Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
A Knowledge Base for Predicting Protein Localization Sites in Eukaryotic Cells

KENTA NAKAI¹ AND MINORU KANEHISA

Institute for Chemical Research, Kyoto University, Uji, Kyoto 611, Japan

Received October 15, 1991; revised September 11, 1992

To automate examination of massive amounts of sequence data for biological function, it is important to computerize interpretation based on empirical knowledge of sequence-function relationships. For this purpose, we have been constructing a knowledge base by organizing various experimental and computational observations as a collection of if–then rules. Here we report an expert system, which utilizes this knowledge base, for predicting localization sites of proteins only from the information on the amino acid sequence and the source origin. We collected data for 401 eukaryotic proteins with known localization sites (subcellular and extracellular) and divided them into training data and testing data. Fourteen localization sites were distinguished for animal cells and 17 for plant cells. When sorting signals were not well characterized experimentally, various sequence features were computationally derived from the training data. It was found that 66% of the training data and 59% of the testing data were correctly predicted by our expert system. This artificial intelligence approach is powerful and flexible enough to be used in genome analyses.

INTRODUCTION

Computational approaches are becoming indispensable components of molecular and cellular biology, especially in the analyses of human and other complex genomes for which massive amounts of sequence data must be examined for biological function. Functional information can be obtained from sequence information not by solving equations of first principles, but by inference based on empirical knowledge. Although the sequence data are now collected and organized in publicly available databases, functional data are not well organized, except, perhaps, in the brain of a human expert. We have been experimenting with an artificial intelligence approach called an expert system (see Waterman, 1986, for example) for collecting and utilizing experts' specific knowledge, as well as computationally acquired knowledge, for the task of protein sorting (Nakai and Kanehisa, 1991). In contrast to the representation of protein sequence data by 20 letters for which no arguments can be made except for possible sequencing errors, the representation of functional data can be far more controversial because it always requires interpretation of observed phenomena.

In eukaryotic cells, the sorting signals that direct proteins to proper subcellular locations are usually encoded in their own amino acid sequences. A growing body of experimental evidence has been clarifying the nature of these signals. There are multiple ways of representing such empirical knowledge for computer processing. A production system utilizes a knowledge base constructed as a collection of "if–then" rules. It is relatively easy to implement knowledge in this manner, and a number of expert systems are of this type. In our previous work, we constructed an expert system for predicting protein localization sites in Gram-negative bacteria by organizing the relationships between amino acid sequence features and functional aspects in the form of if–then rules (Nakai and Kanehisa, 1991). The expert system could discriminate 106 proteins in our database into four localization sites with 83% accuracy. In the present work, our previous system is expanded so that it can also predict the localization sites of eukaryotic proteins, incorporating knowledge of a variety of sorting signals (Verner and Schatz, 1988).

This is a first attempt to systematically allocate various protein localization sites in eukaryotic cells from a theoretical point of view, which requires consideration of the following. First, because protein-sorting signals are mutually related, it does not seem sufficient to examine each feature separately. Second, because of the variations in how such features are encoded, it is not possible to treat all signals uniformly, say, by a single discriminant function. Third, because of the difficulty of interpreting the functional data mentioned above, any theory should make experimentally testable predictions and be able to evolve as new insights are gained. A knowledge-based approach is suitable for solving problems in these circumstances.

¹ Present address: National Institute for Basic Biology, Okazaki 444, Japan.
TABLE 1
The Numbers of Proteins in the Repertoire of Localization Sites

| Localization site             | Abbreviation | Number of proteins |
|-------------------------------|--------------|-------------------|
| Chloroplast (for plants)      | CHST         | 15                |
| Stroma                        | CHTM         | 7                 |
| Thylakoid membrane            | CHTS         | 6                 |
| Thylakoid space               | CP           | 33                |
| Cytoplasm                     |              | 14                |
| Endoplasmic reticulum         | ERL          | 4                 |
| Lumen                         | ERM          | 13                |
| Membrane                      | GG           | 6                 |
| Golgi complex                 | LSL          | 8                 |
| Lysosome (for animals)        | LSM          | 3                 |
| Lumen (planta)                | MMM          | 10                |
| Membrane (planta)             | MTOM         | 2                 |
| Mitochondrion                 | NC           | 43                |
| Extracellular space (outside) | OT           | 50                |
| Plasma membrane               | GPI          | 14                |
| GPI-anchored                  | PM           | 33                |
| Integral                      | PX           | 13                |
| Peroxisome                    | VC           | 9                 |
| Vacuole (for plants and yeasts)| VC           | 255               |

**MATERIALS AND METHODS**

**Localization sites.** The localization sites considered in this study are cytoplasm (one site), nucleus (one site), mitochondrion (four sites), chloroplast (three sites), peroxisome (one site), endoplasmic reticulum (two sites), Golgi complex (one site), lysosome (two sites), vacuole (one site), plasma membrane (two sites), and extracellular space (one site). They are summarized in Table 1. For proteins from animal cells, the lysosome is added to the repertoire, whereas the vacuole (one site), plasma membrane (two sites), and extracellular space (one site) are included in plant proteins. Cytoplasmic proteins include cytoskeletal ones. Peroxosomal proteins include so-called microbody proteins, but those of trypanosomes are excluded. GPI-anchored plasma membrane proteins are treated separately from other integral membrane proteins because the prediction logic is different.

There is some ambiguity in the definition of protein localization sites. For example, some proteins have more than one localization site (see Discussion). Peripheral membrane proteins are distinguished from integral membrane proteins and define their localization sites to be the spaces toward which the membrane surfaces face. Some proteins become soluble after specific proteolytic processing reactions in the membrane-bound precursor form. In the present work, we regard them as membrane proteins unless the reaction is coupled with the translocation process.

**Sequence data.** The amino acid sequence data of known localization sites were collected from the NBRF-PIR database, release 27.0 (Barker et al., 1990). The total of 401 sequences can be obtained from Kenta Nakai (e-mail: nakai @ nibh.ac.jp).

Except for some proteins in which the N-terminal methionine was removed, all sequences were in the genetically coded (unprocessed) form. To eliminate redundant data, these sequences have been selected from the original database such that there were no pairs with more than 50% identical residues. Exceptions were a few isozymes localizing at different sites or with apparently different localization mechanisms. We divided our dataset into training data for extracting knowledge and testing data for evaluating the prediction ability. When a localization site contained more than 10 proteins, 30% of them were used as testing data; the remainder were used as training data. When proteins at a certain site were divided into groups sorted by different pathways, testing data were selected proportionally. The number of proteins in each category of our dataset is summarized in Table 1.

**Terminology for sorting signals.** There is as yet no consensus for the terminology of various sorting signals. We adopted Varshavsky’s (1991) proposal for a systematic naming method. This naming convention, as well as part of our own, is shown in Table 2.

**Expert system.** The architecture of our expert system is the same as that previously reported (Nakai and Kanehisa, 1991). It is a component is divided into three modules, corresponding to the order of reasoning steps. In the first module, rules are sorted for classifying organism name into one of the five categories: Gram-positive bacteria, Gram-negative bacteria, yeast, animal, or plant. The second and third modules contain rules for sorting signals, the second containing general rules and the third containing more specialized rules. This distinction is somewhat arbitrary, but the reasoning with more specialized rules is performed after the general reasoning step. The calculations involving sequence data are written in C language and are called from rules when necessary. The flow of reasoning occurs in a backward fashion; each localization site of the repertoire is activated one by one as a hypothesis and the rules for examining sequence characteristics for each localization site are invoked to verify the hypothesis. Thus, given the amino acid sequence and the organism name, our expert system reports a list of probable localization sites with certainty factors.

**Discriminant analysis.** Occasionally, we use the method called stepwise discriminant analysis (Nakai and Kanehisa, 1988, 1991) for deriving an optimal combination of various sequence features. Suppose, for example, proteins sorted to a certain localization site have amino acid compositions different from those of proteins sorted to another site. Using 20 amino acid contents as variables, the stepwise discriminant analysis determines the set of coefficients that maximally discriminates the two groups. The derived discriminant function is represented by the formula

\[ y = \sum a_i x_i + \text{const}, \]

where \( x_i \) is the variables to be selected (say, 20 amino acid contents) and \( a_i \) and const are the coefficients. When used for prediction, each unknown sequence is classified into one of the two groups according to the sign of this function. To avoid excessive dependence on training data, the stepwise procedure is usually not repeated until all 20 coefficients are determined. It is possible to regard the order of selection of variables as corresponding roughly to the order of their importance.

**Other analytical methods.** (i) Hydrophobic moment. The hydrophobic moment is calculated according to Eisenberg’s (1984) method. The amplitude of the moment with a given angle is calculated as the average value over 11 residues around each position in the sequence. The maximum value and its position may be used as variables for discriminant analysis.

(ii) The ‘apot’ algorithm. The algorithm we used to detect an apolar region is as follows. First, we define, somewhat arbitrarily, an index for the apolar value of each amino acid: 1.0 for A, I, L, F, V; 0.5 for C, G, M; 0.0 for S, T, Y; 1.0 for P, W; -3.0 for N, Q, H; and -5.0 for R, D, K. For each of the positions in the search area, if the apolar value is positive, then the following positions are considered together until the sum of the apolar values becomes equal to or less than the sign of this function. To eliminate redundant data, these sequences have been selected from the original database such that there were no pairs with more than 50% identical residues. Exceptions were a few isozymes localizing at different sites or with apparently different localization mechanisms. We divided our dataset into training data for extracting knowledge and testing data for evaluating the prediction ability. When a localization site contained more than 10 proteins, 30% of them were used as testing data; the remainder were used as training data. When proteins at a certain site were divided into groups sorted by different pathways, testing data were selected proportionally. The number of proteins in each category of our dataset is summarized in Table 1.
than -5.0. The output is the segment corresponding to the best sum value.

(iii) The "alom" algorithm. Transmembrane segments in membrane proteins were detected by the alom program (Klein et al., 1985). It is based on a discriminant function between integral and peripheral membrane proteins, applied to all 17 residue segments in the sequence.

RESULTS

Mitochondrial Presequences

To discriminate mitochondrial proteins, we searched for the M-transferon signal (rule 'mtmod'; see Table 4 for a summary of rules in our knowledge base; see also Table 2 for our naming convention of sorting signals), although several mitochondrial proteins without this signal are known to be sorted by other pathways (Hartl and Neupert, 1990; Baker and Schatz, 1991). Since there had been no proposed method of recognizing unknown M-transferons, we searched for an optimal set of sequence features that best discriminate M-transferons. We utilized the features reported previously. Von Heijne et al. (1989) used comparisons of aligned sequences to indicate sequence features of M-transferons. However, the amplitude and the location of maximal hydrophobic moments at both 95° and 75° were not selected as effective variables in the stepwise discriminant analysis. Gavel and von Heijne (1990b) proposed a method for recognizing the cleavage site motif (rule 'exgavel'). However, we could not raise the prediction accuracy by incorporating the information of predicted cleavage sites (data not shown).

The best discriminant function for our training data was obtained from the amino acid composition of the 20-residue segment at the N-terminus (rule 'mtdisc'). The variables selected and their corresponding coefficients are shown in Table 3a. Variables are listed in the order of stepwise selection, which roughly corresponds to the order of relative importance. In Table 3a we can see that R tends to strongly favored (positive coefficient) but other charged or polar residues are disfavored (negative coefficients) in the N-terminal region of M-transferons. In all, 90% of the training data could be correctly discriminated.

Intramitochondrial Sorting

Since an M-transferon is the signal that brings a newly synthesized protein to the mitochondrial matrix, further sorting signals must exist (Hartl and Neupert, 1990). Indeed, some intermembrane space proteins have a presequence of bipartite structure. The N-terminal half of the presequence is cleaved off at the matrix space and then the C-terminal half (M/IMS-transferon) is used for the translocation signal into the intermembrane space (rule 'mt2nd'). In our data for intermembrane space proteins, however, those that use this conservative pathway seem to be in the minority (two of five). There were two more examples of the bipartite signal in our inner membrane protein data. All but one of the four proteins with the bipartite signal had high discriminant scores of M-transferons at the N-terminus. In addition, using the 'apolar' algorithm, we found typical apolar stretches in the region from the N-terminus to the 70th residue (rule 'mtit'). However, these stretches were not always located near the second cleavage site. Interestingly, one of the two outer membrane proteins, the 45K protein of yeast, showed features characteristic of intermembrane space proteins (rule 'mtom'). In this case, the apolar region turned out to be near the N-terminus, i.e., starting at the fifth residue.

Other proteins that were predicted to have M-transferons were further classified into inner membrane or matrix proteins based on the existence or absence of transmembrane stretches detected by the 'alom' program (rules 'mtim' and 'mtmx'). However, many inner membrane proteins did not have apparent hydrophobic segments.

Nuclear Proteins

The sorting mechanism of nuclear proteins differs from that of other proteins (Silver, 1991). The main dif-
### TABLE 3

**Discriminant Functions**

**Variables: Amino acid contents of the N-terminal region of 20 residues**

| Variable: | R | D | P | E | G | Q |
|-----------|---|---|---|---|---|---|
| Coefficient: | 0.116 | -0.238 | -0.253 | -0.233 | -0.250 | -0.155 |
| H | 0.239 | -0.113 |

- **Data to be discriminated:** Those having M-transferons in MTIM & MTMX
- **Control data:** CP, NC, and PX
- **Discrimination result:** False negative: 1/20; False positive: 10/86

**Total hits:** 89.6%

**Variables: Amino acid contents of the entire sequence**

| Variable: | F | W | N | E | C | Y |
|-----------|---|---|---|---|---|---|
| Coefficient: | 1.384 | 1.905 | -0.462 | -0.177 | -0.585 | 0.457 |
| H | 0.521 | 0.351 |

- **Data to be discriminated:** PX except for JS0371
- **Control data:** NC and CP
- **Discrimination result:** False negative: 1/12; False positive: 13/77

**Total hits:** 84.3%

**Variables: Amino acid contents of the predicted mature portion**

| Variable: | K | C | E | S | D | Const |
|-----------|---|---|---|---|---|-------|
| Coefficient: | -0.442 | -0.535 | -0.598 | -0.506 | -0.351 | 12.625 |
| T | 0.740 | -0.669 | -1.153 | -0.463 |

- **Data to be discriminated:** LSL
- **Control data:** OT
- **Discrimination result:** False negative: 0/8; False positive: 9/50

**Total hits:** 84.5%

**Variables: Amino acid contents of the predicted mature portion**

| Variable: | Q | R | T | W | V | Const |
|-----------|---|---|---|---|---|-------|
| Coefficient: | -1.481 | -0.740 | -0.669 | -1.153 | -0.463 | 16.029 |
| T | 0.521 | 1.153 |

- **Data to be discriminated:** VC with ER-transferon
- **Control data:** OT
- **Discrimination result:** False negative: 0/5; False positive: 4/50

**Total hits:** 92.7%

**Variables: Amino acid contents of the regions of residues 3 to 10 (suffix 1) and 1 to 30 (suffix 3), and maximum hydrophobic moment with 165 degs for the region of residues 25 to 70 (hmx)**

| Variable: | S₃ | A₃ | C₁ | W₃ | Rₙ | Const |
|-----------|----|----|----|----|----|-------|
| Coefficient: | 0.440 | 0.241 | 0.234 | -0.247 | -0.074 | 0.081 |
| T | 0.076 | 0.151 | 0.142 | 0.281 |

- **Data to be discriminated:** Those having S-transferons in CHST and CHTM
- **Control data:** CP, MTIM, MTMX, NC, and PX
- **Discrimination result:** False negative: 1/22; False positive: 8/117

**Total hits:** 93.5%
ference is that proteins do not actually transverse the nuclear membrane at the time of entrance. It seems possible that a protein without its own nuclear localization signal (Nu-transferon) enters the nucleus via cotransport with another protein (Zhang and Padmanabhan, 1988). In addition, the Nu-transferons identified so far are not cleaved off after translocation and their exact positions in the primary sequence are not essential. This situation makes the task of finding Nu-transferons difficult.

The most common Nu-transferon is the SV40 type, which is composed of short stretches rich in basic amino acids and, often, proline residues. We attempted to find sequence patterns that can cover most known SV40-type Nu-transferons: four-residue patterns composed of basic amino acids (K or R) or of three basic amino acids (K or R) and H or P (rule 'nuc1'); and a pattern starting with P and followed within three residues by a basic four-residue segment containing three K or R residues (rule 'nuc2'). Although these patterns match most known SV40-type signals, 8 of 33 cytoplasmic proteins in the training data also had such patterns. However, if we count overlapping patterns separately, most proteins of this type seem to have more than one pattern, giving higher predictability.

Recently, another type of Nu-transferon, consisting of two interdependent basic domains, was discovered (Robbins et al., 1991). The authors proposed a simple scheme for the recognition of this bipartite signal—2 basic residues, a 10-residue spacer, and another basic 5-residue region consisting of at least 3 basic residues (rules 'nuc3' and 'nuc7')—which was rather apparent in our training data; 14 nuclear proteins and only 1 cytoplasmic protein had the pattern.

Since nuclear proteins are generally rich in basic residues, we used this heuristic in addition to the knowledge of Nu-transferons. If the sum of K and R compositions is higher than 20%, then the protein is considered to have a higher possibility of being nuclear than of being cytoplasmic (rule 'nuc4'). In addition, it might be essential to predict RNA-binding ability because it is possible for some RNPs to use targeting signals in the bound RNAs (Hamm et al., 1990). However, simple examination of the RNP consensus motif (Query et al., 1989) was not useful for discriminating nuclear proteins because many cytoplasmic proteins seem to have this motif as well (rules 'nuc5' and 'nuc6').

Peroxisomal Proteins

Some peroxisomal proteins are known to have an un-cleavable sorting signal at the C-terminus: the SKL motif (reviewed in Osumi and Fujiki, 1990). Because their sorting pathway seems to be conserved widely throughout eukaryotes (Gould et al., 1990), we included microbody proteins of various organisms except trypanosomes into our data of peroxisomal proteins. It has been further shown that the SKL motif can be tolerated in a more loose form: (S/A)(K/R/H) L (Gould et al., 1989). We searched our training data for this motif (rules 'pox1' and 'pox4'); 7 of 13 peroxisomal proteins had the motif at their C-terminus, whereas only one extracellular protein, a-amylase A2, had it at that position. In addition, all but one peroxisomal protein, in contrast to 45% of cytoplasmic and nuclear proteins, in our training data had the motif in at least one position in the entire sequence, suggesting the possibility that internal motifs may also play some role in the sorting mechanism.

To supplement the knowledge for prediction, we examined amino acid compositions of different regions of peroxisomal proteins. Although some of them have a cleavable N-terminal portion, it is not known whether it contains any information as a sorting signal (Osumi and Fujiki, 1990). We tested whether the amino acid components of the 20-residue N-terminal segment, the 20-residue C-terminal segment, or the entire sequence could be used effectively as variables of discriminant function. The discrimination accuracy was 78, 79, and 84%, respectively. Because the third case was the best, we used this as a rule for prediction (rules 'pox2', 'pox3', and 'pox5'). It can be seen from the derived discriminant function in Table 3b that the compositions F and W seem especially important. The overall net positive charge, suggested to be characteristic of peroxisomal proteins (Borst, 1986), was a more prominent feature of some nuclear proteins in our data.

Our training data of peroxisomal proteins contained a 70K membrane protein. It was unclear whether our rule could also be applied to this protein, but it had three internal SKL motifs and was positive with the discriminant score even though this protein was not included in the derivation of the function.

ER-Transferons and ER-Compartmentons

In our prediction scheme, proteins sorted along the nonselective bulk flow (Pfeffer and Rothman, 1987) are recognized as follows: First, a protein having an N-terminal signal sequence (ER-transferon) is transported to the endoplasmic reticulum (rule 'rgh1'). Second, if it has any stop-transfer signal, it is integrated into the membrane; if not and if the ER-transferon is cleaved off, it is translocated into a lumen. Third, unless it has any other signals for specific retention or commitment (compartmentons), it will be transported to the cell surface by default; a luminal protein will be secreted constitutively to the extracellular space (rule 'out1') and a membrane protein will reside at the plasma membrane. It is not necessary, however, for membrane proteins to have N-terminal ER-transferons (see below).

As in the case of Gram-negative bacterial proteins (Nakai and Kanehisa, 1991), the methods of McGeoch (1985) and von Heijne (1986) were used for the recognition of N-terminal ER-transferons (rules 'mcg1', 'mcg2', and 'rgh1'). The former, modified by us to be represented as a discriminant function, uses the information from a short N-terminal charged region and a subsequent uncharged region, whereas the latter uses the in-
transferon. In fact, most extracellular proteins predicted by the above method, and there are some soluble proteins is not easy because many membrane proteins have transmembrane segment. Type Ia proteins have a cleavable ER-transferon and are in NexoCyt orientation. Type II proteins do not have a cleavable signal and are in NcytCexo orientation. Type Ib proteins also do not have a cleavable signal but the orientation is NexoCyt. Here, all membrane proteins with more than one transmembrane segment are classified as type III. The most N-terminal segment is shadowed because the charge difference between both of its sides is thought to be important in topogenesis (types IIIa and IIIb).

formation from the region around the cleavage site. The combined use of these two methods was rather effective for detecting ER-transferons as well as bacterial signal sequences. On the one hand, most of the proteins that do not use the secretory pathway are predicted to have no ER-transferons; 94% (145 of 155) were correctly predicted. On the other hand, the evaluation of false negatives is not easy because many membrane proteins have internal start-transfer sequences that are not detected by the above method, and there are some soluble proteins sorted through pathways not requiring an ER-transferon. In fact, most extracellular proteins predicted to have no ER-transferons turned out to belong to such an exceptional class. The cleavage sites of ER-transferons predicted by von Heijne’s method were rather accurate; of the 62 training proteins whose cleavage sites are recorded in the NBRF-PIR database, 45 (73%) are correctly predicted (data not shown).

As a retention signal of ER luminal proteins (ER-comparton), the sequence motif KDEL (HDEL in yeast and some plants) at the C-terminus (rules ‘er1’ and ‘er2’) seems essential (Pelham, 1990). Although some variations of this motif are allowed in some organisms and cell types, they were not required for the discrimination of our current data. We could select all ER luminal proteins by the C-terminal KDEL motif with no false positives.

Membrane Topology

Because sequence features of compartons, except ER-compartons, seem to be weak, the reliability of detecting such weak signals would be greatly enhanced if there were additional contextual features. Thus, the topology of membrane proteins was examined with our expert system. We have adopted the latest definition of the membrane topology by Singer (1990) as shown in Fig. 1, although we do not distinguish type IV (channel) proteins from type III (polytopic) proteins.

Before the prediction of the membrane topology, we located transmembrane segments by the ‘alom’ program (Klein et al., 1985). Because it was difficult to set a single appropriate cutoff value between transmembrane and peripheral segments, we adopted a two-way approach. First, the sequence is examined with a high cutoff value of -2.0 (rule ‘alom3’); then, if more than two transmembrane segments excluding a cleavable ER-transferon are detected, the calculation is repeated with a low cutoff value of 0.5 (rule ‘alom2’). Despite this treatment, it was difficult to locate precisely all transmembrane segments in polytopic proteins, such as the seven transmembrane segments in the rhodopsin family. However, the number of predicted segments was close to that of most models of available polytopic proteins.

The sequence determinants for membrane topology have been studied extensively (von Heijne and Gavel, 1988; Hartmann et al., 1989; Parks and Lamb, 1991). Here we used the Hartmann et al. (1989) method for prediction. It is characterized by both the ‘first helix’ rule and the ‘charge difference’ rule; the overall topology is determined by the charge difference of both sides of 15 residues flanking the most N-terminal transmembrane segment (rule ‘mtp1’). This method could also be applied to usual ER-transferons and gave good results when we changed the critical value from 0 to +1.0. Moreover, it was useful for the detection of internal ER-transferons as shown below.

When our prediction scheme was applied, it showed some sequences with their predicted transmembrane segments located near the N-terminal. Although it was possible to detect uncleavable signal sequences by the combined use of the McGeoch and the von Heijne methods for Gram-negative bacterial proteins (Nakai and Kanehisa, 1991), these ER-transferons were often falsely predicted to be cleaved. To overcome this difficulty, we have added the new rule that if a type Ib protein is predicted to have an NcytCexo configuration, its most N-terminal-sided transmembrane segment is assigned as uncleavable regardless of the von Heijne score (rules ‘sig2’ and ‘sig3’).

The predicted membrane topology was not only useful for limiting the search area for compartons but also suggestive for specific prediction. For example, we have noticed a tendency for type Ib proteins to be favored at the endoplasmic reticulum, whereas type II proteins are favored at the Golgi complex and the plasma membrane (rules ‘er4’, ‘glg1’, and ‘pm3’). The topology of L-gulonolactone oxidase (OXRTGU) is not well studied. If it has a type II topology, as predicted from the charge difference, then it is an unusual protein whose transmembrane segment resides near the center of the sequence. A similar feature can be observed in the type Ib topology of a plasma membrane protein, glycophorin C (GPHUC). We represented these observations as hypothetical rules (rules ‘er5’ and ‘pm4’). In addition, cytochrome b$_5$ (S03373) is an exceptional protein whose hydrophobic
segment may not transverse the membrane (Holloway and Buchheit, 1990). As for type III proteins, we could not find any prominent sequence features that could be used to predict their localization sites, partly because of the difficulty of precisely allocating transmembrane segments; type III proteins with more than three predicted segments were tentatively predicted to be plasma membrane proteins (rule ‘pm5’).

**Signals in Cytoplasmic Tails**

Many studies so far have indicated that compartions of membrane proteins are often found at a cytoplasmic tail, a short terminal region exposed to the cytoplasmic space in type Ia, Ib, and II proteins (Fig. 1). First, two lysines positioned three and four or three and five residues from the C-terminus (rule ‘er6’) are proposed to constitute the retention signal of ER membrane proteins (Jackson et al., 1990). With the constraint of predicted membrane topology, this simple motif was specific enough to detect one ER membrane protein with no false positives. In addition, the existence of the same motif in a type III protein, HMG CoA reductase, has been noted. Since we could not determine the detailed membrane topology of polytopic proteins, we searched for the motif at the C-terminus of all polytopic proteins with more than three predicted segments. With this additional rule, the reductase was selected with one false positive, sodium channel protein I of plasma membrane.

Second, two sequence motifs, NPXY and YXRF, have been identified as signals for rapid internalization into endosomes (Chen et al., 1990; Collawn et al., 1990). The exact position of these signals seems unimportant, provided that there is a spacer from the transmembrane region. We did not distinguish internalized receptors. They are simply treated as plasma membrane proteins and these signals are used as clues for selecting plasma membrane proteins (rules ‘pm6’ and ‘pm7’). In our training data, all proteins that have a single transmembrane segment and either the NPXY or the YXRF motif at the cytoplasmic tail turned out to be plasma membrane proteins. There were two false negatives owing to the failure of predicting correct topology.

Although the nature of Golgi localization signals has not been clarified fully, recent studies suggest that it may reside in the transmembrane domain (Hurtley, 1992). In addition, the membrane-flanking sequences also seem to affect the efficiency of Golgi retention. Apart from these studies, the existence of a consensus motif, (S/T)X(E/Q)(R/K), near the hydrophobic domain (possibly the Golgi lumen) of all Golgi-localized glycosyltransferases has been pointed out, although there is no experimental proof that it is a sorting signal (Bendiak, 1990). To select Golgi proteins, we searched for this motif in the regions flanking the type II transmembrane segment (rule ‘glt1’). Although two Golgi proteins were selected with no false positives, two other Golgi proteins satisfying this condition failed because their N-terminal transmembrane segment was predicted to be cleaved. In the E1 glycoprotein of coronavirus, the first transmembrane domain is required for its retention (Machamer and Rose, 1987). However, since it is the only example of such a signal, we did not incorporate this knowledge. As stated later, a tyrosine residue in the cytoplasmic tail is also important for sorting lysosomal membrane proteins.

**Lipid Anchors**

We considered two distinct cases in which chemical modifications have primary importance in protein sorting. One is the case in which proteins are anchored at a membrane with covalently bound lipid moieties and the other is the case of lysosomal proteins.

There are several ways of lipid anchoring, as illustrated in Fig. 2 (Schultz et al., 1988). Among them, all proteins linked to glycosyl-phosphatidylinositol (GPI) seem to be localized at the extracellular surface of the plasma membrane (Ferguson and Williams, 1988; Cross, 1990). Thus, if we can predict this modification, we can simultaneously predict the localization. Although the signal that leads to the GPI attachment (GPI-modon) is not fully understood, all precursors seem to be type Ia proteins and have cleavable C-terminal sequences. Moreover, their cytoplasmic tails, if present at all, are predicted to be very short. These features were sufficient for the discrimination of our training data (rule ‘pm9’); 12 of 16 members were correctly predicted to have GPI anchors by the criterion of a type Ia protein that has a short cytoplasmic tail (within 10 residues). The false predictions were caused by the failure of predicting topology. There were no type Ia proteins with short cytoplasmic tails in other localization sites.

In eukaryotic cells, palmitic and myristic acids are observed to be bound directly to proteins. However, recent studies suggest that many of them may not take part in lipid anchoring (McIlhinney, 1990; Resh, 1990). Thus, although we could make an efficient discriminant function for potential N-myristoylated sequences based on the substrate specificity of yeast N-myristoyltransferase (Towler et al., 1988), it did not work well for the prediction of localization sites (rule ‘pm8’).
There is another type of lipid linking known as isoprenylation or farnesylation (Maltese, 1990). This modification requires a CaaX motif at the C-terminus; “a” denotes an aliphatic amino acid. Since isoprenylated proteins have been found in the plasma membrane and in the nuclear envelope, another signal is needed for correct sorting (Hancock et al., 1990). In the case of nuclear lamin A, it seems to be an usual Nu-transferon (Holtz, 1989). Scanning our training data with the C-terminal CaaX motif revealed two false proteins. With an additional rule that isoprenylated proteins do not have any transmembrane segments or ER-transferons, only one of them, lipocortin I, remained (rules ‘caax0’, ‘caax1’, and ‘caax2’).

**Lysosomal and Vacuolar Proteins**

Vacuoles, found in plant and yeast cells, have diverse functions, one of which is analogous to that of mammalian lysosomes. Based on our current understanding of the sorting mechanisms, we separated the prediction category into three sites: the lumen of lysosomes in animal cells, the membrane of lysosomes in animal cells, and the vacuoles in yeast and plant cells.

There are at least two distinct mechanisms for sorting lysosomal proteins (Kornfeld and Mellman, 1989). One is dependent on the posttranslational modification of mannose 6-phosphate, used in soluble enzymes. The other is dependent on a tyrosine residue at a particular position in the cytoplasmic tail (Williams and Fukuda, 1990). The formation of mannose 6-phosphate seems conformation-dependent and some sequence segments could contribute to form a recognition domain (Baranski et al., 1990). Thus, we postulated that a soluble lysosomal protein should have a cleavable ER-transferon, have no transmembrane segments, but have at least two potential N-glycosylation sites, i.e., NX(S/T) motifs, in the mature portion (rule ‘lys3’). Since many extracellular proteins (23 of 50) share these features, we examined the differences in amino acid composition by stepwise discriminant analysis (rule ‘lys1’). The predicted region of the ER-transferon was excluded from the calculation of amino acid composition. As shown in Table 3c, up to five variables were chosen because of the small size of the training data. With the derived discriminant function and the criterion of the potential glycosylation site, only 3 of 50 proteins were falsely discriminated.

The discrimination of lysosomal membrane proteins was accomplished as follows. First, we predicted the membrane topology of a protein; then, if it was a type Ia protein and if it contained a GY motif in the cytoplasmic tail within 17 residues of the boundary with the membrane, it was predicted to be a lysosomal membrane protein (rule ‘lyse2’). This procedure was sufficient for discriminating the three proteins from all the training data. Interestingly, they were also positive in the discriminant score for lysosomal soluble proteins. It may be that most parts of the sequence are exposed to the lumen and they are similar in amino acid composition.

The vacuolar sorting signals have been studied in yeast (Rothman et al., 1989) and in plants (Chrispeels and Raikhel, 1992). It is likely that most plant and yeast cells have a common sorting pathway for vacuolar proteins, though they may have other diverse pathways. Some vacuolar proteins have their signals in the preregion, which essentially looks like an ER-transferon, and the presequence, which is needed for specific recognition at the Golgi complex. However, no common sequence features have been discovered in the latter. In addition, there is a protein that does not use even the secretory pathway (Yoshihisa and Anraku, 1990), as well as a membrane protein that uses a distinct pathway (Klionsky and Emr, 1990). We performed a discriminant analysis between vacuolar and extracellular proteins from the amino acid composition of the sequence excluding the presequence (rule ‘vacl’ and ‘vac2’), as shown in Table 3d. The vacuolar sequences that do not have an ER-transferon at the N-terminus were excluded from the analysis because they seemed to have distinct characteristics. Indeed, they could not be correctly discriminated by the derived function. The selected variables were totally different from the ones selected in the analysis of lysosomal proteins. With only a single amino acid content (K), 12 of 13 lysosomal and vacuolar proteins could be correctly distinguished (data not shown), although in our expert system the distinction is made by the organism name.

**Chloroplast Proteins**

For plant proteins, the chloroplast is also a possible localization site. We postulated that all stromal proteins and thylakoid membrane proteins have the same kind of stroma-targeting signal (S-transferon; Hand et al., 1989; Keegstra et al., 1989) and searched for an efficient method of detecting it. Based on the observation that S-transferons have three distinct regions (von Heijne et al., 1989), we performed a stepwise discriminant analysis using as variables the amino acid compositions of the two N-terminal segments, residues 3 to 10 and residues 1 to 30, and the position and the amplitude of maximum hydrophobic moment of 165° for residues 25 to 70 (rule ‘chts0’). As shown in Table 3e, the result of self-discrimination was rather good. In contrast, the loosely conserved motif (V/1)X(A/C)A of S-degrons (Gavel and von Heijne, 1990a) was not as effective. We have also used the knowledge that in most (90.5%) chloroplast targeting peptides the residue at position 2 is alanine (rule ‘chnm02’).

Like some mitochondrial proteins, thylakoid luminal proteins are known to have a targeting sequence of bipartite structure; the N-terminal half is functionally equivalent to an S-transferon and the C-terminal half is required for translocation from the stroma into the thylakoid lumen (S/T-transferon). These proteins showed positive scores with the discriminant function for S-transferons. To detect the latter half of the bipartite signal, we employed two methods. One was the weight ma-
matrix of Howe and Wallace (1990) (rule 'ch2nd2'), which
was derived from the data of thylakoid luminal proteins
by the same method as von Heijne's (1986). The weight
matrix could locate all the cleavage sites of our data, but
it was not sufficient for discriminating thylakoid lu-
minal proteins from others. It is probable that the
weight matrix could only detect the signal recognized by
a specific protease and could not detect the S/T-transfer-
feron itself. The second method we used was the 'apolar'
algorithm applied to the limited region of residues 40 to
90 (rule 'ch2nd1'). From the 'apolar' score and the length
of the apolar region, most thylakoid luminal proteins
could be discriminated from other chloroplast proteins.

Thylakoid membrane proteins were discriminated by
the 'alam' program. The remainder of chloroplast pro-
teins were regarded as stromal proteins (rules 'chtm' and
'chat').

Organization of Knowledge

The results described above were integrated into a set
of rules in our expert system. The list of core rules for
reasoning steps 2 and 3 are summarized in Table 4. The
number of rules is currently 80, excluding those for bacte-
rial sequences. A simplified reasoning tree is given in
Fig. 3. For each possible site, the reasoning procedure is
performed roughly following the tree. Each node is a
checkpoint for a certain sequence feature and certainty
factors are modified according to its result. In principle,
the path of reasoning should follow the real pathway of
sorting in vivo. It is most probable that the first recogni-
tion process for a nascent polypeptide is mediated by an
ER-transferon. If the polypeptide has an ER-transferon,
it will be committed to a vesicle-mediated pathway and
its final localization site will be determined by its com-
parton signals. Otherwise, it will be sorted according to
other transferon signals. If it has no signals at all, it will
become a cytoplasmic protein.

Apart from biological reality, we had to modify this
simple scheme in some minor points. One was the treat-
ment of internal ER-transferons, which cannot be effec-
tively detected at present. For example, since most
plasma membrane proteins do not have N-terminal ER-
transferons, they must be examined in the context of the
cytoplasmic pathway. Another was the independence of
evaluations. As described, the possibility of being sorted
to each localization site is evaluated one by one. In gen-
eral, different evaluations are independent except for
those for cytoplasmic proteins. Finding that one protein
is unlikely to be sorted to one site does not usually raise
the possibility that it may be sorted to any other specific
site. The calculated possibilities are stored as certainty
factors and the site that has the highest certainty is se-
lected as the most probable site. However, we had to
break this principle for some localization sites with
poorly characterized sorting signals (e.g., nuclei and ly-
sosomes), making the evaluation order-dependent, i.e.,
well-characterized sites first.

The assignment of certainty factors was one of the
more difficult aspects. For some rules, the distribution of
score values over the training data was examined. The
score values were usually divided into three classes, posi-
tive, not clear, and negative, and the certainty factor was
assigned for each class. By trial and error, certainty fac-
tors were further adjusted considering the whole predic-
tion accuracy.

Prediction Accuracy

As summarized in Table 5, 66% of the training data
and 59% of the testing data were correctly predicted.
Since the testing data were selected from the localiza-
tion sites that involved more than 10 members, the com-
position of the testing data was not proportional to that
of the training data. It is difficult to compare this predic-
tion accuracy to other standards, but since each protein
has a possibility of 14 to 17 localization sites depending
on the organism, a random guess would result in less
than 10% accuracy. If we simply assume that all proteins
in the testing data belong to the largest site, the extra-
cellular space, the value is 21%. It should be noted that
the upper limit of the value is apparently lower than
100% because of the presence of exceptional sorting
pathways. There was not a marked difference in the pre-
diction accuracy between animal proteins and plant/
yeast proteins.

In the training data, proteins at the lumen of ER and
at the lysosomal membrane were perfectly discrimi-
nated. However, they are very small in size. The next
well-predicted classes were the stroma of chloroplasts,
the matrix of mitochondria, and the plasma membrane
(integral). GPI-anchored proteins had the highest pre-
dictability in the testing data, which was actually higher
than that in the training data. A significant decrease in
the predictability was observed for the stroma of chloro-
plasts, the matrix of mitochondria, and the peroxisomes,
all of which were largely dependent on the results of
discriminant analysis, which apparently had the danger
of overfitting to the training data. The fine structure of
organelles, such as the mitochondrial inner membrane,
was relatively difficult to predict. If we simply consider
all proteins in each organelle as one group, the discrimi-
nation accuracy in the training data is 66% for the mito-
chondrion and 86% for the chloroplast. The prediction
accuracy for the combined mitochondrial proteins be-
comes 64% in the testing data.

Table 5 also contains the number of false positives for
each site. We noticed that many exceptional proteins are
falsely predicted to be cytoplasmic proteins because they
do not have usual targeting signals or other features.
Many testing proteins were falsely predicted to be per-
oxosomal proteins, which implies that current knowl-
edge of peroxisomal targeting signals is not specific
enough.

For practical use, alternative localization sites with
lower certainty factors are also suggestive. When we
took two sites with the two best certainty factors, the
probability of one of them being correct was 71.9% for
the training data and 69.8% for the testing data.
TABLE 4
List of Core Rules in the Knowledge Base

| Rule   | Description                                      | Rule   | Description                                      |
|--------|--------------------------------------------------|--------|--------------------------------------------------|
| einit  | Prepare for the calculations of GvH, McG, and ALOM.| er1    | If it might be an ER luminal protein, the existence of the KDEL (HDEL in yeast) motif around the C-terminus must be examined. |
| gvh1   | If not yet, check the ER-transferon by GvH and store the result. | er2    | If an ER luminal protein, it is likely to have the motif, an ER-transferon, and no TMSs. |
| mcg1   | If not yet, check the ER-transferon by McG and store the result. | er4    | If an ER membrane protein, it may be a type IIb protein whose TMS locates within the 30% region from the N-terminus. |
| mcg2   | If the result of 'mcg1' is obtained, calculate and store the discriminant value. | er5    | If an ER membrane protein, it may be a type II protein whose TMS locates within the 70% region from the C-terminus. |
| alom2  | If not yet, find the TMS by ALOM (threshold 0.5) and store the result. | er6    | If an ER membrane protein, it may be a type Ia protein with a cytoplasmic tail of appropriate length containing a retention signal. |
| mtop1  | If there is at least one TMS, calculate the charge difference around the most N-terminal TMS by MTOP. | er7    | If an ER membrane protein, it may be a type IIIa or IIIb protein but the probability is relatively low. |
| sig2   | If the charge difference predicts the NcytCexo orientation, determine whether there is an ER-transferon and whether it is cleavable from previous results. | er8    | If an ER membrane protein, it may have an uncleavable ER-transferon. |
| sig3   | If the charge difference predicts the NexoCyt orientation, determine whether there is an uncleavable ER-transferon from previous results. | out1   | If an extracellular protein, it has a cleavable ER-transferon and does not have TMSs at all. |
| alom3  | If the number of TMS is less than 3 in the mature sequence, change the threshold of ALOM to -2.0. | pm1    | If a plasma membrane protein, its topology may be type Ia. |
| alom4  | If possible, output the final result of ALOM considering the possibility of cleavage and the variable threshold value. | pm2    | If a plasma membrane protein, its topology may be type Ib. |
| mtop2  | If it has a cleavable ER-transferon and one more TMS, it is type Ia. | pm3    | If a plasma membrane protein, its topology may be type II, its TMS locates within the 40% region from the N-terminus, and there is no M-transferon. |
| mtop3  | If it has one TMS, does not have a cleavable ER-transferon, and the charge balance predicts NcytCexo, it is type II. | pm4    | If a plasma membrane protein, its topology may be type Ib, and its TMS locates within the 40% region from the C-terminus. |
| mtop4  | If it has one TMS and does not have a cleavable ER-transferon and the charge balance predicts NexoCct, it is type Ib. | pm5    | If a plasma membrane protein, its topology may be type IIIa or IIIb and if the number of TMSs exceed 10, the possibility rises. |
| mtop5  | If it has a cleavable ER-transferon and more than one TMSs, it is type IIIa. | pm6    | If a plasma membrane protein, its topology may be type Ia, II, or Ib with a NPXY motif in the cytoplasmic tail. |
| mtop6  | If it does not have a cleavable ER-transferon and has more than one TMSs, it is type IIIa or IIIb according to the charge balance. | pm7    | If a plasma membrane protein, its topology may be type Ia, II, or Ib with a YXR motif in the cytoplasmic tail. |
| aac1   | If the examination of ER-transferon is finished, calculate and store the amino acid composition of the mature portion. | pm8    | If a plasma membrane protein, its N-terminus may be myristylated. |
| rgh1   | If it has an ER-transferon, the sites on the vesicular pathway have some possibility of being selected. | pm9    | If a plasma membrane protein, its topology may be type Ia and the length of the tail is less than 10; that is, it may be GPI-anchored. |
| exgavl | If it might be a mitochondrial protein, examine the possible cleavage site of M-transferon by GAVEL. | caax0  | If it might be a plasma membrane or nuclear protein and if it has no TMSs, the existence of the CaaX motif should be searched for at the C-terminus. |
| mdisc  | If it might be a mitochondrial protein, examine the existence of M-transferon from the AAC of 20 N-terminal residues. | caax1  | If a plasma membrane protein, it might have the C-terminal CaaX motif but does not have TMSs or Nu-transfomers. |
| mntod  | If it has a positive possibility of having an M-transferon and does not have an ER-transferon, it may be targeted to a mitochondrion. | caax2  | If a nuclear protein, it might have the C-terminal CaaX motif and Nu-transfomers but does not have TMSs. |
| chpm   | If it might be a chloroplast protein, calculate the maximum hydrophobic moment in the segment from res. 26 to 70. | glg1   | If a Golgi protein, it is likely a type II protein with the (S/T)X(E/Q)(H/R) motif near the TMS. |
| chlaa1 | If it might be a chloroplast protein, calculate the AAC of res. 3 to 10. | glg2   | If a Golgi protein, its topology might be type IIIa or IIIb. |
| chlaa2 | If it might be a chloroplast protein, calculate the AAC of res. 1 to 30. | lys1   | If it might be a lysosomal protein, the discriminant score must be calculated from the amino acid composition of the mature portion. |
| chldisc| If it might be a chloroplast protein, examine the existence of S-transferon from the results of chpm, chlaa1, and chlaa2. | lys2   | If a lysosomal membrane protein, its topology should be type Ia with the GY motif in the tail near the TMS and have characteristic AAC. |
| chlmod | If it has a positive possibility of having an S-transferon and does not have an ER-transferon, it may be targeted to a chloroplast. | lys3   | If a lysosomal luminal protein, it should have a cleavable ER-transferon, no TMSs, at least two N-glycosylation motifs, and deviated AAC. |
In this work, various experimental and computational observations on protein-sorting signals were organized into a consistent knowledge base that can be used to interpret unknown sequences. The knowledge base was realized as a collection of if-then rules (production rules), and utilized for machine inference based on standard techniques in artificial intelligence. Our system turned out to be flexible enough to incorporate diverse types of sorting signals and could contain ambiguous observations and working hypotheses. Furthermore, its performance could be evaluated by the predictability applied to unknown sequences. There are, however, still problems to be overcome, especially in knowledge acquisition and maintenance. Although rule-based representation is simple enough to update each piece of knowledge, it is time-consuming to revise certainty factors, which requires a global optimization. It is desirable that they be automatically optimized, say, by the neural network method.

One of the difficulties in constructing our knowledge base was the assignment of a single appropriate localization site for each protein. There are many proteins whose localization sites are not confined to a single space. For example, some proteins like NF-κB change their localization sites in a regulated manner (Hunt, 1989). Ribosomal proteins are first sorted from the cytoplasm to the nucleus according to their Nu-transferons, but after their assembly they are transported back to the cytoplasm possibly by a specific mechanism (Underwood and Fried, 1990). We defined ribosomal proteins as nuclear proteins. Future progress in understanding these mechanisms may enable us to make a more detailed prediction of multiple localization sites. Another difficulty was the presence of nonconservative and specific sorting pathways. Some extracellular proteins such as...
as interleukins 1α and 1β do not have N-terminal ER-transferons and are sorted through distinct pathways (Rubartelli et al., 1990). There is also a nuclear transport by specific interaction with a protein with an Nu-transferon (Zhao and Padmanabhan, 1988). Although it may be possible to introduce exception rules dealing with such specific sorting pathways, it is very likely that the ratio of specific sorting mechanisms involved in the total system determines the upper limit of our prediction accuracy.

To evaluate the performance of our expert system, it was applied to the testing data not used for its construction. The prediction accuracy was about 60% when more than 10 sites were distinguished. This result should be considered much better than the accuracy of widely used protein secondary structure prediction, which is also about 60% but which distinguishes only three states (helix, sheet, and coil). Apparently, most sorting signals are confined to limited regions of sequences; if there were many signals that depend on peptide conformation, like M6P-modons, the prediction accuracy would have been much worse. In addition, our combinational approach has advantages over individual prediction of each localization site. For example, the control data used in the von Heijne et al. analysis of M-transferons (1989) were adopted from the mature portion of mitochondrial proteins. It is more natural to use the N-terminal regions of proteins that compete with mitochondrial ones in real cellular recognition processes. In this respect, it may be an important observation that proteins with internal ER-transferons seem to be similar to mitochondrial proteins in N-terminal amino acid composition.

Since experimental knowledge on sorting signals was not always complete, we had to rely on computational results characterizing proteins of given localization sites, which sometimes may not be directly related to sorting signals. The stepwise discriminant analysis could be effectively used for extracting characteristic amino acid components (Table 3). The derived discriminant function for M-transferons shows that the 20-residue segment at the N-terminus is rich in R, but poor in P and acidic residues. The function for S-transferons shows that the segment of residues 1 to 30 is rich in S and A, and the segment of residues 3 to 10 is rich in C. The first variables in both cases were in agreement with the von Heijne et al. observation (1989), which had been based on different control data. The discriminant function for peroxisomal proteins shows the abundance in aromatic F and W, as well as in Y and H, although its biological significance is unclear. It was also found that lysosomal proteins were poor in K, whereas vacuolar proteins were poor in Q, when compared with extracellular proteins. The content of K was almost sufficient for discriminating lysosomal proteins from vacuolar proteins. It is interesting because lysosomes are acidic organelles.

Several hypotheses were also included in the knowledge base to supplement the experimentally proven knowledge. One hypothesis was that the charge difference between both sides of the most N-terminal transmembrane segment would inhibit its cleavage. It is a natural hypothesis because a reversed orientation along the membrane may be difficult for the signal peptidase to approach. However, some ER-transferons of extracellular proteins had an unusual charge balance, which should be further studied. Another hypothesis was that there would be a preference of type Ib or II proteins depending on the localization site. Usually type Ib proteins have larger cytoplasmic domains and type II proteins have larger extracytoplasmic domains. It is possible that
TABLE 5

Results of Prediction

| Localization site                        | Number of hits | Number of data | Number of false positives | Percentage true positives |
|------------------------------------------|----------------|----------------|---------------------------|--------------------------|
| Chloroplast stroma                       | 13             | 15             | 3                         | 86.7                     |
| Chloroplast thylakoid membrane           | 3              | 7              | 1                         | 42.9                     |
| Chloroplast thylakoid space              | 4              | 6              | 0                         | 66.7                     |
| Cytoplasm                                | 20             | 33             | 22                        | 66.0                     |
| Endoplasmic reticulum lumen              | 4              | 4              | 0                         | 100.0                    |
| Endoplasmic reticulum membrane           | 7              | 13             | 0                         | 53.9                     |
| Golgi complex                            | 2              | 6              | 0                         | 33.3                     |
| Lysosomal lumen                          | 5              | 8              | 6                         | 62.5                     |
| Lysosomal membrane                       | 3              | 3              | 0                         | 100.0                    |
| Mitochondrial inner membrane             | 3              | 10             | 1                         | 30.0                     |
| Mitochondrial intermembrane space        | 1              | 5              | 3                         | 20.0                     |
| Mitochondrial matrix space               | 17             | 21             | 7                         | 81.0                     |
| Mitochondrial outer membrane             | 1              | 2              | 0                         | 50.0                     |
| Nucleus                                  | 25             | 43             | 5                         | 58.1                     |
| Extracellular space                      | 36             | 50             | 8                         | 72.0                     |
| Plasma membrane (GPI-anchored)           | 10             | 14             | 0                         | 71.4                     |
| Plasma membrane (integral)               | 28             | 33             | 14                        | 84.9                     |
| Peroxisome                               | 9              | 13             | 13                        | 69.2                     |
| Vacuole                                  | 5              | 9              | 0                         | 55.6                     |
| **Total**                                | **196**        | **295**        | **99**                    | **66.4**                 |

(a) Training data

(b) Testing data

| Localization site                        | Number of hits | Number of data | Number of false positives | Percentage true positives |
|------------------------------------------|----------------|----------------|---------------------------|--------------------------|
| Chloroplast stroma                       | 3              | 6              | 0                         | 50.0                     |
| Cytoplasm                                | 7              | 14             | 7                         | 50.0                     |
| Endoplasmic reticulum membrane           | 4              | 6              | 3                         | 66.7                     |
| Mitochondrial inner membrane             | 1              | 5              | 1                         | 20.0                     |
| Mitochondrial matrix space               | 5              | 9              | 7                         | 55.6                     |
| Nucleus                                  | 12             | 19             | 3                         | 63.2                     |
| Extracellular space                      | 13             | 22             | 2                         | 59.1                     |
| Plasma membrane (GPI-anchored)           | 5              | 6              | 1                         | 83.3                     |
| Plasma membrane (integral)               | 11             | 14             | 7                         | 78.6                     |
| Peroxisome                               | 2              | 5              | 11                        | 40.0                     |
| Other site                               |                |                | 1                         |                          |
| **Total**                                | **63**         | **106**        | **43**                    | **59.4**                 |

ER membrane proteins, which favor type Ib, possess large cytoplasmic domains for their function, whereas for plasma membrane proteins extracellular domains are important. In fact, a majority of plasma membrane proteins have type Ia topology with short cytoplasmic tails. More samples should be collected for further consideration.

Although we attempted to incorporate the most up-to-date knowledge, there are certain areas for improvement. In the current system, knowledge of polarized sorting in the plasma membrane (Simons and Wandinger-Ness, 1990) has not been included. Knowledge of cell type will also be required for precise prediction. The distinction between constitutive and regulated secretions will also become possible. The prediction of intraorganellar sorting is still insufficient. One reason is that their member proteins may not have been allocated exactly by experiments. Another reason is that in both mitochondria and chloroplasts, membrane proteins have relatively low hydrophobicity, reflecting the different natures of organelle membranes. Proteins with higher hydrophobicity may have a higher possibility of being falsely recognized by signal recognition particles. In practice, it may be effective to change the threshold of discrimination of their transmembrane segments. Many comparten signals are still open to future study. Above all, the integration of various kinds of degron (degradation) and modon (modification) signals is a challenging subject, enabling the prediction of the overall metabolic fate of proteins.

It seems evident that any functional implication that can be derived from determining sequence data will become even more necessary with the development of large-scale sequencing projects. One such example has been reported recently (Adams et al., 1992). Of the 2375 partial cDNA sequences that were newly determined, 83% were not related to known sequences in the databases. Our goal is to provide additional clues to the characterization of such unknown sequence data for further investigations. However, there is one difficulty in applying our work to the results of partial cDNA data; our method requires full-length sequences for input because
missing parts can involve important targeting signals. Nevertheless, it may be used to find some signals. Sequencing N-terminal halves seems more preferable to C-terminal halves for our analyses since many trans-membrane N-terminal halves are found in N-terminal parts.

ACKNOWLEDGMENTS

We thank Dr. Koreaki Ito for critically reading the manuscript. This work was supported by a grant from the Human Frontier Science Program.

REFERENCES

Adams, M. D., Dubnick, M., Kerlavage, A. R., Moreno, R. Kelley, J. M., Utterback, T. R., Nagle, J. W., Fields, C., and Venter, J. C. (1992). Sequence identification of 2,575 human brain genes. Nature 359: 632-640.

Baker, K. P., and Schatz, G. (1991). Mitochondrial proteins essential for viability mediate protein import into yeast mitochondria. Nature 349: 205-208.

Baranski, T. J., Faust, P. L., and Kornfeld, S. (1990). Generation of a lysosomal enzyme targeting signal in the secretory protein pepsinogen. Cell 63: 281-291.

Barker, W. C., George, D. G., and Hunt, L. T. (1990). Protein sequence database. Methods Enzymol. 183: 91-49.

Benda, K. (1980). A common peptide stretch among enzymes localized to the Golgi apparatus: Structural similarity of Golgi-associated glycosyltransferases. Biochem. Biophys. Res. Commun. 170: 879-882.

Borst, P. (1986). How proteins get into microbodies (peroxisomes, glyoxysomes, glycosomes). Biochim. Biophys. Acta 866: 179-203.

Chen, W.-J., Goldstein, J. L., and Brown, M. S. (1990). NPXY, a recognition motif for endocytosis. Cell 63: 1061-1072.

Chrispeels, M. J., and Raikhel, N. V. (1992). Short peptide domains target proteins to plant vacuoles. Cell 68: 613-616.

Collawn, J. F., Stangel, M., Kuhn, L. A., Esekogwu, V., Jing, S., Trowbridge, I. S., and Tainer, J. A. (1990). Transferrin receptor internalization sequence YYRF implicates a tight turn as the structural recognition motif for endocytosis. Cell 63: 1061-1072.

Cross, G. A. M. (1990). Glycolipid anchoring of plasma membrane proteins. Annu. Rev. Cell Biol. 5: 1-39.

Dahllof, B., Wallin, M., and Kivist, S. (1991). The endoplasmic reticulum retention signal of the E3/19K protein of adenovirus-2 is mitochondrial targeting peptides. Annu. Rev. Biochem. 59: 969-977.

Hamm, J., Durynskiewicz, E., Tahara, S., and Mattaj, I. W. (1990). The trimethylguanosine cap structure of U1 snRNA is a component of a bipartite nuclear targeting signal. Cell 62: 569-577.

Hancock, J. F., Paterson, H., and Marshall, C. J. (1990). A polybasic domain or palmitoylation is required in addition to the CAAX motif to localize p21ras to the plasma membrane. Cell 63: 133-138.

Hartl, F.-U., and Neupert, W. (1990). Protein sorting to mitochondria: Evolutionary conservations of folding and assembly. Science 247: 930-938.

Hartmann, M. A., and Williams, A. F. (1988). Cell-surface anchoring of surface proteins. Annu. Rev. Biochem. 57: 285-320.

Holloway, P. W., and Buchelt, C. (1990). Topography of the membrane-binding domain of cytochrome b(6) in lipids by Fourier-transform infrared spectroscopy. Biochimica et Biophysica Acta 108: 285-291.

Holtz, D., Tanaka, R. A., Hartwig, J., and McKeon, F. (1989). The CaaX motif of lamin A functions in conjunction with the nuclear localization signal to target assembly to the nuclear envelope. Cell 59: 969-977.

Horta, C. J., and Wallace, T. P. (1990). Prediction of leader peptide cleavage sites for polypeptides of the thylakoid lumen. Nucleic Acids Res. 18: 3417.

Hunt, T. (1989). Cytoplasmic anchoring proteins and the control of nuclear localization. Cell 59: 949-951.

Klionsky, D. J. (1989). A polybasic nuclear localization signal. Trends Biochem. Sci. 17: 2-3.

Jackson, M. R., Nilsson, T., and Peterson, P. A. (1990). Identification of a conserved motif for retention of transmembrane proteins in the endoplasmic reticulum. EMBO J. 9: 3103-3106.

Kekstra, K., Olsen, L. J., and Theg, S. M. (1989). Chloroplastic precursor proteins and their transport across the envelope membranes. Annu. Rev. Plant Physiol. Plant Mol. Biol. 40: 471-501.

Klein, P., Kanehisa, M., and DeLisi, C. (1985). The detection and classification of membrane-spanning proteins. Biochim. Biophys. Acta 815: 468-476.

Kikoinsky, D., and Emr, S. D. (1990). A new class of lysosomal/vacuolar protein sorting signals. J. Biol. Chem. 265: 5349-5352.

Kornfeld, S., and Mellman, I. (1989). The biogenesis of lysosomes. Annu. Rev. Cell Biol. 5: 485-525.

Machamer, C. E., and Rose, J. K. (1987). A specific transmembrane domain of a coronavirus E1 glycoprotein is required for its retention in the Golgi region. J. Cell Biol. 105: 1205-1214.

Maltese, W. A. (1990). Posttranslational modification of proteins by isoprenoids in mammalian cells. FASEB J. 4: 3319-3328.

McGeeoch, D. J. (1985). On the predictive recognition of signal peptide sequences. Virus Res. 3: 271-286.

McClimhiney, R. A. J. (1990). The fats of life: The importance and function of protein acylation. Trends Biochem. Sci. 15: 387-391.

Nakai, K., and Kanehisa, M. (1988). Prediction of in vivo modification sites of proteins from their primary structures. J. Biochem. (Tokyo) 104: 693-699.

Nakai, K., and Kanehisa, M. (1991). Expert system for predicting protein localization sites in Gram-negative bacteria. Proteins 11: 95-110.

Osumi, T., and Fujiki, Y. (1990). Topogenesis of peroxisomal proteins. BioEssays 12: 217-222.

Parks, G. D., and Lamb, R. A. (1991). Topology of Eukaryotic type II membrane proteins: Importance of N-terminal positively charged residues flanking the hydrophobic domain. Cell 64: 777-787.

Pelham, H. R. B. (1990). The retention signal for soluble proteins of the endoplasmic reticulum. Trends Biochem. Sci. 15: 483-486.

Pfeffer, S. R., and Rothman, J. E. (1987). Biosynthetic protein trans-
port and sorting by the endoplasmic reticulum and Golgi. *Annu. Rev. Biochem.* **56**: 829–852.

Query, C., Bentley, R. C., and Keene, J. D. (1989). A Common RNA Recognition motif identified within a defined U1 RNA binding domain of the 70K U1 snRNP protein. *Cell* **57**: 89–101.

Resh, M. D. (1990). Membrane interactions of pp60^src^: A model for myristylated tyrosine protein kinases. *Oncogene* **5**: 1437–1444.

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

Rothman, J. H., Yamashiro, C. T., Kane, P. M., and Stevens, T. H. (1989). Protein targeting to the yeast vacuole. *Trends Biochem. Sci.* **14**: 347–350.

Rubartelli, A., Cozzolino, F., Talio, M., and Sitia, R. (1990). A novel secretory pathway for interleukin-1β, a protein lacking a signal sequence. *EMBO J.* **9**: 1503–1510.

Schultz, A. M., Henderson, L. E., and Oroszlan, S. (1988). Fatty acylation of proteins. *Annu. Rev. Cell. Biol.* **4**: 611–647.

Silver, P. A. (1991). How proteins enter the nucleus. *Cell* **64**: 489–497.

Simons, K., and Wandinger-Ness, A. (1990). Polarized sorting in epithelia. *Cell* **62**: 207–210.

Singer, S. J. (1990). The structure and insertion of integral proteins in membranes. *Annu. Rev. Cell Biol.* **6**: 247–296.

Tanaka, H., and Shimoj, Y. (1987). “Expert System Kochiku-no-hōhō,” Personal Media, Inc. [in Japanese]

Towler, D. A., Gordon, J. I., Adams, S. P., and Glaser, L. (1988). The biology and enzymology of eukaryotic protein acylation. *Annu. Rev. Biochem.* **57**: 69–99.

Underwood, M. R., and Fried, H. M. (1990). Characterization of nuclear localizing sequences derived from yeast ribosomal protein L29. *EMBO J.* **9**: 91–99.

Varshavsky, A. (1991). Naming a targeting signal. *Cell* **64**: 13–15.

Verner, K., and Schatz, G. (1988). Protein translocation across membranes. *Science* **241**: 1307–1313.

von Heijne, G. (1986). A new method for predicting signal sequence cleavage sites. *Nucleic Acids Res.* **14**: 4683–4690.

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

von Heijne, G., Steppuhn, J., and Herrmann, R. G. (1989). Domain structure of mitochondrial and chloroplast targeting peptides. *Eur. J. Biochem.* **180**: 535–545.

Waterman, D. A. (1986). “A Guide to Expert Systems,” Addison-Wesley, Reading, MA.

Williams, M. A., and Fukuda, M. (1990). Accumulation of membrane glycoproteins in lysosomes requires a tyrosine residue at a particular position in the cytoplasmic tail. *J. Cell Biol.* **111**: 955–966.

Yoshihisa, T., and Anraku, Y. (1990). A novel pathway of import of α-mannosidase, a marker enzyme of vacuolar membrane, in Saccharomyces cerevisiae. *J. Biol. Chem.* **265**: 22418–22425.

Zhao, L.-J., and Padmanabhan, R. (1988). Nuclear transport of adenovirus DNA polymerase is facilitated by interaction with preterminal protein. *Cell* **64**: 13–15.