PSORT II Users' Manual

WWW version (date of last revision: Dec. 1, 1998)

Kenta Nakai

Institute for Molecular Biology, Osaka University, Japan

CONTENTS

Introduction

About PSORT II

This is a long-awaited new version of PSORT, a program to predict the subcellular localization sites of proteins from their amino acid sequences. Since the older version of PSORT was mostly written in a commercial language, OPS83, the distribution of the source code was difficult. Furthermore, since its knowledge base was represented as a collection of "if-then" rules and since the assignment of numeric certainty parameters to each rule had to be done manually, I could not add new rules nor tune the entire numeric parameters using newly-obtained sequence data with known localization sites. In this version, these difficulties have been overcome: the entire source code was rewritten by perl which will enable this program to run on almost all kinds of machines. Moreover, the reasoning algorithm was replaced with a very simple one, the *k*-nearest neighbors classifier, based on the collaboration with Paul Horton, Univ. California, Berkeley (Horton and Nakai, 1996; Horton and Nakai, 1997). Therefore, it is now possible to train the program using your own sets of sequence data. In this server, it was trained using the yeast sequences from SWISS-PROT (Bairoch and Apweiler, 1997). We expect the reliability of prediction is improved and will be improved further because much larger training data were used than those for the previous version. However, it runs much slower and I have to emphasize that the subprograms which calculate various scores from the query have not been updated yet in this stage. In other words, the knowledge base is still old although the algorithm to detect signal peptides has been refined (Nakai, 1996). I believe that much more improvements are possible on both the reasoning algorithm and the algorithms for feature detection.

Quick Start (WWW server)

In the input form, select an appropriate button for the source origin of your query sequence and paste that sequence in the field (this means that your WWW browser must support the form-fill feature). If your sequence is already registered in the regular release of SWISS-PROT, you may indicate it by its accession number, instead. Click the "Submit" button, and you will get the output. Its first part is the summary of your input for confirmation. The rest is the result of analyzing various sequence features related to protein sorting signals. The final prediction results, *i.e.*, the percentages of the localization sites of neighbors, are given finally.

Input Information for the WWW Server

Source of Input Sequence

Select one of the radio buttons to specify the source origin of the input sequence. This selection determines the candidate localization-sites for prediction as listed below:

Gram-positive bacterium (not yet supported):
    (cytoplasmic) membrane, cytoplasm, and outside, *i.e.*, the protein will be secreted.
Gram-negative bacterium (not yet supported):
    cytoplasm, inner membrane, periplasm, and outer membrane.
yeast and animal:

cytoskeleton, cytoplasm, nucleus, mitochondria, vesicles of secretory system, endoplasmic reticulum (ER), Golgi, vacuole, plasma membrane, peroxisome, extracellular space including cell wall.

plant (not yet supported):

Sequence Field

Enter the sequence here by direct typing or by copying & pasting. You may specify the accession number of SWISS-PROT but currently that for daily-updated data cannot be used. In case of sequence data, characters except standard one-letter code for 20 amino acids, *e.g.,* spaces, numerals, and carriage returns, will be removed off by the system. Small cases will be changed to capital ones.

The input sequence is treated as a full-length amino acid sequence containing all information for sorting. Thus, a warning message will be issued if it starts from an amino acid except M (methionine).

Output for Bacterial Sequences (not yet supported)

Gram-positive or Gram-negative

In the current version, programs and parameters are the same for both kinds of bacteria. The inner membrane in Gram-negative bacteria is thought to be equivalent to the cytoplasmic membrane of Gram-positive bacteria. Then, the outside in Gram-positive bacteria is further divided into the periplasm and the outer membrane in Gram-negative ones.

Recognition of Signal Sequence

In Gram-negative bacteria, most periplasmic and outer membrane proteins have a signal sequence (also called a leader peptide) in the N-terminus, which is cleaved off after the translocation of the cytoplasmic membrane. Some of the cytoplasmic membrane proteins also have a cleavable signal sequence but some N-terminal signals in the cytoplasmic membrane proteins remain as transmembrane segments. Such a signal is called the "signal-anchor" sequence.

PSORT first predicts the presence of signal sequences by McGeoch's method (D. J. McGeoch, *Virus Res.*, **3,** 271, 1985) modified by Nakai and Kanehisa, 1991 and Nakai, 1996. It considers the N-terminal positively-charged region (N-region) and the central hydrophobic region (H-region) of signal sequences. A discriminant score is calculated from the three values: length of H-region, peak value of H-region, and net charge of N-region. These results are summarized in **"PSG"** (formerly, it was called "McG"). A large positive discriminant score means a high possibility to possess a signal sequence but it is unrelated to the possibility of its cleavage.

Next, PSORT applies von Heijne's method of signal sequence recognition (G. von Heijne, *Nucl. Acids Res.*, **14,** 4683, 1986). It is a weight-matrix method and incorporates the information of consensus pattern around the cleavage sites (the (-3,-1)-rule) as well as the feature of the H-region. Thus it can be used to detect signal-anchor sequences. The output score of this **"GvH"** is the original weight-matrix score (for prokaryotes) subtracted by 7.5. A large positive output means a high possibility that it has a cleavable signal sequence. The position of possible cleavage site, *i.e.,* the most C-terminal position of a signal sequence, is also reported.

Recognition of Transmembrane Segments

In general, hydrophobic transmembrane segments exist in the cytoplasmic membrane proteins only.

Thus, these segments can be regarded as a sorting signal into the cytoplasmic membrane.

PSORT employs Klein et al.'s method (ALOM, also called KKD) to detect potential transmembrane segments (P. Klein, M. Kanehisa, and C. DeLisi, *Biochim. Biophys. Acta*, **815,** 468, 1985) modified by Nakai and Kanehisa, 1992. It repeats to identify the most probable transmembrane segment from the average hydrophobicity value of 17-residue segments, if any. It predicts whether that segment is a transmembrane segment (INTEGRAL) or not (PERIPHERAL) comparing the discriminant score (reported as 'ALOM score') with a threshold parameter(see below). For an integral membrane protein, position(s) of transmembrane segment(s) are also reported. Their length is fixed to 17 but their extension, *i.e.,* the maximal range that satisfies the discriminant criterion, is also given in parentheses. The discrimination step mentioned above is continued after leaving out the detected segment till there remains no predicted transmembrane segment. The item 'number of TMSs' is the number of predicted transmembrane segments. Since this algorithm is applied to a predicted mature sequence, *i.e.,* cleavable signal sequence is not included, this number is expected to be the one for mature proteins.

The modification by Nakai and Kanehisa, 1992 was to employ two kinds of threshold values because ALOM is not very accurate to predict the exact number of transmembrane segments of polytopic, *i.e.,* multiple membrane-spanning, proteins. The rationale of our approach is that less hydrophobic segments are likely to be more easily integrated into the membrane once a part of the polypeptide is integrated. Specifically, PSORT first tentatively evaluates the number of TMSs using less stringent value (0.5). Then, it re-evaluates the number by using a more stringent threshold (-2.0). If it is still predicted to have at least one TMS, the former threshold value is used. This modified algorithm is named **"ALOM2"** and the perl program is also available.

Analysis of Lipoproteins

The signal sequence of lipoproteins, *i.e.,* proteins with a covalently attached lipid molecule in their mature N-terminus, are essentially the same as those of usual proteins except the region around their cleavage sites. Thus, they can be recognized by the combination of McGeoch's method and the consensus motif around the cleavage site formulated by von Heijne (G. von Heijne, *Protein Eng.*, **2,** 531, 1989). The program is named as **"Lipop"** here. It gives the possible modification site around the end position of preceding H-region defined in McGeoch's method for a probable lipoprotein; otherwise, it returns a dummy modification site, -1.

Since the N-terminal lipid moieties of lipoproteins are thought to be integrated into membranes, they are predicted to be membrane-associated proteins. For Gram-negative bacterial proteins, further discrimination between the cytoplasmic membrane and the outer membrane is needed. It is done as follows based on the experiment of Yamaguchi *et al.* (K. Yamaguchi, F. Yu, and M. Inoue, *Cell*, **53,** 423, 1988): If a lipoprotein has a negatively charged residue at the second or the third position of the mature part, it is sorted to the inner membrane; otherwise, it is sorted to the outer membrane.

Analysis of Amino Acid Composition

In Gram-negative bacteria, although outer membrane proteins are integrated into the membrane, they do not have any hydrophobic segments which characterize usual integral membrane proteins. It is interpreted that their membrane-spanning parts consist of beta-strands. In addition, the sorting signal which discriminates outer membrane proteins from periplasmic proteins is not well characterized. Therefore, PSORT uses the information of amino acid composition of the predicted mature portion for their discrimination (Nakai and Kanehisa, 1991). That is, a discriminant score is calculated by the linear combination of the percentage of 10 amino acids for the amino acid sequence except for the predicted N-terminal signal sequence, if any. Large positive scores mean the tendency to be an outer membrane protein. In addition, PSORT also examines whether the C-terminal residue is phenylalanine or not because its role on sorting to the outer membrane has been shown for some proteins (M. Struyve *et al.*, *J. Mol. Biol.*, **218**, 141, 1991).

Output for Eukaryotic Information

Yeast, Animal, (or Plant)

In this version of PSORT, parameters for analyzing yeast (or plant) sequences are the same with parameters for animal sequences. Yeast (and plant) have a candidate site, vacuole, instead of lysosome in animal. In yeast, the consensus sequence for ER-lumen retention is HDEL rather than KDEL in others. Lastly, plants have a chloroplast as an extra-candidate (not yet supported).

Recognition of Signal Sequence

In eukaryotes, proteins sorted through the so-called vesicular pathway (bulk flow) usually have a signal sequence (also called a leader peptide) in the N- terminus, which is cleaved off after the translocation through the ER membrane. Some N-terminal signal sequences are not cleaved off, remaining as transmembrane segments but it does not mean these proteins are retained in the ER; they can be further sorted included in vesicles.

PSORT first predicts the presence of signal sequences by McGeoch's method (D. J. McGeoch, *Virus Res.*, **3,** 271, 1985) modified by Nakai and Kanehisa, 1991 and Nakai, 1996. It considers the N-terminal positively-charged region (N-region) and the central hydrophobic region (H-region) of signal sequences. A discriminant score is calculated from the three values: length of H-region, peak value of H-region, and net charge of N-region. These results are summarized in **"PSG"** (formerly, it was called "McG"). A large positive discriminant score means a high possibility to possess a signal sequence but it is unrelated to the possibility of its cleavage.

Next, PSORT applies von Heijne's method of signal sequence recognition (G. von Heijne, *Nucl. Acids Res.*, **14,** 4683, 1986). It is a weight-matrix method and incorporates the information of consensus pattern around the cleavage sites (the (-3,-1)-rule) as well as the feature of the H-region. Thus it can be used to detect signal-anchor sequences. The output score of this **"GvH"** is the original weight-matrix score (for eukaryotes) subtracted by 3.5. A large positive output means a high possibility that it has a cleavable signal sequence. The position of possible cleavage site, *i.e.,* the most C-terminal position of a signal sequence, is also reported.

Recognition of Transmembrane Segments

The current version of PSORT assumes that all integral membrane proteins have hydrophobic transmembrane segment(s) which are thought to be alpha- helices in membranes.

PSORT employs Klein et al.'s method (ALOM, also called KKD) to detect potential transmembrane segments (P. Klein, M. Kanehisa, and C. DeLisi, *Biochim. Biophys. Acta*, **815,** 468, 1985) modified by Nakai and Kanehisa, 1992. It repeats to identify the most probable transmembrane segment from the average hydrophobicity value of 17-residue segments, if any. It predicts whether that segment is a transmembrane segment (INTEGRAL) or not (PERIPHERAL) comparing the discriminant score (reported as 'ALOM score') with a threshold parameter(see below). For an integral membrane protein, position(s) of transmembrane segment(s) are also reported. Their length is fixed to 17 but their extension, *i.e.,* the maximal range that satisfies the discriminant criterion, is also given in parentheses. The discrimination step mentioned above is continued after leaving out the detected segment till there remains no predicted transmembrane segment. The item 'number of TMSs' is the number of predicted transmembrane segments. Since this algorithm is applied to a predicted mature sequence, *i.e.,* cleavable signal sequence is not included, this number is expected to be the one for mature proteins.

The modification by Nakai and Kanehisa, 1992 was to employ two kinds of threshold values because ALOM is not very accurate to predict the exact number of transmembrane segments of polytopic, *i.e.*, multiple membrane-spanning, proteins. The rationale of our approach is that less hydrophobic segments are likely to be more easily integrated into the membrane once a part of the polypeptide is integrated. Specifically, PSORT first tentatively evaluates the number of TMSs using less stringent value (0.5). Then, it re-evaluates the number by using a more stringent threshold (-2.0). If it is still predicted to have at least one TMS, the former threshold value is used. This modified algorithm is named **"ALOM2"** and the perl program is also available.

Prediction of Membrane Topology

Every membrane protein has its own orientation to be integrated in the membrane; in other words, membrane proteins know at which side (cytoplasmic or exo-cytoplasmic) its N-terminus should be located. Such an orientation is called the membrane topology. We used Singer's classification for membrane topology (S. J. Singer, *Ann. Rev. Cell Biol.*, **6,** 247, 1990). Prediction of membrane topology is important because some sorting signals exist at specific positions, *e.g.*, the cytoplasmic tail, in a certain topology (see below).

PSORT uses Hartmann et al.'s method (E. Hartmann, T. A. Rapoport, and H. F. Lodish, *Proc. Natl. Acad. Sci. USA*, **86,** 5786, 1989); called **"MTOP"** in PSORT) for the prediction of membrane topology. MTOP assumes that the overall topology of eukaryotic membrane proteins is determined by the net charge difference of 15 residues flanking the most N-terminal transmembrane segment on both sides. The central residue of such a segment is first reported.

If a protein is predicted to have a cleavable signal sequence and one transmembrane segment, its topology is '1a'. If a protein is predicted to have no cleavable signal sequence but has one transmembrane segment, its position is examined. If it exists near its C-terminus. Its topology is assigned as 'Nt (N-tail)' (see U. Kutay *et al., Trends Cell Biol.*, **3,** 72-75, 1993). Otherwise, its topology is assigned to '1b' or '2' depending on the charge difference reported by MTOP. For polytopic proteins, their topology is simply predicted by MTOP (bug?).

There seems to be a preference of membrane topology at each localization site. For example, type Ib proteins are favored at the ER while type II tend towards the Golgi complex and the plasma membrane. Thus, dummy variables representing the topology are used for later prediction.

Recognition of Mitochondrial Proteins

Although many proteins which engage in mitochondrial protein targeting have been characterized, their exact pathways has not been fully understood. Many proteins transported to mitochondria have a mitochondrial targeting signal on their N-terminus. Some seem to have internal signals but they may be recognized by a common cytosolic factor (MSF).

PSORT employs a very simple method to recognize mitochondrial targeting signals: the discriminant analysis (called **"MITDISC"**) whose variables are the amino acid composition of the N-terminal 20 residues (Nakai and Kanehisa, 1992).

PSORT also reports some consensus patterns around the cleavage sites (**"Gavel"**), reported in Y. Gavel and G. von Heijne, *Prot. Eng.*, **4,** 33, 1990). However, the result is not used in the prediction because of its lack of reliability.

Hopefully, more reliable prediction methods such as (Claros and Vincens, 1996 and Fujiwara *et al.*, 1997 should be incorporated in the near future.

In this version, further discrimination of signals directing to the substructures of mitochondria, *e.g.*,

intermembrane space, is not attempted although proteins which are predicted to have both the mitochondrial targeting signal and transmembrane segment(s) are likely to be localized at the inner membrane.

Recognition of Nuclear Proteins

Although it seems possible that a protein without its own nuclear localization signal (NLS) enters the nucleus via cotransport with a protein that has one, many nuclear proteins have their own NLSs. Presently, NLSs are classified into three categories (reviewed in G. R. Hicks and N. V. Raikhel, *Ann. Rev. Cell Dev. Biol.*, **11,** 155, 1995).

The classical type of NLSs is that of SV40 large T antigen. PSORT uses the following two rules to detect it: 4 residue pattern (called 'pat4') composed of 4 basic amino acids (K or R), or composed of three basic amino acids (K or R) and either H or P; the other (called 'pat7') is a pattern starting with P and followed within 3 residues by a basic segment containing 3 K/R residues out of 4.

Another type of NLS is the bipartite NLS, first found in *Xenopus* nucleoplasmin by Robbins *et al*. (J. Robbins, S. M. Dilworth, R. A. Laskey, and C. Dingwall, *Cell*, **64,** 615, 1991). The pattern (called 'bipartite') is: 2 basic residues, 10 residue spacer, and another basic region consisting of at least 3 basic residues out of 5 residues.

The last category of NLS is the type of an N-terminal signal found in yeast protein, Mat alpha2. However, PSORT doesn't try to find it because this type of signal has not been well studied. Nor the knowledge of Nuclear Export Signals (NESs) has not been incorporated yet.

In the yeast genome, nuclear proteins occupy the majority. Since the precise discrimination of NLSs is presently difficult, the prediction of nuclear proteins affect much to the total prediction accuracy. Then, PSORT uses a heuristic that nuclear proteins are generally rich in basic residues: If the sum of K and R compositions are higher than 20%, then the protein is considered to have higher possibility of being nuclear than cytoplasmic. Moreover, a score (called "**NNCN**") which discriminates the tendency to be at either the nucleus or the cytoplasm is calculated based on the amino acid composition according to the neural network constructed by Reinhardt (A. Reinhardt and T. Hubbard, *Nucl. Acids Res.* **26**, 2230, 1998). This routine was originally given from Dr. Reinhardt and was translated into perl by K. Nakai. You can test the full program developed by Reinhardt and Hubbard from here.

Most scores mentioned above are combined by a discriminant function to give the 'NLS score'. In addition, PSORT examines the presence of RNP (ribonucleoprotein) consensus motif (called 'RNA-binding motif') (K. Nagai, *et al*., *Trends Biochem. Sci*., **20,** 235, 1995) because some RNPs are transported to the nucleus by signals existing in the bound RNAs. However, it is apparently insufficient for actual prediction.

In this version, we classify ribosomal proteins as cytoplasmic proteins although some of them have NLSs and are once transported into the nucleus.

Recognition of Peroxisomal Proteins

Peroxisomes, sometimes called glyoxisomes, glycosomes, or microbodies, are organelles found in almost every eukaryotic cell. Several sorting signals into peroxisomes have been characterized (see J. A. McNew and J. M. Goodman, *Trends Biochem. Sci*., **21,** 54, 1996). Obviously there are rooms for further improvement. As for peroxisomal-matrix targeting sequences (PTSs), two kinds of them are known. One is the tripeptide, (S/A/C)(K/R/H)L, at the C-terminus, known as PTS1 or the SKL-motif. PSORT calculates score of a given sequence empirically. The other is the N-terminal segment known as PTS2. The consensus patterns of known PTSs, (R/K)(L/I)xxxxx(H/Q)L is searched in the present version, although its importance has not been fully verified.

The sorting signal of peroxisomal membrane proteins (mPTSs) is not well understood although the importance of a hydrophilic loop of 20 residues facing the matrix in Pmp47 is known.

Recognition of Chloroplast Proteins (not yet supported)

Proteins targeted to chloroplasts have cleavable signals in the N-terminus, the chloroplast (stroma) targeting signals. PSORT postulates that all stromal proteins and thylakoid membrane proteins have this kind of signal. It uses a discriminant score calculated from partial amino acid compositions (positions 3-10 and 1-30) and from the amplitude of maximum hydrophobic moment of 165 degrees (potential beta-structure) for residues 25 to 70 (Nakai and Kanehisa, 1992). The form of discriminant function shows the abundance of alanine and serine residues in the N-terminal 30 residues. In addition, the observation that the second residue is often alanine is also used.

Like some mitochondrial proteins, proteins of chloroplast thylakoid lumen have a bipartite signal in their N-terminus. Its N-terminal half is essentially the same as a stroma targeting signal and the C-terminal half is used for the translocation from the stroma to the thylakoid lumen. For the detection of latter signal, another clue, PSORT uses both the result of APOLAR algorithm applied to the limited region of residues 40 to 90 and a weight matrix score around the cleavage sites (C. J. Howe and T. P. Wallace, *Nucl. Acids Res.*, **18,** 3417, 1990).

Thylakoid membrane proteins were discriminated by ALOM. The remainder of chloroplast proteins are tentatively regarded as stromal proteins.

Recognition of ER (endoplasmic reticulum) Proteins

PSORT postulates that the proteins with N-terminal signal sequence will be transported to the cell surface by default unless they have any other signals for specific retrieval, retention, or commitment; a luminal protein will be secreted constitutively to the extracellular space and a membrane protein will reside at the plasma membrane. For a recent review of ER retrieval signals, see R. D. Teasdale and M. R. Jackson, *Ann. Rev. Cell Dev. Biol.*, **12**, 27, 1996.

The retrieval signal of ER luminal proteins from the bulk flow is the consensus motif, KDEL (HDEL in yeast), in the C-terminus. In addition, these proteins should have a cleavable signal sequence in their N-terminus but the existence of KDEL is often practically sufficient. Although PSORT only recognizes the (K/H)DEL pattern, it is known that some variations of this motif are allowed in some organisms and/or cell types.

The retrieval signals for ER membrane proteins appear more complex. Two kinds of signals are known; one is the di-lysine motif (the KKXX motif) which exist near the C-terminus of type Ia proteins and the other is the di-arginine motif (the XXRR motif) which exist near the N-terminus of type II proteins. Note that both of these motifs exist close to the terminus of the cytoplasmic tail. However, for the practical prediction, the existence of these motifs itself is not necessary nor sufficient for the localization at the ER membrane. Thus, the reliability of prediction is not high in this stage.

Analysis of Proteins in Vesicular Pathway

The signals that govern the protein sorting through the vesicle transport are complex because they are inter-related and they seem to be recognized in various situation. Moreover, there are redistribution processes via endocytosis and a pathway to lysosomes. Thus, we, should also consider these signals.

As already exemplified above, many sorting signals of membrane proteins transported through these pathways exist in the cytoplasmic tail, a short terminal segment exposed to the cytosol in type Ia, Ib, and II proteins (J. E. Rothman and F. T. Wieland, *Science*, **272,** 227, 1996; for the terminology of membrane

topology, see above).

For membrane proteins with (usually) a single transmembrane segment, PSORT tries to find the following motifs in the cytoplasmic tail: the YQRL motif which directs the transport from cell surface to Golgi; the tyrosine-containing motif and the dileucine motif for selective inclusion in clathrin-coated vesicles (endocytosis) and lysosomal targeting; (recently, the role of another motif, di-acidic signal, has been also shown in Nishimura and Balch, 1997). However, the subprograms to detect these motifs are rather primitive. Further elaboration is clearly needed.

Lysosomal and Vacuolar Proteins

Lysosomes are acidic organelles that contain numerous hydrolytic enzymes. In yeast and plant cells, similar activities are seen in vacuoles (lysosome-like vacuoles) and the protein sorting mechanisms for these organelles are conserved to some degree (reviewed in J. H. Stack, B. Horazdovsky, and S. D. Emr, *Ann. Rev. Cell Dev. Biol.*, **11**, 1, 1995).

In mammalian lysosomes, one sorting signal of soluble (lumenal) proteins is a post-translational modification, addition of mannose-6-phosphate, but there is also a pathway which is independent of mannose-6-phosphate. Two kinds of mannose-6-phosphate receptor are known but the substrate specificity of the enzyme which adds mannose-6-phosphate is not well understood although the importance of a specific conformation, a beta-hairpin motif, has been implicated.

Although yeast vacuoles have been studied as a model system of mammalian lysosomes, soluble proteins of yeast vacuole do not use the mannose-6-phosphate dependent pathway. Analyses of several yeast proteins suggest that the pro-peptides, which are exposed to the N-terminus after the cleavage of 'pre' signal sequence, work as a sorting signal. However, there are not apparently conserved motifs except for a weak motif, (T/I/K)LP(L/K/I), which PSORT searches for.

The sorting mechanism of lysosomal membrane proteins seems different from that of lysosomal luminal proteins. The existence of the GY motif within 17 residues from the membrane boundary in the cytoplasmic tail seems to be important for some of them. The signals for vacuolar membrane proteins have not been clarified yet.

In this version of PSORT, the prediction accuracy for lysosomal/vacuolar proteins is disastrous. No doubt are further efforts needed.

Lipid Anchors

The protein modification reactions which bind lipid molecules to proteins are important because a linked lipid moiety can be integrated into various membranes and can anchor the bound protein.

For example, myristoylations occur at the consensus sequence in the N-terminal 9 residues (reviewed in D. R. Johnson, *et al.*, *Ann. Rev. Biochem.*, **63**, 869, 1994). PSORT predicts its presence by the 'NMYR' program but note that the N-myristoylated proteins are not always anchored to the membrane.

In contrast, all proteins linked to the glycosyl-phosphatidylinositol (GPI) molecules are thought to be anchored at the extracellular surface of the plasma membrane. In addition, GPI anchor plays some roles on the protein sorting in polarized cells. Although much is known about the biosynthesis of GPI-anchor (reviewed in J. Takeda and T. Kinoshita, *Trends Biochem. Sci.*, **20**, 367, 1995), PSORT predicts GPI-anchored proteins by empirical knowledge that most of them are the type Ia membrane proteins with very short cytoplasmic tail (within 10 residues).

Lastly, there is a lipid modification known as (iso)prenylation (*i.e.*, farnesylation or geranylgeranylation; reviewed in F. L. Zhang and P. J. Casey, *Ann. Rev. Biochem.*, **65**, 241, 1996). This modification

requires a CaaX motif in the C-terminus, where 'a' denotes an aliphatic amino acid. Prenylated proteins have been found in the plasma membrane and the nuclear envelope. Since such knowledge may not be reflected in the *k*-NN method, users should notice the report from PSORT.

Miscellaneous Motifs

Experimentally the knowledge of various protein functional motifs in the PROSITE database (Bairoch, *et al.*, 1997) has been included. To discriminate cytoskeletal proteins, two motifs of actin-binding proteins were examined but apparently our knowledge is insufficient.

Ideally, the prediction of PSORT should be done based on the knowledge on various sorting signals only and should be independent from the knowledge on the functions of query proteins because it will be hopefully applied for engineered/artificial sequences. However, since our knowledge on the NLSs seems to be too little to make reliable prediction, these PROSITE motifs are experimentally introduced: 63 DNA binding motifs, which may be useful to distinguish nuclear proteins; 71 ribosomal protein motifs, which may be necessary because the sorting processes of ribosomal proteins are complex; 33 prokaryotic DNA binding motifs, which might be useful for the prediction of bacterial sequences.

Coiled-coil Structure

Coiled-coil structures are found in some structural proteins, *e.g.*, myosins, and in some DNA-binding proteins as the so-called leucine-zipper. In this structure, two alpha-helices bind each other making a coil, where these two alpha-helices show a 3.5-residue periodicity which is slightly different from the typical value, 3.6. Thus, the detection of coiled-coil structure by searching for 7-residue periodicity is relatively more accurate than usual secondary structure prediction. Although the presence of coiled-coil structure in a protein itself does not indicate its subcellular localization, such information might be useful for the people who try to characterize ORFs of unknown function. Currently, a classical detection algorithm developed by A. Lupas is used (Lupas *et al.*, 1991). The sequences which are likely to be in a coiled-coil conformation are reported. For more recent progresses of the coiled-coil prediction, see this Web site.

Notes on the Reasoning System

k-Nearest Neighbors Classifier

As stated before, the most drastic change in PSORT II is the use of *k*-nearest neighbor (*k*-NN) algorithm for assessing the probability of localizing at each candidate sites (Horton and Nakai, 1997). Namely, for each query protein, the output values of the subprograms mentioned above are normalized and simple Euclid distances to all of the data points contained in the training data are calculated. Then, the prediction is performed using the *k*-nearest data points, where *k* is a predefined integer parameter. If these *k* data points contain, say, nuclear proteins with 50%, the query is predicted to be localized to the nucleus with the probability of 50%.

This prediction algorithm seems to be problematic for one point in our analysis. Because the data size for each localization site varies very much, the sites with smaller samples would be hard to be predicted if a large *k* value is used. Therefore, an experimental modification, two-fold *k*-NN, is employed. Namely, two different *k* values ($k1 < k2$) are used and the localization sites are classified into two categories according to their data size. First, a prediction is performed using the smaller *k1* value. If its predicted site belongs to the smaller category, the algorithm terminates; otherwise, the prediction is redone using the larger *k2*. Currently, *k1* and *k2* are set to 9 and 23, respectively. Unfortunately, this two-fold *k*-NN does not seem to be effective so much. More studies should be needed.

Training Data

In this distribution package, the program uses a reference data set for distance calculation, collected from 1080 yeast sequences from SWISS-PROT (Bairoch and Apweiler, 1997). Because most sequences are expected to be coded in different genes, the redundancy caused from similar sequences is ignored. However, proteins coded in the mitochondrial genome were omitted (I would like to thank Les Grivell for pointing out my careless mistakes). The ratio of protein numbers for each localization site can be regarded as *a priori* probability for prediction.

Reliability of Prediction Result

The superiority of the *k*-NN algorithm has been shown in Horton and Nakai, 1997. At that time, its prediction accuracy evaluated by the cross-validation procedure was approximately 57% for yeast sequences and 86% for *E. coli* sequences. However, the labels of candidate localization sites, the number of data used for distance calculation, and the variables used are not the same with those in this version. So, I would like to emphasize that it is a beta-version whose performance has not been explored. We plan to distribute the whole program and data under the GNU copyleft agreement when the tuning of the algorithm is done. Any comments, criticisms, and bug reports are welcome.

---

nakai@imcb.osaka-u.ac.jp