular.

INTRODUCTION

In 1999, the Nobel prize in Physiology or Medicine was awarded to Günther Blobel for "the discovery that proteins have intrinsic signals that govern their transport and localization in the cell." As the subcellular localization of a protein is an important clue to its function, the characterization and prediction of these intrinsic signals—the "zip codes" of proteins—has become a major task in bioinformatics.

Over the last few years, several methods to determine the subcellular localization of proteins using high-throughput experiments have been developed. One of the first attempts was by Burns *et al.*[1], who randomly inserted the *lacZ* reporter gene into the yeast genome, and were able to determine the subcellular localization of a total of 245 yeast fusion proteins. This approach was further refined by the use of the green fluorescent protein[2] or simian virus V5 epitope to tag cDNAs and localize the resulting fusion proteins through fluorescence screening of the transfected yeast cells[3,4] and later also human cells[5]. Although these methods allow the study of protein localization *in vivo*, the presence of the fusion/tagging protein may interfere with sequence or structural signals necessary to direct the protein of interest to its proper compartment.

The other main line of large-scale subcellular localization determination, enabled in part by the presence of whole-genome DNA sequences, is to homogenize and fractionate (most commonly through centrifugation) the cell and use mass spectrometry[6] to identify the proteins in the various fractions[7,8]. With proper purification and fractionation techniques, the contents of a particular fraction will correspond to a particular organelle, but the approach is sensitive to contaminations. Andersen *et al.*[9,10] presented a method, protein correlation profiling, which allows the simultaneous analysis over several fractions, thus reducing the need for a complete purification of each fraction. In a recent study, this technique was used to map the subcellular localization of 1,404 mouse liver proteins[11]. Using mass-spectrometry-based organellar proteomics and stable isotope labeling, it has also recently become possible to map quantitatively the dynamics of protein trafficking in and out of entire compartments, for example, the human nucleolus[12].

A third approach was described by Agaton *et al.*[13]. They demonstrated that protein-specific antibodies can be designed and used, through immunohistochemistry, to map tissue specificity and subcellular localization of human proteins. Although applied to fixed, and not living cells, and presently focusing on different tissues rather than different subcellular localizations[14], this is a promising approach to detect the whereabouts of unaltered proteins at their naturally occurring expression levels.

As it is inevitable that high-throughput experimental techniques produce some false-positive assignments (and also false negatives), there is a strong degree of complementarity between the use of experimental methods and computational tools that can score the likelihood that a protein belongs to a given compartment. Such scores can be used to improve the quality of high-throughput data, and also their subsequent use as starting points for identification of compartment-specific protein complexes and networks[15].

Signal peptides

The best known protein "zip code" is the secretory signal peptide (SP), which is found in all the three domains of life. It targets a protein for translocation across the plasma membrane in prokaryotes and across the endoplasmic reticulum (ER) membrane in eukaryotes[16]. It is an N-terminal peptide, typically 15–30 amino acids long, which is cleaved off during translocation of the protein across the membrane. There is no simple consensus sequence for SPs, but they typically show three distinct compositional zones: an N-terminal region (n-region) which often contains positively charged residues, a hydrophobic region (h-region) of at least six residues and a C-terminal region (c-region) of polar uncharged residues with some conservation at the −3 and −1 positions relative to the cleavage site (**Table 1**).

It should be noted that far from all proteins with secretory SPs are actually secreted to the outside of the cell. In Gram-negative

bacteria, they by default end up in the periplasmic compartment, and a separate mechanism is needed to secrete them into the growth medium[17]. In eukaryotes, proteins translocated across the ER membrane are by default transported through the Golgi apparatus and exported by secretory vesicles, but some proteins have specific retention signals that hold them back in the ER or the Golgi or divert them to the lysosomes. In general, these retention signals are poorly characterized, one exception being the ER lumen retention signal, which has the consensus sequence KDEL or HDEL[18].

Not all secretory proteins have SPs—alternative secretion pathways exist in both prokaryotes and eukaryotes. In eukaryotes, a few proteins are secreted without an SP through the so-called nonclassical pathway, which does not follow the classical route through the ER and Golgi[19]. Well-known examples include fibroblast growth factors, interleukins and galectins. In Gram-negative bacteria, as many as five secretion systems are known, of which type II and V depend on cleaved SPs, whereas type I, III and IV do not[17,20–22]. Type I secretion exists also in Gram-positive bacteria, as well as a recently discovered signal-peptide-independent secretion system, which is not identical to any of the five Gram-negative types[23].

TABLE 1 | Examples of signal peptides[a].

Human α-1-antichymotrypsin precursor (ACT):
MERMLPLLALGLLAAGFCPAVLC↓HPNSPLDEEN...
Escherichia coli class B acid phosphatase precursor:
MRKITQAISAVCLLFALNSSAVALA↓SSPSPLNPGT...
Clostridium perfringens ε-toxin type B precursor:
MKKNLVKSLAIASAVISIYSIVNIVSPTNVIA↓KEISNTVSNE...

[a]Three examples of secretory signal peptides from a eukaryote (Human), a Gram-negative bacterium (*E. coli*) and a Gram-positive bacterium (*Clostridium perfringens*). The cleavage sites are marked by arrows and the hydrophobic regions are underlined. Note that the Gram-positive signal peptide is considerably longer than the Gram-negative, which is slightly longer than the eukaryotic one—these examples have been selected to represent the average length for each organism group.

Transit peptides

The targeting peptides of chloroplasts and mitochondria are also N-terminal peptides[24], and similar to the SPs for secretion, these presequences are cleaved off upon entry into their final compartment. Their sequence features are, however, less well characterized and the reported sequence motifs are even less conserved than those of the secretory SP.

The chloroplast transit peptide (cTP), which directs nuclear-encoded proteins into the chloroplast stroma, tends to be rich in hydroxylated residues, in particular serines, and is devoid of acidic residues[25]. A motif, VRA|AAV, has been detected around the cleavage site (|)[26], but the signal is relatively weak. The most conserved residue is an alanine directly after the N-terminal methionine. Like SPs, cTPs have been characterized as having a three-domain structure: an uncharged N-terminal domain of approximately ten residues beginning with MA- and terminating with a G/P, a central domain lacking acidic residues but enriched in S/T and, finally, a C-terminal domain enriched in arginines and potentially forming an amphiphilic β-strand. Various regions of cTPs have also been predicted to form amphipathic α-helices[27]. cTPs from different proteins vary considerably in length (20–100 residues).

The mitochondrial targeting peptide (mTP), which directs nuclear-encoded proteins into the mitochondrial matrix, is enriched in Arg, Ser and Ala, whereas negatively charged residues are rare[25]. The sequence conservation around the cleavage site is low, with an Arg in position −2 or −3 relative to the cleavage site as the most common motif[28,29]. For some mitochondrial proteins, an additional 8–9 residues are removed after initial cleavage of the mTP[30,31]. The mTP is structurally versatile and is known to form an amphipathic α-helix when bound to its receptor on the mitochondrial surface[32], whereas it adopts an extended structure during mTP cleavage[33]. The length of reported mTPs spans from 6 residues up to 85.

In both mitochondria and, in particular, chloroplasts, there is also a significant intraorganellar sorting taking place. Owing to the endosymbiotic origin of chloroplasts and mitochondria, the process of translocating a protein from the stroma of the chloroplast or the matrix of the mitochondrion, to the thylakoid lumen and the mitochondrial intermembrane space (IMS; i.e., between the inner and outer mitochondrial membranes), respectively, is topologically equivalent to and evolutionary reminiscent of prokaryotic protein secretion[34,35]. In particular, the corresponding signals are similar to the SP for secretion. The signals are arranged in a so-called bipartite presequence with the cTP or mTP at the N-terminus, followed by the SP-like signal, which is exposed only after the cleavage of the cTP/mTP part. For proteins destined for the thylakoid lumen, this signal has been termed lumenal transfer peptide (lTP) and it shares as expected significant features with the SP for secretion, for example, the three-region composition and a strongly conserved −3, −1-motif[16,36–38].

Some proteins are dually targeted to both chloroplasts and mitochondria using the same targeting sequence[39–41]. Finally, there are examples of other, less common and less well characterized, pathways for organellar protein sorting[42,43].

Prediction of sorting signals

Methods for computational prediction of protein subcellular localization can roughly be divided into two categories: those that exclusively use the amino-acid sequence of the protein as the input and those that also require other input data, for example, expression levels[44], phylogenetic profiles[45], lexical context in database entries[46] or Gene Ontology numbered terms (GO-numbers)[47–49]. When additional information is provided, reported performance values will in general be higher than for "*ab initio*" sequence-based methods, whereas the applicability will be more limited, as these methods can only be used for examples where this additional information is available. Thus, the use of lexical context or GO-terms requires that the proteins are already to some degree annotated, and this annotation may often include information on subcellular localization, either explicitly (as in the "cellular component" GO-numbers) or implicitly (as in the SWISS-PROT description lines or keywords). The methods using GO-numbers are technically able to accept input consisting of sequence only, but it should be noted that their high reported performances have only been measured on data sets where almost all sequences had GO-numbers, and a high proportion of these were of the "cellular component" type[48].

There are also methods that need only the sequence as input but then use it to search databases for homologs or to look for cooccurrence of certain protein domains[50,51]. In general, annotation by homology to proteins with known subcellular localization

can yield as good a predictive performance as machine learning-based methods, even better if the homolog is very similar. Nair and Rost[52] carried out an extensive analysis of how close two proteins should be in sequence space to have the same subcellular localization. If pairwise identity is used as a measure of homology, the conclusion was that over 70% identical residues in a pairwise BLAST search[53] are needed to correctly infer localization for 90% of all proteins. The BLAST expectation value (*E*-value) or the so-called HSSP distance (a measure developed for protein structure prediction) was found to be better a estimator than percent identities for localization predictive accuracy—regarding *E*-value, its natural logarithm should be below approximately -80 to obtain 90% accuracy. On the other hand, Yu *et al.*[54] found that localization prediction by homology was better than a machine learning method above a pairwise identity cutoff as low as 30%.

This protocol deals primarily with sequence-based tools, which can be used even if no close homologs are found. These can also be divided into two groups: those that predict the actual sorting signals and those that base their prediction on global properties of the sequence, for example, the amino-acid composition.

One of the first attempts at predicting subcellular localization was a weight matrix for secretory SPs[55]. A weight matrix is a simple sequence profile built from an ungapped multiple sequence alignment, where the weights are calculated from the counts of each amino acid at each position in a window around the site of interest. The SP weight matrix has in the past found extremely wide usage. It does not exist as a stand-alone WWW-server, but it is included in PSORT[56] (see below), and it is still used in the tools SPScan (in the GCG commercial package) and SigCleave (in the Emboss public domain package).

Later, several machine learning methods have been used. Among these are neural networks (NNs), hidden Markov models (HMMs) and support vector machines (SVMs). The common feature is that they are data driven, that is, they adjust their free parameters gradually by repeated presentation of a data set, and thereby learn to generalize from the examples they have been trained on. NNs[57] are inspired by the way networks of biological neurons are connected; the input patterns are presented to one or more layers of artificial "neurons" that compute a weighted sum of their inputs and apply a nonlinear function to the sum. When used on biological sequences, NNs typically, like weight matrices, treat the sequence as a series of overlapping windows, calculating a score for each window from a number of position-specific weights; but unlike weight matrices the calculation of the score can be nonlinear, allowing correlations between positions to influence the prediction.

HMMs[58] of the well-known profile architecture can also be seen as an extension of weight matrices, where the alignment used to calculate the weights can contain gaps, so that motifs of varying length can be represented. In addition to the profile architecture, there are many ways to build an HMM; for example, a branched model can represent a choice between alternative patterns, while a cyclic model can represent a repeated pattern. Finally, SVMs[59] treat each input pattern as a set of numbers which is mapped onto a high-dimensional space by the so-called kernel function, and then define an optimal separating hyperplane in that space which classifies the patterns into two categories and maximizes their distance from the hyperplane.

SignalP[60], a predictor of secretory SPs, was among the first to use NNs for sorting signal prediction. Later, in version 2, an HMM was added to the method[61]. In version 3, the input to the NN part has been extended by including, in addition to the moving amino-acid windows themselves, the relative position of each window and the overall amino-acid composition of the entire sequence[62].

An early method to predict mTPs was Mitoprot[63]. It is a feature-based method, using a linear combination of a number of sequence characteristics such as amino-acid abundance, maximum hydrophobicity and maximum hydrophobic moment (α-helix amphiphilicity), which are combined into an overall score. For cTPs, one of the first methods was the NN-based ChloroP[26]. A successor of ChloroP is TargetP, which provides prediction of cTPs, mTPs and secretory SPs[64]. Both ChloroP and TargetP use a combination of NNs to calculate a transit or SP score, and a weight matrix to locate the transit peptide cleavage sites.

Prediction by global sequence properties

In addition to the recognition of the sorting signals, prediction of protein sorting can exploit the fact that proteins of different subcellular compartments differ in global properties, reflected in their amino-acid composition. It has been shown that the signal in the total amino-acid composition, which makes it possible to identify the subcellular localization, is due almost entirely to surface residues[65]. Although the signal prediction methods are probably closer to mimicking the actual information processing in the cell, methods based on global properties can work also for genomic or EST sequences where the N-terminus of the protein has not been included or correctly predicted. Another advantage is that they provide the opportunity to predict localizations for which the sorting signals are not known or not adequately defined. One drawback is that such methods will not be able to distinguish between very closely related proteins or isoforms that differ in the presence or absence of a sorting signal.

Nakashima and Nishikawa[66] pioneered the prediction of protein sorting by global properties by using a simple odds-ratio statistics to discriminate between soluble intracellular and extracellular proteins on the basis of amino-acid composition and residue-pair frequencies. The NNPSL method by Reinhardt and Hubbard[67] was the first to use NNs trained on overall amino-acid composition to predict localization. The method distinguished between three bacterial compartments (cytoplasmic, periplasmic and extracellular) and four eukaryotic compartments (cytoplasmic, extracellular, mitochondrial and nuclear). Interestingly, plant proteins were found to be very poorly predicted and were not included in the final method.

Cedano *et al.*[68] used a simplified version of the nonlinear Mahalanobis distance between amino-acid compositions to discriminate between five possible localizations: intracellular, extracellular, transmembrane, membrane-anchored and nuclear. This approach has been refined by Chou and Elrod[69,70], who used a covariant discriminant method incorporating the Mahalanobis distance to discriminate between three localizations in bacteria and as many as 12 in eukaryotes. A number of SVM-based methods using amino-acid composition also exist, the first being SubLoc by Hua and Sun[71].

Using amino-acid composition as the only input of course discards all information about the sequence order. Some aspect of this order can be incorporated into global property methods, for example by using the frequencies of pairs of amino acids, either consecutive (dipeptide composition) or separated by a certain

number of positions. Nakashima and Nishikawa[66], as mentioned above, found that using pairs with a separation of up to four positions improved performance considerably. The SVM-based methods PLOC[72] and ESLpred[73] also include amino-acid pair frequencies in their input. The problem with this approach is that it creates a high number of input parameters—400 for each separation distance—which can make machine learning methods overfit (see next subsection). A compromise is represented by the so-called pseudo-amino-acid composition approach introduced by Chou[74], where a mathematical function is applied to the amino-acid pairs and then summed over the entire sequence, yielding only one or two extra parameters for each separation distance. The exact nature of this function can vary, but it is most often based on physicochemical properties like the hydrophobicity or side-chain mass. Another possibility is to calculate amino-acid composition separately for a number of subregions of the input sequence. Thus, Esub8 (ref. 75) uses the composition of the first and last half of each sequence, whereas BaCelLo[76] calculates the composition of N- and C-terminal regions of varying width.

Another approach is represented by the feature-based SecretomeP program, which is designed to predict non-classically secreted proteins[19,23]. SecretomeP is built upon the idea that extracellular proteins share certain characteristics regardless of which mechanism was used to secrete them. As it is not possible to collect a sufficiently large data set of non-classically secreted proteins, SecretomeP is trained on a positive data set of classically secreted proteins with the SP removed. Several sequence-derived features were tested for correlation to localization, and selected features were combined using an NN. Some of the features were simple, like number of atoms or number of positively charged residues, others consisted of the outputs of other prediction softwares like ProP or PSORT. Interestingly, protein disorder as predicted by the DisEMBL program[77] turned out to be a feature with large discriminatory power in both Gram-positive and Gram-negative bacteria—secreted proteins generally show a higher degree of disorder. After training and testing on the truncated secretory proteins, the SecretomeP methods have been tested on small test sets of non-classically secreted proteins. The mammalian version was able to predict ten out of 13 human examples, and the Gram-positive version predicted 14 out of 35. Apart from predicting non-classically secreted proteins, SecretomeP can also be useful for predicting subcellular localization of proteins with an incorrectly assigned N-terminus.

PSORT is another early and widely used predictor of subcellular localization using both sorting signals and global features[78], which has subsequently grown to an entire family of prokaryotic and eukaryotic protein localization predictors. The original version of PSORT uses a decision tree to make the prediction: a set of sequence-derived parameters are computed and compared to a representation of a number of "localization rules" that have been collected from the literature. Many of these rules concern the presence of various sequence motifs that enable proteins to be localized to a certain compartment; others deal with amino-acid content in certain regions. PSORT discriminates between 17 different compartments for plants, 14 for animals and 13 for yeast. Later, the method was developed into PSORT II (ref. 79), which uses a statistical classification technique, the *k*-nearest neighbors algorithm, to integrate scores from all the features and arrive at a prediction. Some of the least predictable compartments were dropped in PSORT II, leaving 12 for yeast and animals (plant sequences are not supported in that version). More recent members of the PSORT family will be described later (see **Box 1**).

Data sets and overfitting

A key element when constructing any prediction method is the quality of the data. Extracting a training set from available databases implies a large amount of work and carries a number of critical decisions and pitfalls. As an example, if the "subcellular location" comment in SWISS-PROT[80] contains "endoplasmic reticulum", the protein may be dissolved in the ER lumen, embedded in the ER membrane or even associated with the cytoplasmic face of the ER membrane, and these alternatives are quite different with respect to the sorting signals involved.

Ideally, all training examples should have experimental evidence and not be inferred by similarity or existing prediction methods. When working with rare or poorly characterized sorting signals, this may not be possible to achieve, and examples with labels such as "potential" may have to be included. In that case, care must be taken to avoid circular reasoning where a number of methods simply reproduce the predictions of each other.

Furthermore, errors in the databases are not infrequent and add an element of noise that may pose a serious problem for the training. Fortunately, machine learning methods are potentially capable of dealing with noisy data, for example, it is often seen that NNs learn erroneous examples later than others or even refuse to learn them at all[81]. The way in which noisy data are handled, however, is critically dependent on the learning procedure and the complexity of the model.

One problem that is rarely properly addressed, especially when training global property methods, is homology in the data. All machine-learning methods have many free parameters and therefore have the potential to overfit; that is, instead of learning the general pattern in the training data, they learn each example "by heart", including any noise they might contain. An overfitted machine-learning method that is able to reproduce all its training input/output patterns exactly will typically have a bad generalization ability, that is, it will show a low predictive performance on patterns it has not seen before. The performance of a prediction method must therefore always be measured on a test set that is different from the data set used to train it. If the test set has sequences that are significantly homologous to sequences in the training data, the measurement will use patterns that are very similar to those the predictor has seen before, and the apparent performance of the method will be an overestimate. To compute a true generalizable performance, the data set should be reduced so that no detectably homologous pairs remain[82].

As an example, when training NNPSL, Reinhardt and Hubbard[67] did reduce their data set, but only removed sequences with more than 90% identity to each other, which is clearly much higher than the detectable homology threshold. Several newer methods, including the above-mentioned SubLoc[71], are trained with the NNPSL data set, which thereby unfortunately has received a kind of benchmark status. Park and Kanehisa[72], when training the 12-localization eukaryotic SVM-based predictor PLOC, constructed a data set in a similar way, but from a more recent SWISS-PROT version. In this data set, which has also been used in other methods, for example, LOCSVMPSI (ref. 83), they allowed at most 80% pairwise identity between sequences, thus reducing the impact of

sequence homology on the performance evaluation a bit. However, studies by Nair and Rost[52] and Yu *et al.*[54] indicate that subcellular localization is significantly conserved even when the sequence identity is at 25–40%. Thus, the aforementioned data sets are not sufficiently homology reduced. In fact, Pierleoni *et al.*[76] recently showed that the prediction of subcellular localization in the NNPSL data set could be carried out with a BLAST search[53], where the localization of each protein was simply predicted to be that of the closest homolog within the data set. The performance of this simple procedure was actually better than that of the machine-learning-based methods NNPSL and SubLoc and at the same level as two newer methods (LOCSVMPSI (ref. 83) and ESLpred(ref. 73)). Therefore, care should be taken when comparing performance values reported in the literature.

MATERIALS

EQUIPMENT

A computer with access to the internet and a web browser.

EQUIPMENT SETUP

Data Your input protein sequences should be written in the standard one-letter code. Thus, the allowed characters are ACDEFGHIKLMNPQRSTVWY and also X (unknown). Spaces and line breaks will be ignored and will not affect the predictions. On the CBS (Center for Biological Sequence Analysis) localization prediction servers and most other servers described here, you can choose to upload your sequences either through copy-and-paste in an input window or by submitting an entire sequence file (which must be in plain text format). If submitting more than one sequence, the file must be in FASTA format: each sequence should be preceded by a line beginning with the ">" sign followed by the sequence name (it does not have to be a unique name). Two example data sets, one prokaryotic and one eukaryotic, that the reader may use for testing the methods described here are found at the supplementary webpage: http://www.cbs.dtu.dk/suppl/natureprotocols/. These two files also serve as an example of the FASTA format.

Programs The following localization prediction programs will be described in this protocol—see PROCEDURE and TROUBLESHOOTING for usage and web addresses. A list of the programs with the web addresses as clickable links can also be found at the supplementary webpage: http://www.cbs.dtu.dk/suppl/natureprotocols/. Programs hosted at the Center for Biological Sequence Analysis are marked "(CBS)":

TargetP 1.1 (CBS): secretory SPs, mTPs and cTPs in eukaryotes.
ChloroP 1.1 (CBS): cTPs in plants.
SignalP 3.0 (CBS): secretory SPs in eukaryotes and Gram-negative and Gram-positive bacteria.
TatP (CBS): twin-arginine translocation SPs in bacteria.
TMHMM 2.0 (CBS): transmembrane α-helices.
B2TMR and HMM-B2TMR: transmembrane β-barrels.
big-Pi: glycosylphosphatidylinositol (GPI) membrane anchors in eukaryotes.
NMT: myristoyl membrane anchors in eukaryotes.
Myristoylator: myristoyl membrane anchors in eukaryotes.
LipoP (CBS): lipoprotein SPs in bacteria.
PROSITE pattern PS00014: ER lumenal retention motifs in eukaryotes.
Golgi predictor: Golgi retention signals in eukaryotes.
PredictNLS: nuclear localization signals (NLSs) in eukaryotes.
NucPred: NLSs in eukaryotes.
NetNES (CBS): leucine-rich nuclear export signals in eukaryotes.
PeroxiP: C-terminal peroxisomal targeting signals in eukaryotes.
PTS1: C-terminal peroxisomal targeting signals in eukaryotes.
SecretomeP (CBS): SP-less secretion in mammals and bacteria.
NetStart (CBS): start codon prediction in eukaryotic DNA sequences.
ProP (CBS): propeptide cleavage in eukaryotes.
Phobius: combined prediction of SPs and transmembrane α-helices.

In addition, a number of non-CBS multicategory localization prediction programs will be described in **Box 1**. **! CAUTION** On the CBS servers, there are certain restrictions on the size of each submission and the number of submissions allowed per 24-h period, to prevent server overload. All normal and even most heavy users will easily be accommodated within these limits, although you may have to split up your sequence file in some cases.

For those who have problems with these restrictions or who would prefer to keep their sequence data in-house, most of the CBS programs are also available as stand-alone program packages for local use. More information is available on the server pages.

PROCEDURE

1| If you have eukaryotic sequences, go to TargetP 1.1 (http://www.cbs.dtu.dk/services/TargetP/; **Fig. 1**) (ref. 64). If you have bacterial sequences, proceed directly to Step 7. TargetP will give a prediction of which kind, if any, of N-terminal sorting signals your sequences have: SP, mTP, cTP (if applicable) or as a last option "no sorting signal" ("other").

2| Customize TargetP. Pick the Plant version for sequences from higher plants and the Non-plant version for all other eukaryotic proteins. If you would like TargetP to predict the cleavage sites of the predicted presequences, tick the "Perform cleavage site predictions" box. For SPs and cTPs, SignalP 1.1 (ref. 60) and ChloroP 1.1 (ref. 26), respectively, are automatically used to predict the cleavage sites and the results incorporated into the TargetP output. You are also asked to choose cutoffs for the predictions. The default is to simply let the highest score determine the prediction, without requiring this highest score to be above any particular value. We recommend that you use this default. If you still choose to specify your own cutoff, or to use either of the two predefined sets corresponding to a specificity of 0.90 or 0.95, note that the predicted score has to be both the highest and above the specified cutoff in order for a prediction to be made.

! CAUTION Algal sequences were excluded from the training set used to construct TargetP. If you wish to use TargetP on algal sequences, use the Plant version, but please be cautious in the interpretation of the results. TargetP is also only able to predict nuclear-encoded proteins, and has not been tested on organelle-encoded proteins.

3| Run TargetP. Paste in your sequence or upload your sequence file. Make sure that, if possible, each protein in your set contains its 130 N-terminal residues. Missing N-terminal residues or a too short sequence make the prediction less reliable. If a protein is shorter than 130 residues, include the entire sequence. Click "Submit" to initiate the TargetP prediction.

4| Examine the output (**Fig. 2**). For every submitted protein, TargetP reports a prediction score for each of the four (Plant version) or three (Non-plant version) possible outcomes (SP, mTP, cTP, "other"), and assigns a predicted localization in the "Loc" column: (C)hloroplast, (M)itochondrial, (S)ignal peptide or "_" ("other" localizations). If you specified your own cutoffs (see

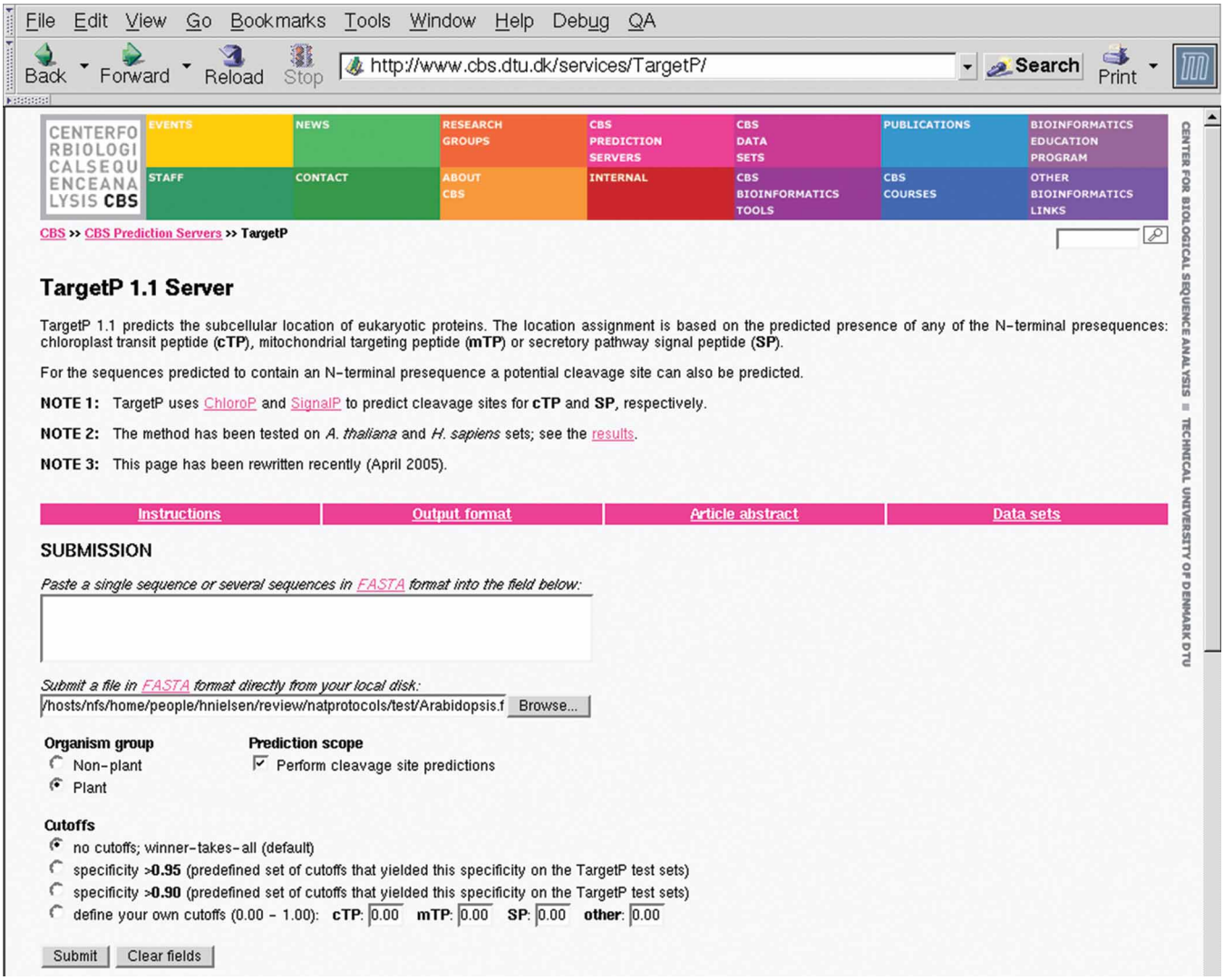

Figure 1 | A screenshot of the TargetP server.

Step 2) and the highest score is below the cutoff for that particular presequence, a "don't know" symbol ("*") will be output. Based on the scores, TargetP provides a "reliability coefficient", RC, which is a measure of how confident TargetP is in each prediction. The RC ranges from 1 (very reliable prediction; virtually no false positives detected in the TargetP test set) to 5 (not reliable prediction; many false positives detected)—see the detailed discussion in ANTICIPATED RESULTS. If you chose to include cleavage site predictions, the predicted length of amino-acid residues of the presequence is reported in the "TPlen" column. TargetP also outputs the length of your submitted sequences in the "Len" column.

If TargetP predicts an SP, go to SignalP 3.0 to have a detailed prediction of the SP and its cleavage site (Step 7).

If TargetP predicts another category but the reliability is low (i.e., the RC is 4 or 5), you should also consult SignalP 3.0 (Step 7) to get a second opinion about whether your sequence could be an SP. SignalP 3.0 is newer and based on a larger data set than the SP part of TargetP 1.1.

If TargetP predicts a cTP, go to ChloroP (Step 5) if you wish a detailed report of the cTP scores along the sequence. This might help you to decide how much you trust the cTP (and the cleavage site) prediction.

If TargetP predicts an mTP or a cTP, and you wish to investigate whether your sequence contains either an IMS signal in addition to a predicted mTP or a thylakoid lTP in addition to a predicted cTP, then go to Step 9.

If TargetP predicts "other", proceed to Step 11 to investigate if your protein contains predicted transmembrane regions.

If you think your protein might be secreted although you found no SP, see TROUBLESHOOTING point 1.

? TROUBLESHOOTING

5| Go to ChloroP 1.1 (http://www.cbs.dtu.dk/services/ChloroP/)[26]. Note: use ChloroP only if you have plant (or algal) sequences and would like to get a detailed report of the scores of a predicted cTP. Paste in your sequence or upload your

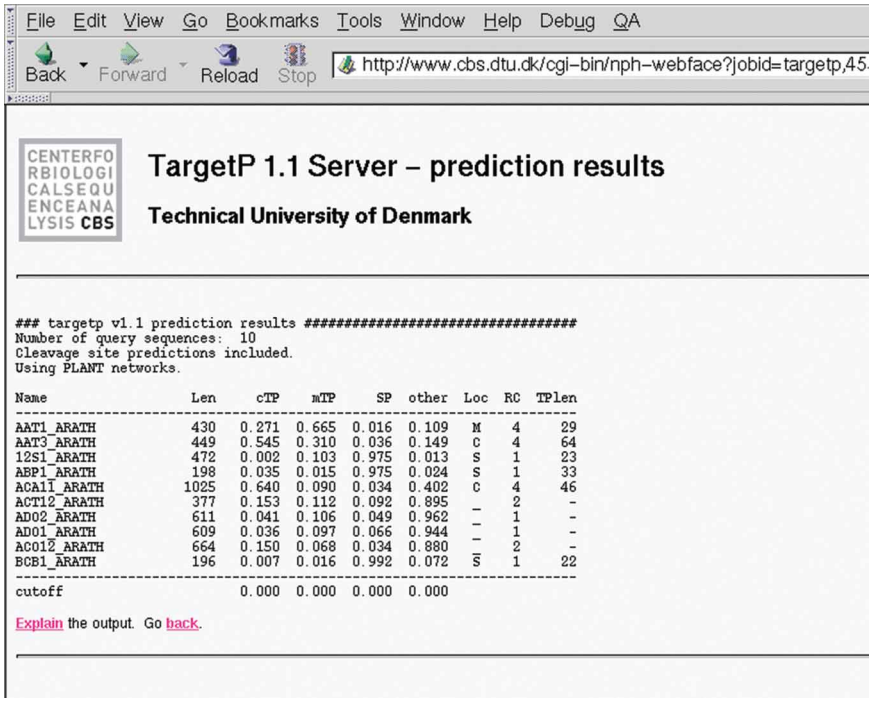

Figure 2 | Output from the TargetP server, tested on ten sequences from *Arabidopsis thaliana*. The "Loc" column contains the prediction: "C" for chloroplast, "M" for mitochondrion, "S" for secretory or "_" for "other". Nine of the answers are correct, whereas the transmembrane protein ACA11_ARATH should have had a "_". The proteins shown are as follows: AAT1: aspartate aminotransferase, mitochondrial precursor; AAT3: aspartate aminotransferase, chloroplast precursor; 12S1: 12S seed storage protein CRA1 precursor; ABP1: auxin-binding protein 1 precursor; ACA11: putative calcium-transporting ATPase isoform 11; ACT12: actin-12; AD02: adagio protein 2 (LOV kelch protein 2); AD01: adagio protein 1 (LOV kelch protein 1); AC012: putative acyl-coenzyme A oxidase 1.2, peroxisomal; BCB1: blue copper protein precursor (phytocyanin 1).

sequence file. Include if possible the 100 N-terminal residues (this is the region on which ChloroP was trained). Tick the "Detailed output" box if you wish the detailed score report (see below). Click "Submit".

6| Examine the output from ChloroP. ChloroP will report for each protein a "Score" indicating how likely it is that the protein has a cTP. Note that the range of this score is rather compressed, and usually lies between 0.4 and 0.6. The cutoff for predicting a protein to have a cTP is 0.5 (a score above this value will yield a "Y" in the "cTP" column), and any score above 0.55 is a quite strong prediction. The "CS-score" is the maximum score of a weight matrix used to predict the cleavage site, and is used to decide the "cTP length", which is the predicted length of the cTP. If the "Detailed output" option was chosen, ChloroP will also report three different scores for each amino-acid residue in each submitted sequence: the residue-wise NN score ("Raw"), a derivative of this score ("Deriv."), used to define the region within which to apply the weight matrix and search for the cleavage site, and the cleavage site score for each position ("CS-score"). The cleavage site is directly N-terminal of the residue with the highest CS-score. If you have found a cTP, go to Step 9 in order to investigate whether your sequence contains a thylakoid lTP in addition to the cTP. Otherwise, proceed to the next step.

7| Go to SignalP 3.0 (http://www.cbs.dtu.dk/services/SignalP/; **Fig. 3**) (ref. 62). It is important that you choose the correct organism group (Eukaryotes, Gram-negative bacteria or Gram-positive bacteria), as the SPs from these groups are quite different. If in doubt whether a bacterium is Gram-positive or Gram-negative, check the NCBI Taxonomy database (http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html): Gram-positive bacteria comprise the groups Firmicutes and Actinobacteria, whereas all other bacteria are basically Gram-negative. However, it should be noted that the training set of Gram-negatives is heavily dominated by Proteobacteria, and we actually do not know whether some of the rarer groups have SPs that differ from those of Proteobacteria.

The other options can safely be left at their default values. Paste in or upload your sequences and click "Submit".

! CAUTION By default, SignalP truncates each sequence to 70 residues to make the graphical output more readable and to avoid false positive cleavage site predictions from downstream regions of the sequence. If you choose to disable truncation, note that although SignalP-NN performance is only marginally affected, the sensitivity of SignalP-HMM will be drastically reduced.

8| Examine the output from SignalP. The many scores in the output can be overwhelming, so we will give an explanation here. SignalP-NN is a system of two artificial NNs, one trained to distinguish between SP and non-SP windows, and the other trained to distinguish SP cleavage sites from everything else. The output scores from the networks are called S-score and C-score, respectively. The S-score can be interpreted as an estimate of the likelihood of the position belonging to the SP, whereas the C-score can be interpreted as an estimate of the likelihood of the position being the first in the mature protein (position +1 relative to the cleavage site). If there are several C-score peaks of comparable strength, the true cleavage site may often be found by inspecting the S-score curve in order to see which of the C-score peaks coincides best with the transition from a high to a low S-score. To formalize this and improve the prediction, we defined the Y-score, which is the geometric average of the C-score and a smoothed slope of the S-score. The Y-score gives the best estimate of where the SP is cleaved. The S-, C- and Y-scores are shown in the graphical output of SignalP-NN (**Fig. 4**).

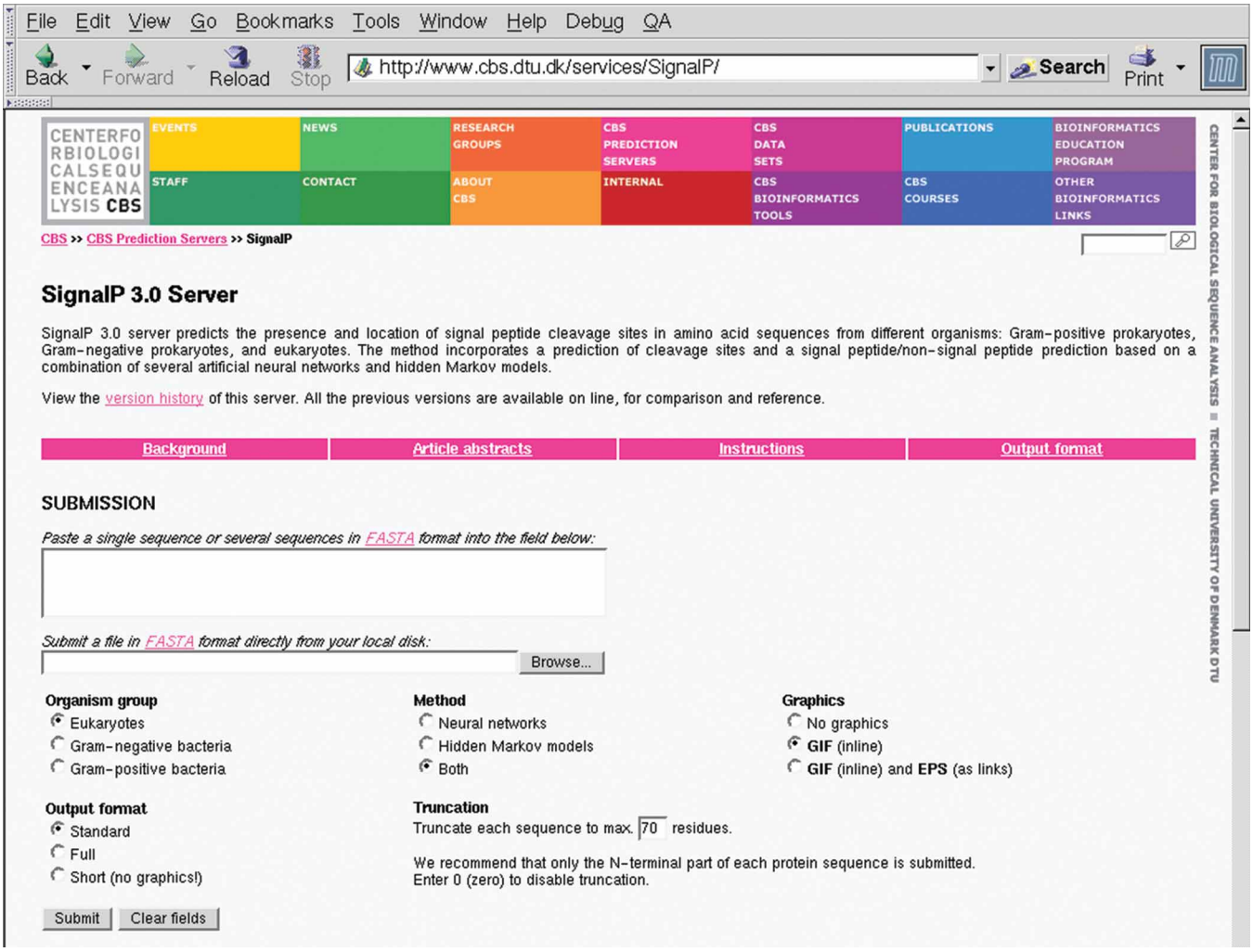

Figure 3 | A screenshot of the SignalP server.

For the discrimination between SPs and non-SPs, we have defined the D-score, which is simply the average of the maximal Y-score and the mean S-score (calculated from the N-terminus to the position of the highest Y-score). The D-score was found to give a better discrimination than S-score or Y-score alone.

While SignalP-NN looks at the sequence as a series of overlapping windows, SignalP-HMM fits the entire sequence onto a model of an SP. The model has separate sub-states representing the n-, h- and c-regions and the cleavage site. For each position in the sequence, SignalP-HMM outputs the posterior probability that the position is in any of these sub-states. These probabilities are shown in the graphical output of SignalP-HMM (**Fig. 5**). SignalP-HMM is a branched HMM, which, in the eukaryotic version, distinguishes between three types of sequence: signal peptide (S), signal anchor (A) and other (Q). A signal anchor, just like an SP, contains a region of hydrophobic residues close to the N-terminus, but it is not cleaved. Instead, it remains as a transmembrane α-helix in the membrane and anchors the protein to the membrane in an N-in/C-out orientation. A protein anchored in this way is called a type II membrane protein[84]. For bacteria, too few type II membrane proteins were found in the database to train a signal anchor branch.

There are also inverted signal anchors that insert in the membrane with the N-terminus first and anchor the protein in an N-out/C-in orientation; such a protein is known as a type III membrane protein[84]. SignalP-HMM is not trained to recognize inverted signal anchor sequences, but TMHMM and other membrane protein topology predictors (see Step 11) can handle both signal anchor and inverted signal anchor sequences.

It is always a good idea to check the graphical output of SignalP. It can be seen from the graphs whether the prediction shows a typical SP with one clear cleavage site or a more ambiguous picture. If there is more than one peak in SignalP-NN's Y-score and SignalP-HMM's cleavage site probability, the predicted cleavage site should be taken with caution. Of course, checking the graphical output of every sequence may be unrealistic if you have submitted an entire proteome—in that case,

you would probably prefer to use the "short" output option—but it is still possible to assess the strength of the individual predictions by looking at the value of the D-score; see ANTICIPATED RESULTS.

If you think your protein might be secreted although you found no SP, see TROUBLESHOOTING point 1.

If the graphical output from SignalP-NN shows a low S-score in the first part of the sequence but a higher value downstream, see TROUBLESHOOTING point 2.

If you have found an SP in a eukaryotic sequence but do not trust the cleavage site prediction, consult TROUBLESHOOTING point 3.

Otherwise, proceed to Step 10 for prokaryotes or Step 11 for eukaryotes.

? TROUBLESHOOTING

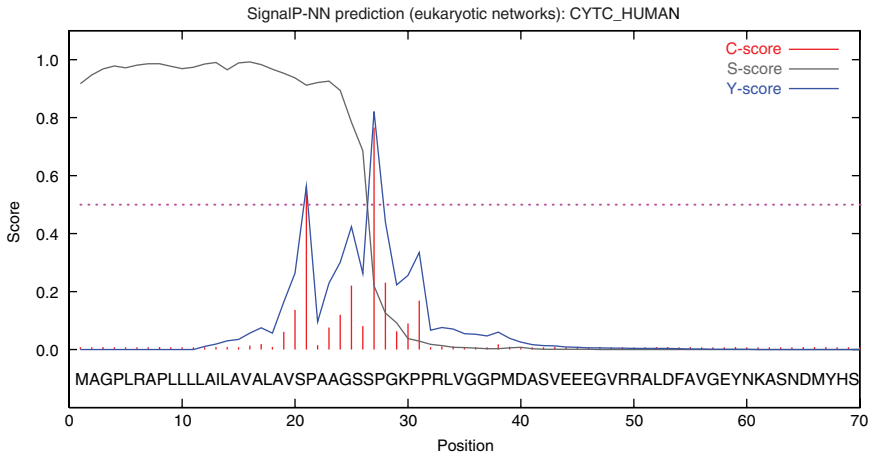

Figure 4 | The graphical output of SignalP-NN, showing C-, S- and Y-score. The cleavage site is predicted to be at the position of maximal Y-score. The example shown here is human cystatin C precursor.

9| For proteins that TargetP predicted to be in the chloroplast or mitochondrion, try also to submit the sequence with the transit peptide removed to SignalP (Step 7), to see if it might be localized to the IMS (mitochondrion) or thylakoid lumen (chloroplast). You should use the prokaryotic version of SignalP but there is no clear consensus whether the Gram-positive or Gram-negative version is preferable. Note that in particular the cTP cleavage site prediction is relatively unreliable, and instead of just removing the predicted cTP you should consider submitting several truncated versions of your protein to SignalP. An ambitious approach tested previously for the lTP prediction[85] is to remove the 20–80 N-terminal residues (very few cTPs fall outside of this length interval) in steps of five residues, and submit each of these truncated sequence versions to SignalP. If at least one truncated sequence is predicted to have an SP, and also fulfill certain other requirements (e.g., overall presequence length), this is interpreted as the protein having a bipartite cTP + lTP presequence. The mTP cleavage site prediction is better than the cTP cleavage site prediction, and this elaborate scheme may not be warranted to predict an IMS signal. Note, however, no systematic study that uses this scheme for predicting IMS signals has been published to date.

For chloroplast proteins, go to the next step to check for Tat SPs—this special translocation system, known mainly from Gram-negative bacteria, is also found in thylakoid membranes. For mitochondrial proteins, proceed to Step 11 to see if your protein is a membrane protein.

10| If your sequence is from a bacterium (or a plant protein with a cTP; see previous step), there is a possibility that it might be using the so-called Tat (for twin-arginine translocation) pathway[86]. To translocate proteins across the plasma membrane (inner membrane in Gram-negative bacteria or thylakoid membrane in chloroplasts), this pathway uses a secretion pore that is separate from the standard translocon, which is built from products of the Sec genes. The signal peptidase, however, is the same in the two pathways. One surprising aspect of the Tat pathway is that it can apparently transport fully folded proteins, whereas proteins using the Sec pathway must be kept in an unfolded conformation. In Gram-negative bacteria, the Tat pathway is preferentially used by periplasmic proteins with metal cofactors. Their SPs are longer but less hydrophobic than Sec SPs[87] and they carry a sequence motif with two consecutive conserved arginines in the n-region. Tat SPs can

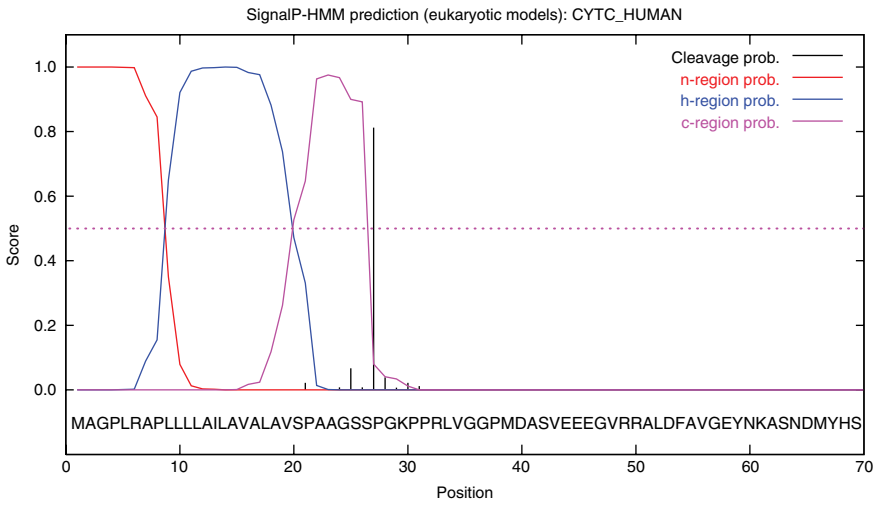

Figure 5 | The graphical output of SignalP-HMM, showing the posterior probabilities for n-, h- and c-region and cleavage site. The example is the same as in **Figure 4**.

be predicted with our tool TatP (http://www.cbs.dtu.dk/services/TatP/)[88], which is based on an NN in combination with a simple pattern matching of the twin-arginine motif. Note: You should consult TatP even if your SignalP prediction was negative, as the lower hydrophobicity of Tat SPs sometimes makes them go undetected by SignalP.

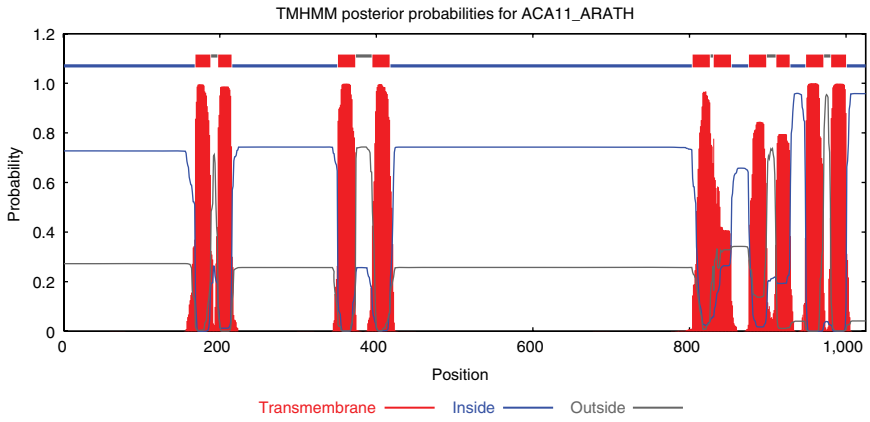

Figure 6 | The graphical output of TMHMM, showing the posterior probabilities for transmembrane, inside (i.e., cytoplasmic), and outside (i.e., lumenal or exterior) regions. In this example (*Arabidopsis thaliana* putative calcium-transporting ATPase isoform 11), ten transmembrane regions are predicted.

11| To check whether your protein is an integral membrane protein, try a transmembrane α-helix predictor. There are several of these available on the net (a list is found at http://www.psort.org/), but we recommend TMHMM 2.0 (http://www.cbs.dtu.dk/services/TMHMM/)[89], an HMM predictor from our own group. In 2001, a comparative study of transmembrane α-helix predictors found TMHMM 2.0 to be the best performing program out of 14 (see ref. 90). Another comparative study from 2002, which analyzed as many as 27 different methods, found that advanced (machine learning) methods were more accurate than simple hydrophobicity-based predictions, but no single advanced method was best by all scores[91]. In this study, TMHMM 1.0 was found to be among the most representative methods (version 2.0 was not tested). A comparison from 2005 placed TMHMM 2.0 as one of the four methods that consistently performed well[92]. Even more recently, a combined review and benchmark placed TMHMM 2.0 as number three in a comparison of ten methods regarding the consistency of their predictions, but ended by concluding that "if one single server is to be recommended, it is probably TMHMM"[93]. Note that both proteins with and without SPs or transit peptides can have transmembrane helices. If an SP is followed by a single transmembrane helix further downstream, this transmembrane region is referred to as a stop-transfer sequence. In this case, the mature protein has an N-out/C-in orientation and is called a type I membrane protein[84]. An example of the graphical output of TMHMM 2.0 is shown in **Figure 6**.

If TMHMM predicts a transmembrane helix in the same region where SignalP predicts an SP, see TROUBLESHOOTING point 4.

? TROUBLESHOOTING

If you have found one or more transmembrane helices in a eukaryotic sequence, proceed to Step 14 to see if your protein is retained in a compartment along the secretory pathway.

If you have found one or more transmembrane helices in a prokaryotic sequence, proceed directly to Step 17.

Otherwise, proceed to the next step.

12| Not all transmembrane proteins have α-helical transmembrane domains—some have a transmembrane β-barrel that forms a pore with a hydrophobic outside and a hydrophilic inside[94]. These proteins are found in the outer membranes of Gram-negative bacteria, chloroplasts and mitochondria, and they are generally more difficult to predict than the α-helical proteins. A number of predictors are available (see http://www.psort.org/ for a list), for example the NN-based B2TMR (ref. 95) and the HMM-based HMM-B2TMR (ref. 96) (both found at http://gpcr.biocomp.unibo.it/predictors/—note that the website requires a free registration).

If you have found a transmembrane β-barrel, go directly to Step 17. Otherwise, proceed to the next step.

13| Even if your protein does not have a transmembrane sequence, there is still a possibility that it might be anchored to a membrane by a covalently attached lipid group. For eukaryotes, perform option A, for prokaryotes perform option B.

(A) *Sequences from eukaryotes*: Check for the presence of a GPI anchor near the C-terminus or a myristoyl anchor at the N-terminus, depending on whether you have found an SP or not. Proteins with GPI anchors also have SPs and are thus anchored to the lumenal or outside face of the membrane. The GPI moiety is attached to the new C-terminus after cleavage of a C-terminal propeptide. Myristoylated proteins do not have SPs and are anchored to the cytoplasmic face of the membrane, with the myristoyl moiety attached to an N-terminal glycine. At present, we do not have publicly available in-house methods to predict these two lipid modifications at CBS, but they can be predicted with the big-Pi and NMT tools (http://mendel.imp.univie.ac.at/mendeljsp/sat/index.jsp)[97], which are based on sequence profiles and physical properties of the amino-acid side chains in various regions around the modification site. Myristoylation can also be predicted with the NN-based Myristoylator (http://www.expasy.org/tools/myristoylator/)[98].

(B) *Sequences from prokaryotes*: Check whether your sequence could be a prokaryotic lipoprotein by using LipoP (http://www.cbs.dtu.dk/services/LipoP/)[99]. These proteins have a special SP that is cleaved by lipoprotein signal peptidase (also known as signal peptidase II) instead of the standard signal peptidase, and their cleavage site has a characteristic

consensus signal with a 100% conserved cysteine at position +1, to which the lipid group is attached[100]. SignalP will typically predict these sequences as SPs, but with an incorrect cleavage site. For the development of LipoP, both an NN and an HMM approach were tried and found to perform well, but only the HMM version is implemented in the server. The HMM is a branched architecture which is able to distinguish four types of sequence: signal peptidase I-cleaved SPs, signal peptidase II-cleaved lipoprotein SPs, N-terminal transmembrane helices or cytoplasmic sequences.

LipoP was trained only on Gram-negative bacteria, as most of the known examples are from these, but nevertheless it was able to correctly predict 26 out of 28 lipoproteins in a test set from Gram-positive bacteria. In Gram-negative bacteria, lipid-anchored proteins bind to the periplasmic face of either the plasma membrane or the outer membrane. The choice between these two membranes seems to depend on the second amino acid after the cleavage site. For prokaryotic proteins, proceed directly to Step 17. For eukaryotic proteins, proceed to the next step.

14| If you have found an SP or a signal anchor in a eukaryotic sequence, remember that it does not necessarily imply that the protein is secreted or integral to the plasma membrane—it may be retained in a compartment along the secretory pathway (ER or Golgi) or sorted to the lysosome. ER lumenal proteins have a C-terminal retention signal with the consensus sequence KDEL (where the K and the D are not totally conserved)[18]. These can be predicted by the PROSITE (ref. 101) pattern PS00014/ER_TARGET. To scan your protein sequence for PROSITE patterns, go to http://www.expasy.org/prosite/, paste in your sequence and click "Quick Scan".

ER membrane proteins have two less well defined retention motifs, one is the dilysine motif, which is located close to the C-terminus of type I membrane proteins, the other is the diarginine motif, which is located close to the N-terminus of type II membrane proteins[18]. Both of these motifs are found in the cytoplasmic part of the protein. PSORT I and II (http://psort.ims.u-tokyo.ac.jp/)[79] make an attempt at predicting ER-resident membrane proteins based on these two signals, but with low reliability.

Golgi-resident proteins are most often transmembrane proteins, but their retention signals are not well characterized and vary between protein families[18]. A large group of Golgi proteins, including glycosyltransferases, are type II membrane proteins, where the retention signal seems to reside within the transmembrane helix. A predictor for this signal exists at http://ccb.imb.uq.edu.au/golgi/golgi_predictor.shtml[102].

15| If TargetP predicted "other" and you found no transmembrane domains, it basically means that the protein can be cytoplasmic, nuclear or peroxisomal. TargetP in its current version cannot distinguish between these three localizations as it recognizes only N-terminal sorting signals. Consult the prediction servers mentioned in this step for nuclear localization, and in the next step for peroxisomal localization.

Nuclear localisation signals (NLSs) are generally rich in lysine and arginine. They can occur internally in the sequence and are not cleaved. Some of them are called bipartite NLSs, as they consist of two basic regions separated by a spacer of approximately ten residues[103]. The method PredictNLS (http://cubic.bioc.columbia.edu/predictNLS/)[103] predicts nuclear localization by comparing the query sequence to a database of nuclear motifs. This database initially consisted of experimentally verified NLS sequences collected from the literature, and was subsequently expanded by two procedures: first by including matching sequences from homologs with more than 80% identity, then by "*in silico* mutagenesis" where residues in the given motifs were changed or removed, and the mutated motifs were accepted if they matched at least two distinct families of nuclear proteins but no non-nuclear proteins. NucPred (http://www.sbc.su.se/~maccallr/nucpred/)[104] also uses a set of nuclear-specific patterns (regular expressions), but instead of being based on experimentally known NLSs, the patterns are found directly from the data sets of nuclear and non-nuclear proteins by a genetic programming algorithm. Thus, the patterns recognized by NucPred do not necessarily function as localization signals but may include other motifs specific to nuclear proteins, such as DNA-binding domains. A paper describing NucPred is available from the server page.

NLSs are generally weaker than the cleaved sorting signals recognized by TargetP and SignalP. To complicate matters further, not all nuclear proteins have an NLS—some are imported by binding to other proteins having a signal[105]. Therefore, nuclear localization predictors cannot be expected to have as high a predictive performance as secretory pathway, mitochondrial or chloroplast predictors.

Unlike most other protein sorting mechanisms, nuclear import is not a one-way process. Many proteins spend only some of their time in the nucleus and can shuttle back and forth between the nuclear and cytoplasmic compartments many times during their life cycle[106]. Several pathways of nuclear export have been identified, but the best known export pathway requires a leucine-rich nuclear export signal. This can be predicted with our tool NetNES (http://www.cbs.dtu.dk/services/NetNES/)[107], which is based on a combination of NNs and HMMs.

16| Peroxisomes also have their own protein import machinery. Two signals for peroxisomal soluble proteins are known: the uncleaved C-terminal PTS1 and the N-terminal PTS2, which is sometimes cleaved[108]. PTS1 is a relatively conserved C-terminal tripeptide (most common sequence is SKL) directing proteins to the peroxisome, and two specialized predictors, PeroxiP and PTS1, are available. As the known PTS1 tripeptides are also abundant in known non-peroxisomal proteins, these predictors look

at a larger C-terminal region (12 residues). PTS2 is rarer and less well characterized, and only PSORT II makes an attempt at predicting it. Peroxisomal membrane proteins have been proposed to have either of two targeting signals, mPTS-1 and mPTS-2 (ref. 108), but these are even less characterized, and to our knowledge no available method predicts them. PeroxiP (http://www.bioinfo.se/PeroxiP/)[109] is a tool from our group even though it is not hosted at the CBS. It is based on a set of accepted PTS1 tripeptides and an NN- and SVM-based machine-learning module, which is trained to discriminate on the nine additional residues and on the amino-acid composition of entire protein (allowing at most 25% sequence identity between any two proteins in the data set). The combination of the motif module and the machine-learning module yields a localization prediction. The signal is still rather weak and the performance significantly lower than what is possible to obtain for most other compartments. To improve the performance, the web server uses TargetP and TMHMM in a prescreening step to remove highly confident predictions of secreted and transmembrane proteins before submitting the sequence to PeroxiP.

The PTS1 predictor (http://mendel.imp.univie.ac.at/PTS1/)[110] is based on sequence profiles and physical properties of the amino-acid side chains in various regions within the 12 most C-terminal residues of the protein.

17| A sequence-based prediction of subcellular localization should always be complemented by a database search. First, search the annotated SWISS-PROT database (http://expasy.org/sprot/)[80] with your sequence and check the annotation of the closest homologs. The search can be performed with BLAST (ref. 53), for example, at NCBI (http://www.ncbi.nlm.nih.gov/BLAST/)—click "Protein-protein BLAST (blastp)" and choose "swissprot" as the database. If you do not find annotated close homologs, search the entire protein sequence universe (go to NCBI BLAST as before, but choose "nr" as the database). Then run the best hits through the predictors. If all, or almost all, of the best hits yield the same predicted localizations as your original sequence, you can be more confident in the prediction.

18| Do not believe blindly in the CBS tools; go for a second opinion from other subcellular localization predictors available on the web. A fairly complete list can be found at the PSORT main page (http://www.psort.org/). Here (**Box 1**) we have selected a few that we can recommend, starting with the recent members of the PSORT family. In making the selection, we have looked at the general usability of the websites and at the quality of the data set construction, including homology reduction. Note, however, that if you require a eukaryotic prediction with more than six localization categories, there is no available predictor that is based on properly homology reduced data.

? TROUBLESHOOTING

1. If you think your protein might be secreted although you found no SP, consider the possibility that it might be secreted via a non-classical pathway. The SecretomeP server (http://www.cbs.dtu.dk/services/SecretomeP/) offers a feature-based prediction of extracellular localization for proteins without an SP from mammals[19] and bacteria[23]; see also "Prediction by global sequence properties" in the INTRODUCTION.

2. If the graphical output from SignalP-NN shows a low S-score in the first part of the sequence but a higher value downstream, check whether there is a second methionine in the sequence, which could be the actual start codon. Coding regions predicted from genomic DNA sequence quite often have wrong start codons assigned. To get a prediction of a eukaryotic start codon based on the DNA sequence, you might want to try the NN-based NetStart server (http://www.cbs.dtu.dk/services/NetStart/)[111].

3. For eukaryotic proteins with SPs: if you have determined the N-terminus of your mature protein by experimental means but SignalP reports a cleavage site upstream of that, consider that your protein might have an N-terminal propeptide that is cleaved after SP cleavage. There are many types of propeptides, and they are cleaved by a wide range of peptidases, but one of the more common forms in eukaryotes is the N-terminal propeptide cleaved at mono- or dibasic sites by subtilisin/kexin-like proprotein convertases (PCs), of which the best characterized is furin. These cleavage sites can be predicted with our NN-based tool ProP (http://www.cbs.dtu.dk/services/ProP/)[112]. The method includes two networks, one specific to furin and another for PCs in general. The sensitivity is much higher for the furin-specific network, and the default is to run only this predictor; if you also want results for the general PC network, remember to select that option before submitting your sequence.

Bacteria, especially Gram-positive bacteria, also have periplasmic or extracellular peptidases, which cleave N-terminal propeptides, but to our knowledge, no predictor is available for these.

4. If TMHMM predicts a transmembrane helix in the same region where SignalP predicts an SP, this means that there is a false-positive prediction from either SignalP or TMHMM. Note that SignalP-NN in its current version was not trained with membrane proteins in the negative set and therefore has a higher false-positive rate on those, whereas TMHMM similarly was not trained to avoid predicting SPs as transmembrane. In this case, we suggest that you try Phobius (http://phobius.cgb.ki.se/)[113]. This is an HMM that combines the architectures of SignalP-HMM and TMHMM. It is less accurate than SignalP-HMM in finding the correct cleavage site, partly because it does not distinguish between eukaryotes and Gram-positive and Gram-negative bacteria, but it has a better discrimination between SPs and transmembrane helices than the present version of SignalP.

In addition, you should search the databases with your sequence; see Step 17 of the PROCEDURE. If your putative SP matches a region that is clearly transmembrane in another protein (e.g., if it is annotated as such without "Potential" or "By similarity", or if it occurs downstream of a highly confident SP prediction), it is probably a transmembrane helix and not an SP.

BOX 1 | A SELECTION OF NON-CBS MULTICATEGORY PROTEIN SUBCELLULAR LOCALIZATION PREDICTION PROGRAMS

This list with the web addresses as clickable links can also be found at the supplementary webpage http://www.cbs.dtu.dk/suppl/natureprotocols/.

WoLF PSORT (http://wolfpsort.org/)[123] is a successor to PSORT II for eukaryotes. Predictions are based on PSORT II features (including known sorting signal motifs), iPSORT features (see below), amino-acid content and sequence length. Features are selected and weighted in a learning procedure and combined using the *k*-nearest neighbors classifier. WoLF PSORT predicts 11 localizations for animals and plants, and ten for fungi. In addition, it is capable of predicting some dual localizations such as "nuclear and cytoplasmic". It achieves a significant improvement in prediction accuracy over PSORT II, while actually being simpler in the sense that it uses fewer features for classification. However, the comparison is difficult as the data sets have not been homology reduced. From the output page, there is a link to a very detailed report of each prediction, showing all features used for classification, so that the evidence can be evaluated. A paper describing WoLF PSORT is available from the server page.

PSORTb v.2.0 (http://www.psort.org/psortb/)[124] is a predictor specifically designed for bacteria, discriminating between five possible localizations in Gram-negatives (cytoplasm, inner membrane, periplasm, outer membrane and extracellular) and four in Gram-positives (cytoplasm, membrane, cell wall and extracellular). It uses a combination of BLAST homology searches to proteins of known localization, PROSITE motifs and profiles[101], a collection of outer membrane-specific motifs, SP and transmembrane helix predictors based on HMMs, and one SVM-based prediction module for each localization using the occurrence of frequent subsequences. The data set has not been homology reduced. PSORTb has been designed to yield as high a precision level as possible, at the expense of recall; so in some cases it will output an "unknown" prediction.

iPSORT (http://biocaml.org/ipsort/iPSORT/)[125] note: the URL given in the paper is no longer valid) specifically recognizes eukaryotic N-terminal sorting sequences. The data set is from TargetP. The authors tested a large number of physicochemical features of N-terminal parts of proteins with signal or transit peptides and obtained a combination of simple rules that yielded a discriminative performance fairly close to that of TargetP. Interestingly, a simple hydrophobicity scale even outperformed the NN-based TargetP on plant SPs.

LOCtree (http://cubic.bioc.columbia.edu/services/loctree/)[126] is a hierarchical system combining several SVMs. The hierarchy is designed to mimic the sorting process in the cell, with the binary SVMs organized in the same order as the sorting decisions. It comes in three versions: plant (six localizations: secreted, organelles of the secretory pathway, nuclear, cytoplasmic, mitochondrion and chloroplast), eukaryotic non-plant (five localizations) and bacteria (three localizations: secreted, periplasm and cytoplasm). Transmembrane proteins are not considered. Data were derived from SWISS-PROT and homology reduced to a threshold below which homology-based annotations of subcellular localization were found to be unreliable[52]. In addition to the "subcellular location" comments, other keywords in SWISS-PROT were used for annotating localization through the LOCkey algorithm[46], thereby enlarging the data set considerably. Each input sequence is used to search the UniProt database[80], creating a profile of homologous sequences. The amino-acid composition of the profile is then used as input to the SVMs, for the whole sequence, for the N-terminal 50 residues and for three predicted secondary structure states separately. In addition, the SignalP output is used as an input to the SVM at the top node for the eukaryotic versions. Note: LOCtree can have very long response times.

BaCelLo (http://gpcr.biocomp.unibo.it/bacello/)[76] is a recently published eukaryotic predictor with very promising performance values. It is based on a number of SVMs organized in a decision tree, very similar to that of LOCtree. Sequences are represented in a very simple way, using just the amino-acid composition of the raw sequence and of sequence profiles found with BLAST, both for the whole sequence and for N-terminal and C-terminal regions separately. There are separate predictors for animals, plants and fungi. In plants, there are five localizations: secretory pathway, cytoplasm, nucleus, mitochondrion and chloroplast; in animals and fungi, there are four. Transmembrane proteins are not considered. Data are derived from SWISS-PROT and homology reduced to 30% identity.

Protein Prowler v. 1.2 (http://pprowler.imb.uq.edu.au/)[127] comes in a Plant and a Non-plant version, and is based on the same data sets as TargetP. It predicts eukaryotic proteins to the same three or four localizations, through a combination of different types of NNs (feed-forward, recurrent[128] and sequential-cascaded[129]), which are used to generate residue-wise scores, and SVMs, which are used to generate a prediction as to the presence or absence of a presequence. The performance is reported to be better than TargetP. An accompanying method on the same server, PTS1Prowler, predicts peroxisomal C-terminal targeting signals[130].

CELLO v.2.5 (http://cello.life.nctu.edu.tw/)[54] is an SVM-based predictor for both eukaryotes (12 localizations), Gram-negative bacteria (five localizations) and Gram-positive bacteria (four localizations). The SVMs are organized in a two-level system, where the first level contains a number of SVMs trained on various sequence encodings, and the second layer is a "jury SVM", which decides on the prediction based on the outputs of the first-layer SVMs. The sequences are encoded by total amino-acid composition, dipeptide composition and amino-acid composition (in some cases with a reduced alphabet) in a number of partitions of each sequence. The data sets are those of PSORTb 2.0 (ref. 124) (which has not been homology reduced) for bacteria and PLOC[72] (where only sequences above 80% identity have been removed) for eukaryotes. The authors report prediction accuracy for a number of different sequence identity values, and they found that above 30% identity, a global alignment was better than the SVM-based prediction, whereas below 30% the prediction was best. Performance improvements relative to PLOC and PSORTb 2.0 were reported, but the homology in the data sets makes it problematic to compare performances.

PA-SUB v2.5 (http://www.cs.ualberta.ca/~bioinfo/PA/Sub/)[131] is a homology-based subcellular localization predictor, available in five versions: animal (nine localizations), plant (ten localizations), fungi (nine localizations), Gram-negative bacteria (six localizations) and Gram-positive bacteria (four localizations). Like WoLF PSORT, PA-SUB has a certain emphasis on trying to explain what features were important for each protein localization prediction. For each query sequence, BLAST is used to find its SWISS-PROT homologs. The presence/absence pattern of a number of annotation key words or phrases among these homologs is logged, and from this pattern the subcellular localization is inferred. The pattern to localization mapping is learned using Naïve Bayes (NB) classifiers. A crucial advantage of NB classifiers is the possibility to determine the importance of every feature for a particular prediction. Other machine-learning techniques were tried by the authors, and NNs and

BOX 1 | CONTINUED

SVMs were actually found to yield better performance, but they were abandoned as the authors decided that the ability of NB classifiers to explain a prediction outweighed the performance increase recorded for NN/SVM. Performance is difficult to compare with most other methods, which do not use homology searches.

MultiLoc/TargetLoc (http://www-bs.informatik.uni-tuebingen.de/Services/MultiLoc/)[132]; TargetLoc is another predictor based on the TargetP data sets, and it is also present in Plant and Non-plant versions. The authors have built a two-layer system of SVMs that take into account the N-terminal part of the sequence (looking for the presence of N-terminal sorting signals), the overall amino-acid composition and the occurrence of certain motifs from PROSITE[101] and NLSdb[103] in the sequence, creating an intermediate "protein profile vector". The protein profile vector is then fed into the top layer SVMs making the localization prediction. The performance as measured on the TargetP sets is reported as better than TargetP, but seems to be slightly lower than Protein Prowler (above). The extended version, MultiLoc, predicts up to ten eukaryotic localizations, but it is trained on sequences allowing up to 80% sequence identity, which we believe is a too permissive threshold; so the reported overall accuracy of approximately 75% (considerably better than PSORT II) is probably an overestimate.

ANTICIPATED RESULTS

Measuring the performance of a prediction method is not trivial. Simply counting the percentage of correct answers can be misleading if the sizes of the classes in the data set differ greatly (e.g., if there are 99 negative examples for each positive example, simply predicting everything as negative will give 99% correct answers). Two of the most often cited prediction measures are sensitivity and specificity, which are calculated for each class separately. Sensitivity expresses how many of the positive examples (e.g., SPs) are found (predicted correctly). Specificity is slightly more complicated, as there are two competing definitions: (i) the proportion of the positive predictions that are true; or (ii) the proportion of the negative examples that are predicted correctly. Here, we use definition (i). Instead of definition (ii), we prefer to report the false positive rate, that is, how many positive predictions are there on negative examples. See Baldi *et al.*[114] for an extensive discussion of performance measures.

There is always a trade-off between sensitivity and specificity; if you want a higher sensitivity, you must also be prepared for a higher number of false positives. The balance between the two kinds of errors, false negatives and false positives, can often be adjusted by changing the threshold (also known as cutoff) for positive prediction.

In general, we have decided not to include performance data for all the prediction methods we describe. Performance measures for multicategory methods are difficult to compare, partly because the data sets, as described, are often not properly homology reduced, and partly because the number of predicted classes differ between the methods. Discriminating between a higher number of location classes will always be more difficult and result in a lower overall percentage of correct predictions. Therefore, we will focus on the performance of the two central tools of this protocol, TargetP and SignalP, in this section. But first a couple of general remarks about multicategory methods.

Note that membrane proteins are often ignored by the multicategory methods, with the notable exception of PSORT II and WoLF PSORT. In general, it is fairly easy to predict whether a protein is transmembrane or not (see Steps 11 and 12 in PROCEDURE), but it is far more difficult to predict exactly which membrane in a eukaryotic cell a given transmembrane protein is integral to. For example, as described in Steps 14 and 16, the sorting signals for membrane proteins of the ER and peroxisomes are less well characterized and harder to predict than those of their lumenal counterparts. However, the membrane protein-interested user is not totally lost. For Golgi type II membrane proteins, there is a predictor available; see Step 14. Proteins of chloroplast thylakoid membranes and mitochondrial inner membranes are also fairly unproblematic; they can be expected to have a transit peptide followed by either a signal anchor or an SP plus a stop-transfer sequence, and should therefore be predictable by combining TargetP with SignalP, TMHMM and/or Phobius, although this scheme has never been rigorously tested. Outer membrane proteins of chloroplasts and mitochondria do not have transit peptides, but most of them can be recognized because they are of the β-barrel type; see Step 12 (how to distinguish between these two membranes in a plant cell is quite another matter, we are not aware of any predictor capable of this).

Another class of proteins that are often ignored are those found in more than one compartment. Most multicategory methods simply exclude those proteins when constructing their data sets (WoLF PSORT is again an exception, as mentioned). But also in this case, combining predictors can sometimes help. NucPred and PredictNLS (see Step 15) recognize proteins with NLSs, and this includes both those proteins that stay in the nucleus and those that are re-exported to the cytoplasm. When combining these two predictors with NetNES, you get a prediction covering "cytoplasmic and nuclear", probably the most important dual location. The methods ISort[115] and PLPD[116] both predict as many as 22 different subcellular locations determined in a yeast high-throughput experiment, taking multiple locations into account. ISort uses pseudo-amino-acid composition, matches to a database of functional domains and GO-numbers (see "Prediction of sorting signals"), whereas PLPD reports a higher performance without using GO-numbers. However, their results are limited to yeast, and neither is implemented as an available web server.

Recently, a review compared the performance of five multicategory methods, which were all capable of predicting at least nine locations, on two data sets of mouse proteins[117]. The selected methods were CELLO, MultiLoc, Proteome Analyst (PA-SUB) and WoLF PSORT (see **Box 1**) plus pTarget[118] (which are all trained without sufficient homology reduction). The conclusion was that "No individual method had a sufficient level of sensitivity across both evaluation sets that would enable reliable application to hypothetical proteins"—an important point to keep in mind when using multicategory subcellular localization methods.

TargetP performance

The outcome from TargetP is a prediction for each submitted protein of whether an N-terminal presequence (SP, mTP or cTP, for Plant version) is present. The predictive performance varies between the three different signals it handles; see **Table 2**. Although the exact figures depend on what test set is used for the evaluation, it is clear that the prediction of SPs is the most accurate. It is also the most conserved and most studied of the three signals. Both sensitivity and specificity of SP prediction have been measured to 90–95% for Plant and Non-plant TargetP versions. For cTPs and mTPs, sensitivities are at 80–90% and specificities at approximately 70% (except for plant mTPs where specificity is 90%). This means that there are more false positive predictions of cTPs and mTPs than of SPs.

The "reliability class" (RC) value associated with each prediction signifies how confident TargetP is in each prediction. The RC is based on the difference between the highest and the second highest TargetP output scores: if this difference is larger than 0.8, then RC = 1; if it is between 0.6 and 0.8, then RC = 2, and so on. An RC value of 1 means that the prediction belongs to the most reliable class of predictions, and 5 that it belongs to the least reliable class. **Table 2** shows that, as expected, RC = 1 gives the best performance in terms of the specificity, which is 100% for all tested sets (except plant SP where it is 99%). The specificity then drops as the RC increases, but remains above what would be expected by random chance even for RC = 5 for all sets.

If opted for, the results also include prediction of cleavage sites for the predicted presequences. The cleavage site prediction uses SignalP and ChloroP for SP and cTP cleavage site predictions, respectively. The mTP cleavage site prediction is a TargetP-unique feature. The cleavage site prediction is less accurate than the corresponding presequence presence prediction. The SP cleavage site prediction is the most accurate whereas in particular the cTP cleavage site predictions should be interpreted cautiously (only 44.8% correct predictions, within two residues). The mTP cleavage site results are in-between. These differences reflect the varying degrees of conservation around the annotated cleavage sites for the three presequences.

SignalP performance

Before discussing the reliability of SignalP, one remark about the choice of organism group: note that there is no SignalP version for Archaea. There are simply too few experimentally verified secretory proteins available for Archaea to train a specific version. This is a problem that SignalP shares with other prokaryotic subcellular localization predictors such as PSORTb (see Step 18). In a preliminary analysis of predicted SPs in the genome of the archaeon *Methanococcus jannaschii*[119], we found that they had eukaryotic-looking cleavage sites, a bacterial-looking charge distribution and a unique composition of

TABLE 2 | TargetP performance values[a].

		Plant		Non-plant		
Category	**RC**	**Cumul. sens. (%)**	**Spec. (%)**	**Cumul. sens. (%)**	**Spec. (%)**	**CS loc. % (RC not recorded)**
cTP	1	24	100	—	—	44.8
	2	50	95	—	—	
	3	62	81	—	—	
	4	79	80	—	—	
	5	85	42	—	—	
mTP	1	22	100	28	97	80.8
	2	44	94	59	95	
	3	59	82	75	82	
	4	66	56	84	71	
	5	78	62	89	53	
SP	1	66	99	62	100	74.9
	2	81	100	85	94	
	3	88	100	90	87	
	4	92	67	94	88	
	5	94	33	95	55	

[a]Performance of TargetP 1.1, Plant and Non-Plant versions, for the various presequence categories. Discrimination performance is given by cumulative sensitivity ("cumul. sens.") and specificity ("spec."). Cumulative sensitivity is measured as the percentage of evaluated proteins predicted to the particular reliability class (RC) or better. The specificity is measured as the percentage of correctly predicted proteins within the particular RC, and is the best measure of how well TargetP is able to predict proteins within a certain RC. The localization prediction performance is measured on plant and non-plant protein sets of 555 and 1,110 proteins, respectively[64]. The cleavage site ("CS") location performance is given as the percentage of positive predictions that have a correctly located cleavage site ("CS loc."), except for cTP CS prediction where the predicted CS is allowed to deviate with up to two residues from the annotated cleavage site (see ChloroP paper for a discussion of this[26]). SP CS prediction in TargetP is performed by an early version of SignalP[60]. There is no RC value associated with a CS prediction.

the hydrophobic region. Therefore, the current recommendation for archaeal sequences must be to try all three versions and use a consensus of the three predictions.

SignalP does not have reliability classes like TargetP, but it is still easy to see whether the prediction is more or less reliable. Simply compare the D-score with the D-score cutoff: if the output value is barely above the cutoff, the prediction is less reliable than if it is far above the cutoff. In **Table 3**, we have shown the sensitivity and false positive rate (measured on nuclear and cytoplasmic proteins) for varying values of the D-score cutoff. If you prefer a different trade-off between sensitivity and specificity than what is given by the default cutoff, simply look up your desired sensitivity or false positive rate in the table and find the corresponding cutoff value.

The cleavage site location performance is also given in **Table 3**. This is calculated as the percentage of correct cleavage site predictions among predictions with a D-score above the chosen cutoff. In other words, this number expresses how much you can trust your cleavage site prediction given the observed D-score.

It may be tempting to assess reliability by looking at how many of the five scores reported by SignalP-NN are above their default cutoff values (counting the number of "YES" answers). Of course, requiring all answers to be "YES" will yield a better specificity at the cost of reduced sensitivity. However, as the D-score provides the best discrimination, you actually get a better performance (better sensitivity and cleavage site location for a given false positive rate) by using a higher D-score cutoff than by looking at the other scores.

HMM performance is also given in **Table 3**. For eukaryotes, SignalP-HMM is clearly not as good as SignalP-NN, having many more false positives. For bacteria, SignalP-HMM is slightly better at discriminating SPs from cytoplasmic sequences, but worse at locating cleavage sites precisely. It is a good idea to check whether the NN and HMM predictions agree, especially if you want reliable cleavage site predictions. The row in **Table 3** labeled "NN + HMM" reports the performance if you accept only those predictions where SignalP-NN and SignalP-HMM are both positive and predict the same cleavage site. As you can see, sensitivity

TABLE 3 | SignalP performance values[a].

Method	Sens. (%)	FP rate (%)	CS loc. (%)
Eukaryotes			
NN, D > 0.30	99.4	3.1	76.3
NN, D > 0.40	98.8	1.4	78.0
NN, D > 0.43*	98.6	1.2	78.4
NN, D > 0.50	98.2	0.8	78.9
NN, D > 0.60	95.1	0.4	79.3
NN, D > 0.70	85.1	0.2	81.0
NN, D > 0.80	57.8	0.0	85.8
HMM	98.4	3.4	75.8
NN+HMM	85.0	0.4	84.7
Gram-negative bacteria			
NN, D > 0.30	99.2	8.7	85.6
NN, D > 0.40	97.8	3.9	89.9
NN, D > 0.44*	97.0	2.1	91.3
NN, D > 0.50	95.1	1.8	92.2
NN, D > 0.60	91.4	0.6	93.5
NN, D > 0.70	83.0	0.0	95.1
NN, D > 0.80	64.6	0.0	95.4
HMM	98.4	2.4	87.4
NN + HMM	91.1	0.6	94.1
Gram-positive bacteria			
NN, D > 0.30	98.8	6.0	82.2
NN, D > 0.40	97.6	0.0	86.5
NN, D > 0.45*	97.6	0.0	86.5
NN, D > 0.50	96.4	0.0	86.3
NN, D > 0.60	92.8	0.0	87.1
NN, D > 0.70	75.4	0.0	89.7
NN, D > 0.80	53.3	0.0	91.0
HMM	98.2	0.7	82.4
NN + HMM	85.6	0.0	90.9

[a]Performance of SignalP 3.0. Discrimination performance is given by sensitivity ("sens.") and false positive rate ("FP rate"). FP rate is measured on cytoplasmic and (for eukaryotes) nuclear proteins. Cleavage site location performance is given as the percentage of positive predictions that have a correctly located cleavage site ("CS loc."). Results are reported for SignalP-NN with varying values of D-score cutoff (where the default cutoff is marked by an "*"), for SignalP-HMM and for a combination of these (counting only predictions where both methods agree on classification and cleavage site).

drops considerably, but false positive rate is low and cleavage site precision is good (better than what you can get by increasing D-score cutoff).

In two comparisons of SP prediction methods, SignalP has been favorably evaluated. In 2000, Menne *et al.*[120] compared SignalP version 1.1 and 2.0 to SPScan and SigCleave, which as mentioned in INTRODUCTION are both based on the older weight matrix method[55]. Both bacterial and eukaryotic sequences were used. They found SignalP to be better than the weight matrix-based methods, and SignalP 2.0 to be better than SignalP 1.1. Within SignalP 2.0, the NN part was found to predict cleavage sites more precisely, whereas the HMM part had fewer false positives. However, they found much higher false-positive rates (up to 18.3%) than we had measured for SignalP, which is probably due to their negative set containing some membrane proteins with a transmembrane region within the 60 N-terminal residues.

In 2005, Klee and Ellis[121] compared SignalP 3.0, SignalP 2.0 and TargetP to PrediSi[122], Phobius[113] and ProtComp 6.0 (a commercially available program from Softberry Inc.) on a set of exclusively mammalian proteins. For the discrimination between secretory and non-secretory proteins (without transmembrane proteins), they found SignalP 3.0 D-score to be the most accurate single score. Furthermore, they found that the predictive accuracy could be substantially improved by combining scores from multiple methods into a single composite prediction. The best combinations turned out to consist of TargetP prediction, SignalP 2.0 maximum Y-score, SignalP 3.0 maximum S-score and a choice of other SignalP scores as the fourth component. Interestingly, SignalP 3.0 D-score was not included here.

Concluding remarks

Although there has recently been much progress in the field of experimental high-throughput determination of subcellular localization of proteins, we believe that reliable localization prediction is still an invaluable (and inexpensive) tool to get a quick and often correct suggestion for the localization of a protein.

No doubt, the subcellular localization tools described in this article will be further developed and refined as the size and accuracy of the protein annotation databases increase, and as the knowledge of protein sorting mechanisms expands.

The servers at the CBS web pages are currently being changed such that they also function as proper SOAP (simple object access protocol)-based web services (implemented by the Web Services Description Language). The SOAP server processes requests and returns data in an XML-formatted SOAP packet using an agreed XML schema format. This means that other computational methods can interact directly with the methods without human intervention. As these standards and schemas change quite rapidly, we refer to the CBS web pages (http://www.cbs.dtu.dk/ws/) for additional details and help on the current setup.

COMPETING INTERESTS STATEMENT The authors declare competing financial interests (see the HTML version of this article for details).

Published online at http://www.natureprotocols.com
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Burns, N. *et al.* Large-scale analysis of gene expression, protein localization, and gene disruption in *Saccharomyces cerevisiae*. *Genes Dev.* **8**, 1087–1105 (1994).
2. Chalfie, M., Tu, Y., Euskirchen, G., Ward, W.W. & Prasher, D.C. Green fluorescent protein as a marker for gene expression. *Science* **263**, 802–805 (1994).
3. Sawin, K.E. & Nurse, P. Identification of fission yeast nuclear markers using random polypeptide fusions with green fluorescent protein. *Proc. Natl. Acad. Sci. USA* **93**, 15146–15151 (1996).
4. Kumar, A. *et al.* Subcellular localization of the yeast proteome. *Genes Dev.* **16**, 707–719 (2002).
5. Simpson, J.C., Wellenreuther, R., Poustka, A., Pepperkok, R. & Wiemann, S. Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. *EMBO Rep.* **1**, 287–292 (2000).
6. Shevchenko, A. *et al.* Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. *Proc. Natl. Acad. Sci. USA* **93**, 14440–14445 (1996).
7. Peltier, J.-B. *et al.* Proteomics of the chloroplast: systematic identification and targeting analysis of lumenal and peripheral thylakoid proteins. *Plant Cell* **12**, 319–341 (2000).
8. Yates, J.R., Gilchrist, A., Howell, K.E. & Bergeron, J.J. Proteomics of organelles and large cellular structures. *Nat. Rev. Mol. Cell Biol.* **6**, 702–714 (2005).
9. Andersen, J.S. *et al.* Directed proteomic analysis of the human nucleolus. *Curr. Biol.* **12**, 1–11 (2002).
10. Andersen, J.S. *et al.* Proteomic characterization of the human centrosome by protein correlation profiling. *Nature* **426**, 570–574 (2003).
11. Foster, L.J. *et al.* A mammalian organelle map by protein correlation profiling. *Cell* **125**, 187–199 (2006).
12. Andersen, J.S. *et al.* Nucleolar proteome dynamics. *Nature* **433**, 77–83 (2005).
13. Agaton, C. *et al.* Affinity proteomics for systematic protein profiling of chromosome 21 gene products in human tissues. *Mol. Cell. Proteomics* **2**, 405–414 (2003).
14. Uhlen, M. *et al.* A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol. Cell. Proteomics* **4**, 1920–1932 (2005).
15. Hinsby, A.M. *et al.* A wiring of the human nucleolus. *Mol. Cell* **22**, 285–295 (2006).
16. von Heijne, G. The signal peptide. *J. Membr. Biol.* **115**, 195–201 (1990).
17. Pugsley, A.P., Francetic, O., Possot, O.M., Sauvonnet, N. & Hardie, K.R. Recent progress and future directions in studies of the main terminal branch of the general secretory pathway in Gram-negative bacteria—a review. *Gene* **192**, 13–19 (1997).
18. van Vliet, C., Thomas, E.C., Merino-Trigo, A., Teasdale, R.D. & Gleeson, P.A. Intracellular sorting and transport of proteins. *Prog. Biophys. Mol. Biol.* **83**, 1–45 (2003).
19. Bendtsen, J.D., Jensen, L.J., Blom, N., von Heijne, G. & Brunak, S. Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng. Des. Sel.* **17**, 349356 (2004).
20. Binnewies, T.T. *et al.* Genome update: protein secretion systems in 225 bacterial genomes. *Microbiology* **151**, 1013–1016 (2005).
21. Henderson, I.R., Navarro-Garcia, F., Desvaux, M., Fernandez, R.C. & Ala'Aldeen, D. Type V protein secretion pathway: the autotransporter story. *Microbiol. Mol. Biol. Rev.* **68**, 692–744 (2004).
22. Ghosh, P. Process of protein transport by the type III secretion system. *Microbiol. Mol. Biol. Rev.* **68**, 771–795 (2004).
23. Bendtsen, J.D., Kiemer, L., Fausbøll, A. & Brunak, S. Non-classical protein secretion in bacteria. *BMC Microbiol.* **5**, 58 (2005).
24. Schatz, G. & Dobberstein, B. Common principles of protein translocation across membranes. *Science* **271**, 1519–1526 (1996).
25. von Heijne, G., Steppuhn, J. & Hermann, S.G. Domain structure of mitochondrial and chloroplast targeting peptides. *Eur. J. Biochem.* **180**, 535–545 (1989).
26. Emanuelsson, O., Nielsen, H. & von Heijne, G. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci.* **8**, 978–984 (1999).

27. Bruce, B.D. Chloroplast transit peptides: structure, function and evolution. *Trends Cell Biol.* **10**, 440–447 (2000).
28. Emanuelsson, O., von Heijne, G. & Schneider, G. Analysis and prediction of mitochondrial targeting peptides. *Methods Cell Biol.* **65**, 175–187 (2001).
29. Schneider, G. *et al.* Feature-extraction from endopeptidase cleavage sites in mitochondrial targeting peptides. *Proteins* **30**, 49–60 (1998).
30. Kalousek, F., Hendrick, J.P. & Rosenberg, L.E. Two mitochondrial matrix proteases act sequentially in the processing of mammalian matrix enzymes. *Proc. Natl. Acad. Sci. USA* **85**, 7536–7540 (1988).
31. Isaya, G. & Kalousek, F. Mitochondrial intermediate peptidase. in *Signal Peptidases* (ed. von Heijne, G.) 87–103 (R.G. Landes Company, Austin, 1994).
32. Abe, Y. *et al.* Structural basis of presequence recognition by the mitochondrial protein import receptor Tom20. *Cell* **100**, 551–560 (2000).
33. Taylor, A.B. *et al.* Crystal structures of mitochondrial processing peptidase reveal the mode for specific cleavage of import signal sequences. *Structure* **9**, 615–625 (2001).
34. Bonen, L. & Doolittle, W.F. On the prokaryotic nature of red algal chloroplasts. *Proc. Natl. Acad. Sci. USA* **72**, 2310–2314 (1975).
35. Moreira, D., Guyader, H.L. & Philippe, H. The origin of red algae and the evolution of chloroplasts. *Nature* **405**, 69–72 (2000).
36. Robinson, C., Hynds, P.J., Robinson, D. & Mant, A. Multiple pathways for the targeting of thylakoid proteins in chloroplasts. *Plant Mol. Biol.* **38**, 209–221 (1998).
37. Shackleton, J.B. & Robinson, C. 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.* **266**, 12152–12156 (1991).
38. Robinson, C. & Bolhuis, A. Protein targeting by the twin-arginine translocation pathway. *Nat. Rev. Mol. Cell. Biol.* **2**, 350–356 (2001).
39. Chabregas, S.M. *et al.* Dual targeting properties of the N-terminal signal sequence of *Arabidopsis thaliana* THI1 protein to mitochondria and chloroplasts. *Plant Mol. Biol.* **46**, 639–650 (2001).
40. Small, I., Wintz, H., Akashi, K. & Mireau, H. Two birds with one stone: genes that encode products targeted to two or more compartments. *Plant Mol. Biol.* **38**, 265–277 (1998).
41. Zhang, X.P. & Glaser, E. Interaction of plant mitochondrial and chloroplast signal peptides with the Hsp70 molecular chaperone. *Trends Plant Sci.* **7**, 14–21 (2002).
42. Kleffmann, T. *et al.* The *Arabidopsis thaliana* chloroplast proteome reveals pathway abundance and novel protein functions. *Curr. Biol.* **14**, 354–362 (2004).
43. Villarejo, A. *et al.* Evidence for a protein transported through the secretory pathway en route to the higher plant chloroplast. *Nat. Cell Biol.* **7**, 1224–1231 (2006).
44. 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.* **301**, 1059–1075 (2000).
45. Marcotte, E.M., Xenarios, I., van Der Bliek, A.M. & Eisenberg, D. Localizing proteins in the cell from their phylogenetic profiles. *Proc. Natl. Acad. Sci. USA* **97**, 12115–12120 (2000).
46. Nair, R. & Rost, B. Inferring sub-cellular localization through automated lexical analysis. *Bioinformatics* **18**, S78–S86 (2002).
47. Chou, K.C. & Shen, H.B. Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. *Biochem. Biophys. Res. Commun.* **347**, 150–157 (2006).
48. Chou, K.C. & Shen, H.B. Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic *k*-nearest neighbor classifiers. *J. Proteome Res.* **5**, 1888–1897 (2006).
49. Chou, K.C. & Shen, H.B. Large-scale plant protein subcellular location prediction. *J. Cell. Biochem.* **100**, 665–678 (2007).
50. Mott, R., Schultz, J., Bork, P. & Ponting, C.P. Predicting protein cellular localization using a domain projection method. *Genome Res.* **12**, 1168–1174 (2002).
51. Scott, M., Thomas, D. & Hallett, M. Predicting subcellular localization via protein motif co-occurrence. *Genome Res.* **14**, 1957–1966 (2004).
52. Nair, R. & Rost, B. Sequence conserved for subcellular localization. *Protein Sci.* **11**, 2836–2847 (2002).
53. McGinnis, S. & Madden, T.L. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* **32**, W20–W25 (2004).
54. Yu, C.S., Chen, Y.C., Lu, C.H. & Hwang, J.K. Prediction of protein subcellular localization. *Proteins* **64**, 643–651 (2006).
55. von Heijne, G. A new method for predicting signal sequence cleavage sites. *Nucleic Acids Res.* **14**, 4683–4690 (1986).
56. Nakai, K. & Kanehisa, M. A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics* **14**, 897–911 (1992).
57. Baldi, P. & Brunak, S. *Bioinformatics: The Machine Learning Approach* (MIT Press, Cambridge, MA, USA, 1998).
58. Durbin, R.M., Eddy, S.R., Krogh, A. & Mitchison, G. *Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids* (Cambridge University Press, Cambridge, U.K. 1998).
59. Vapnik, V. *The Nature of Statistical Learning Theory* (Springer, NY, USA, 1995).
60. Nielsen, H., Brunak, S., Engelbrecht, J. & von Heijne, G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10**, 1–6 (1997).
61. Nielsen, H. & Krogh, A. 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* (eds. Glasgow, J. *et al.*) 122–130 (AAAI Press, Menlo Park, CA, USA, 1998).
62. Bendtsen, J.D., Nielsen, H., von Heijne, G. & Brunak, S. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* **340**, 783–795 (2004).
63. Claros, M.G. & Vincens, P. Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur. J. Biochem.* **241**, 779–786 (1996).
64. 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.* **300**, 1005–1016 (2000).
65. Andrade, M.A., O'Donoghue, S.I. & Rost, B. Adaptation of protein surfaces to subcellular location. *J. Mol. Biol.* **276**, 517–528 (1998).
66. Nakashima, H. & Nishikawa, K. Discrimination of intracellular an extracellular proteins using amino acid composition and residue-pair frequencies. *J. Mol. Biol.* **238**, 54–61 (1994).
67. Reinhardt, A. & Hubbard, T. Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res.* **26**, 2230–2236 (1998).
68. Cedano, J., Aloy, P., Pérez-Pons, J.A. & Querol, E. Relation between amino acid composition and cellular location of proteins. *J. Mol. Biol.* **266**, 594–600 (1997).
69. Chou, K.-C. & Elrod, D.W. Using discriminant function for prediction of subcellular location of prokaryotic proteins. *Biochem. Biophys. Res. Commun.* **252**, 63–68 (1998).
70. Chou, K.-C. & Elrod, D.W. Protein subcellular location prediction. *Protein Eng.* **12**, 107–118 (1999).
71. Hua, S. & Sun, Z. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* **17**, 721–728 (2001).
72. Park, K.-J. & Kanehisa, M. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics* **19**, 1656–1663 (2003).
73. Bhasin, M. & Raghava, G.P.S. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res.* **32**, W414–W419 (2004).
74. Chou, K.C. Prediction of protein cellular attributes using pseudo amino acid composition. *Proteins* **43**, 246–255 (2001).
75. Cui, Q., Jiang, T., Liu, B. & Ma, S. Esub8: a novel tool to predict protein subcellular localizations in eukaryotic organisms. *BMC Bioinformatics* **5**, 66 (2004).
76. Pierleoni, A., Martelli, P.L., Fariselli, P. & Casadio, R. BaCelLo: a balanced subcellular localization predictor. *Bioinformatics* **22**, e408–e416 (2006).
77. Linding, R. *et al.* Protein disorder prediction: implications for structural proteomics. *Structure* **11**, 1453–1459 (2003).
78. Nakai, K. & Kanehisa, M. Expert system for predicting protein localization sites in Gram-negative bacteria. *Proteins* **11**, 95–110 (1991).
79. Horton, P. & Nakai, K. 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* (eds. Gaasterland, T. *et al.* 147–152 (AAAI Press, Menlo Park, CA, USA, 1997).
80. Bairoch, A. *et al.* The universal protein resource (UniProt). *Nucleic Acids Res.* **33**, D154–D159 (2005).
81. Brunak, S. Doing sequence analysis by inspecting the order in which neural networks learn. in *Computation of Biomolecular Structures—Achievements, Problems and Perspectives* (eds. Soumpasis, D.M. & Jovin, T.M.) 43–54 (Springer-Verlag, Berlin, 1993).
82. Hobohm, U., Scharf, M., Schneider, R. & Sander, C. Selection of representative protein data sets. *Protein Sci.* **1**, 409–417 (1992).
83. Xie, D., Li, A., Wang, M., Fan, Z. & Feng, H. LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. *Nucleic Acids Res.* **33**, W105–W110 (2005).
84. von Heijne, G. Transcending the impenetrable: how proteins come to terms with membranes. *Biochim. Biophys. Acta* **947**, 307–333 (1988).

85. Peltier, J.-B. *et al.* Central functions of the lumenal and peripheral thylakoid proteome of *Arabidopsis* determined by experimentation and genome-wide prediction. *Plant Cell* **14**, 211–236 (2002).
86. Berks, B.C. A common export pathway for proteins binding complex redox cofactors? *Mol. Microbiol.* **22**, 393–404 (1996).
87. Cristóbal, S., de Gier, J.-W., Nielsen, H. & von Heijne, G. Competition between Sec- and TAT-dependent protein translocation in *Escherichia coli*. *EMBO J.* **18**, 2982–2990 (1999).
88. Bendtsen, J.D., Nielsen, H., Widdick, D., Palmer, T. & Brunak, S. Prediction of twin-arginine signal peptides. *BMC Bioinformatics* **6**, 167 (2005).
89. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E.L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001).
90. Möller, S., Croning, M.D. & Apweiler, R. Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* **17**, 646–653 (2001).
91. Chen, C.P., Kernytsky, A. & Rost, B. Transmembrane helix predictions revisited. *Protein Sci.* **11**, 2774–2791 (2002).
92. Cuthbertson, J.M., Doyle, D.A. & Sansom, M.S. Transmembrane helix prediction: a comparative evaluation and analysis. *Protein Eng. Des. Sel.* **18**, 295–308 (2005).
93. Sadovskaya, N.S., Sutormin, R.A. & Gelfand, M.S. Recognition of transmembrane segments in proteins: review and consistency-based benchmarking of internet servers. *J. Bioinform. Comput. Biol.* **4**, 1033–1056 (2006).
94. Schulz, G. β-barrel membrane proteins. *Curr. Opin. Struct. Biol.* **10**, 443–447 (2000).
95. Jacoboni, I., Martelli, P.L., Fariselli, P., de Pinto, V. & Casadio, R. Prediction of the transmembrane regions of β-barrel membrane proteins with a neural network-based predictor. *Protein Sci.* **10**, 779–787 (2001).
96. Martelli, P.L., Fariselli, P., Krogh, A. & Casadio, R. A sequence-profile-based HMM for predicting and discriminating β-barrel membrane proteins. *Bioinformatics* **18**, S46–S53 (2002).
97. Eisenhaber, F. *et al.* Prediction of lipid posttranslational modifications and localization signals from protein sequences: big-Π, NMT and PTS1. *Nucleic Acids Res.* **31**, 3631–3634 (2003).
98. Bologna, G., Yvon, C., Duvaud, S. & Veuthey, A.-L. N-Terminal myristoylation predictions by ensembles of neural networks. *Proteomics* **4**, 1626–1632 (2004).
99. Juncker, A.S. *et al.* Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci.* **12**, 1652–1662 (2003).
100. von Heijne, G. The structure of signal peptides from bacterial lipoproteins. *Protein Eng.* **2**, 531–534 (1989).
101. Hulo, N. *et al.* Recent improvements to the PROSITE database. *Nucleic Acids Res.* **32**, D134–D137 (2004).
102. Yuan, Z. & Teasdale, R.D. Prediction of Golgi type II membrane proteins based on their transmembrane domains. *Bioinformatics* **18**, 1109–1115 (2002).
103. Cokol, M., Nair, R. & Rost, B. Finding nuclear localization signals. *EMBO Rep.* **1**, 411–415 (2000).
104. Heddad, A., Brameier, M. & MacCallum, R.M. Evolving regular expression-based sequence classifiers for protein nuclear localisation. in *Applications of Evolutionary Computing, EvoWorkshops2004: EvoBIO, EvoCOMNET, EvoHOT, EvoIASP, EvoMUSART, EvoSTOC, vol. 3005 of LNCS* (eds. Raidl, G.R. *et al.*) 31–40 (Springer-Verlag, Berlin, Germany, 2004).
105. Zhao, L.-J. & Padmanabhan, R. Nuclear transport of adenovirus DNA polymerase is facilitated by interaction with preterminal protein. *Cell* **55**, 1005–1015 (1988).
106. Pemberton, L.F. & Paschal, B.M. Mechanisms of receptor-mediated nuclear import and nuclear export. *Traffic* **6**, 187–198 (2005).
107. la Cour, T. *et al.* Analysis and prediction of leucine-rich nuclear export signals. *Protein Eng. Des. Sel.* **17**, 527–536 (2004).
108. Olivier, L.M. & Krisans, S.K. Peroxisomal protein targeting and identification of peroxisomal targeting signals in cholesterol biosynthetic enzymes. *Biochim. Biophys. Acta* **1529**, 89–102 (2000).
109. Emanuelsson, O., Elofsson, A., von Heijne, G. & Cristóbal, S. *In silico* prediction of the peroxisomal proteome in fungi, plants and animals. *J. Mol. Biol.* **330**, 443–456 (2003).
110. Neuberger, G., Maurer-Stroh, S., Eisenhaber, B., Hartig, A. & Eisenhaber, F. Prediction of peroxisomal targeting signal 1 containing proteins from amino acid sequence. *J. Mol. Biol.* **328**, 581–592 (2003).
111. Pedersen, A.G. & Nielsen, H. Neural network prediction of translation initiation sites in eukaryotes: perspectives for EST and genome analysis. in *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology* Gaasterland, T. *et al.* (eds.) 226–233 (AAAI Press, Menlo Park, CA, USA, 1997).
112. Duckert, P., Brunak, S. & Blom, N. Prediction of proprotein convertase cleavage sites. *Protein Eng. Des. Sel.* **17**, 107–12 (2004).
113. Käll, L., Krogh, A. & Sonnhammer, E.L.L. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **338**, 027–1036 (2004).
114. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A. & Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **16**, 412–424 (2000).
115. Chou, K.-C. & Cai, Y.-D. Predicting protein localization in budding yeast. *Bioinformatics* **21**, 944–950 (2005).
116. Lee, K., Kim, D.-W., Na, D., Lee, K.H. & Lee, D. PLPD: reliable protein localization prediction from imbalanced and overlapped datasets. *Nucleic Acids Res.* **34**, 4655–4666 (2006).
117. Sprenger, J., Fink, J.L. & Teasdale, R.D. Evaluation and comparison of mammalian subcellular localization prediction methods. *BMC Bioinformatics* **7**, S3 (2006).
118. Guda, C. pTARGET: a web server for predicting protein subcellular localization. *Nucleic Acids Res.* **34**, W210–W213 (2006).
119. Nielsen, H., Brunak, S. & von Heijne, G. Machine learning approaches to the prediction of signal peptides and other protein sorting signals. *Protein Eng.* **12**, 3–9 (1999).
120. Menne, K., Hermjakob, H. & Apweiler, R. A comparison of signal sequence prediction methods using a test set of signal peptides. *Bioinformatics* **16**, 741–742 (2000).
121. Klee, E.W. & Ellis, L.B. Evaluating eukaryotic secreted protein prediction. *BMC Bioinformatics* **6**, 256 (2005).
122. Hiller, K., Grote, A., Scheer, M., Munch, R. & Jahn, D. PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Res.* **32**, W375–W379 (2004).
123. Horton, P., Park, K.-J., Obayashi, T. & Nakai, K. Protein subcellular localization prediction with WoLF PSORT. *In Proceedings of the 4th Annual Asia Pacific Bioinformatics Conference APBC06* 39–48 (Taipei, Taiwan, 2006).
124. Gardy, J.L. *et al.* PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics* **21**, 617–623 (2005).
125. Bannai, H., Tamada, Y., Maruyama, O., Nakai, K. & Miyano, S. Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics* **18**, 298–305 (2002).
126. Nair, R. & Rost, B. Mimicking cellular sorting improves prediction of subcellular localization. *J. Mol. Biol.* **348**, 85–100 (2005).
127. Hawkins, J. & Bodén, M. Detecting and sorting targeting peptides with neural networks and support vector machines. *J. Bioinform. Comput. Biol.* **4**, 1–18 (2006).
128. Baldi, P., Brunak, S., Frasconi, P., Soda, G. & Pollastri, G. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics* **15**, 937–946 (1999).
129. Pollack, J.B. The induction of dynamical recognizers. *Mach. Learn.* **7**, 227 (1991).
130. Wakabayashi, M., Hawkins, J., Maetschke, S. & Bodén, M. Exploiting sequence dependencies in the prediction of peroxisomal proteins. in *Intelligent Data Engineering and Automated Learning—Vol 3578 of LNCS* (eds. Gallagher, M., Hogan, J. & Maire, F.) 454–461 (Springer–Verlag, Berlin, Germany, 2005).
131. Lu, Z. *et al.* Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics* **20**, 547–556 (2004).
132. Höglund, A., Dönnes, P., Blum, T., Adolph, H.-W. & Kohlbacher, O. MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinformatics* **22**, 1158–1165 (2006).