Bull. Inst. Chem. Res., Kyoto Univ., Vol. 69, No. 3, 1991

REVIEW

Predicting Various Targeting Signals in Amino Acid Sequences

Kenta NAKAI*

Received June 24, 1991

KEY WORDS: Prediction/ Protein targeting signal/ Sequence analysis/ *In-vivo* protein fate

1. INTRODUCTION

One of the ultimate goals of modern biology is to clarify how a variety of biological functions are encoded as genetic information. Although experimental works have been the major contributing force for that clarification, theoretical or computational works to interpret genetic information have also been tried and have made significant contributions. It is expected that the role of such studies will become more important because both our knowledge for interpretation and the sequence data to be examined are increasing in a surprising rate. That is, computers should be used for the storage of massive sequence data, for the construction of a consistent knowledge base, for the examination of the generality of experimental observation, and for the creation of hypotheses or predictions to be verified by experiments.

There are many ways that genetic information is encoded in DNA sequences. Some information is confined to a local region of the sequence while other information is distributed in many segments or the entire sequence. Generally, the theoretical recognition of the former information is easier than that of the latter. The present limitation of protein secondary structure prediction methods seems to imply this: secondary structures are determined from their local sequence to some extent and it is rather easy to predict them to that level, however, it is very difficult to incorporate the long-range interaction effects of sequences (for a recent review of secondary structure prediction, see 1). Similarly, the prediction of protein functions using the sequence motifs is based on the observation that the functionally most important region is strongly conserved in a few short segments of the sequence.[2 3)] Although it is easy to search such motifs for unknown sequences, it seems rather difficult to predict protein functions from combining multiple existing motifs.

In this review, I want to focus on another kind of sequence information, the

* 中井謙太 : Laboratory of Molecular Design for Physiological Functions, Institute for Chemical Research, Kyoto University, Uji, Kyoto-fu 611

signals which determine various *in-vivo* fates of nascent proteins, that is, sorting signals for the localization in cells, degradation signals, and (some typical examples of) modification signals. All of these signals are likely to have a common feature: They are recognized by specific molecular mechanism in cells. Moreover, many of them exist as relatively short segments although global conformational effects also exert influence to a certain degree. Thus, it seems rather promising to interpret such signals and predict the *in-vivo* fate of proteins. Here, I describe sequence features of each signal and the result or prospect of theoretical attempts to recognize them. It seems to be convenient to use the terminology recently introduced by Varshavsky[4] because it covers our subject consistently except for the sorting signals in bacterial proteins. Related subjects, such as proposed sequence features recognized by molecular chaperones[5], are not treated here because there is few experimental evidence.

2. SORTING OF BACTERIAL PROTEINS

2. a. General aspects

Although there are usually no organelles in bacterial cells, the basic membrane translocation problem of the cytoplasmic membrane remains. That is, cytoplasmic membrane proteins should be inserted into the membrane and proteins to be excreted, *i.e.,* secreted into the extracellular space, should go through the membrane. In Gram-negative bacteria, since there is an additional outer membrane, the sorting problem becomes more complicated; the outer membrane proteins and the periplasmic proteins, which exist in the space bounded by both the outer and inner (cytoplasmic) membranes, are to be sorted and appropriately localized (Figure 1). There are also some other minor localization sites such as pili and flagella.

From the viewpoint of interpreting unknown amino acid sequence information,

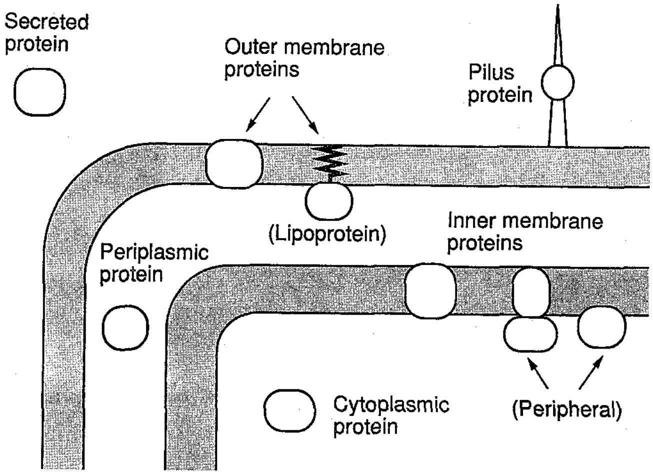

Fig. 1. Various protein localization sites in Gram-negative bacteria (adopted from 11). Basically, proteins localize at either the cytoplasm, the cytoplasmic (inner) membrane, the periplasmic space, or the outer membrane.

it is important to know how general the protein sorting mechanisms are. If they were specific for each protein to be sorted, the knowledge of their mechanism could not be used for prediction. Of course, there is the possibility that some proteins specifically sorted for each share a certain feature unrelated to the sorting signals. However, finding such common features from limited data is dangerous because there is no effective way to test their generality. In view of the known sorting mechanisms of bacterial proteins, many excreted proteins in Gram-negative bacteria[6], proteins at *pili*[7] and *flagella*[8] are sorted through somewhat specific mechanisms. On the other hand, there are also common mechanisms which serve for the sorting of most proteins. It has been postulated that the translocation machinery of the cytoplasmic membrane of Gram-positive bacteria is essentially the same as that of Gram-negative bacteria because their signals can be recognized mutually[9]. Thus, the excreted proteins in Gram-positive bacteria are likely to be regarded as periplasmic and outer membrane proteins in Gram-negative bacteria. We focus our attention on the basic sorting mechanisms in the latter bacteria.

2. b. Analyses of signal sequences

In Gram-negative bacteria, the major four sorting sites are: the cytoplasm, the cytoplasmic (inner) membrane, the periplasmic space, and the outer membrane[10]. Nakai and Kanehisa systematically examined the sequence features of these four types of proteins in their database[11]. In general, the 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 cleavable N-terminal signal sequences but some N-terminal signal sequences in the cytoplasmic membrane proteins are not cleaved off, remaining as transmembrane segments. The other cytoplasmic membrane proteins do not have N-terminal signal sequences except lipoproteins (see below). However, most of them seem to have internal signal sequences, instead. The remainders are thought to be peripheral membrane proteins, *i.e.*, proteins which loosely associate with the membrane and are not integrated in it. Many of them are thought to be anchored by specific integral membrane proteins as components of a membrane protein complex. Therefore, it is presently impossible to discriminate them from cytoplasmic proteins. It may be more appropriate to classify these peripheral membrane proteins as a distinct class or to merge them into the class of cytoplasmic proteins.

Theoretical works have been done to analyze the sequence features of N-terminal signal sequences. In his series of works,[12~16] von Heijne clarified their typical features: First, they consist of three domains (a basic N-terminal region, a central hydrophobic region, and a more polar C-terminal region). Second, a net charge imbalance between the N-terminal and the C-terminal regions is usually observed in bacterial sequences. Third, the amino acids with positions -3 and -1 relative to the cleavage site are loosely conserved ("(-3, -1)-rule"). And fourth, the signal sequences of lipoproteins are essentially the same as those of usual proteins except for the region around their cleavage sites. These features are also observed in the signal sequences (ER-transferons; see below) of eukaryotic proteins to some degree.

However, there is still a controversy over whether signal recognition particles have a primary importance in the protein translocation through the cytoplasmic membrane as in eukaryotic cells.[17] Based on these results, von Heijne proposed a simple method to recognize signal sequence cleavage sites.[18] It is a weight-matrix method and is largely dependent on the information of the (-3, -1)-rule. On the other hand, McGeoch published another method to recognize signal sequences, which only considers the N-terminal and central regions.[19] According to our recent analysis,[11] both methods could be effectively used for the recognition of signal sequences. The fact that no prokaryotic sequences were used for the derivation of McGeoch's method implies that they are functionally common. Interesingly, by exploiting the different nature of the two methods, we could predict where some uncleavable signal sequences appeared in cytoplasmic membrane proteins. In addition, the result of the discrimination of lipoproteins, proteins with a covalently attached lipid molecule in the mature N-terminus, was rather satisfactory. They are recognized by the combination of McGeoch's method and the consensus motif around the cleavage site formulated by von Heijne.[16]

Another interesting analysis was made by Sjĕström *et al.*[20]. Using a multivariate method they have developed themselves, they claimed that there are significant differences in the signal sequences with different localization sites. Although it may be possible that there are distinct sequence features reflecting the difference of evolutionary origin, they are all likely to be functionally equivalent because they can be swapped experimentally (reviewed in 9).

2. c. Other signals

The N-terminal lipid moieties of lipoproteins are thought to be integrated into membranes. Thus, they are membrane-associated proteins. Furthermore, they are specifically sorted into either the cytoplasmic membrane or the outer membrane. As for its sorting signal, the importance of an N-terminal segment, essencially the second residue from the N-terminus, was demonstrated[21], that is, if a lipoprotein has a negatively charged residue at the second or third position of the mature part, it is sorted to the inner membrane; otherwise, it is sorted to the outer membrane. We could use this rule effectively for the further discrimination of lipoproteins although there were very few examples.

According to our observation,[11] hydrophobic transmembrane segments exist in cytoplasmic membrane proteins only. Thus, these segments can be regarded as the sorting signal into the cytoplasmic membrane. On the other hand, outer membrane proteins do not have any hydrophobic segments which characterize usual integral membrane proteins. This phenomenon is usually interpreted by the hypothesis that the membrane-spanning part of the outer membrane proteins consist of β strands[22]. In fact, the 16-strand β-barrel structure of porin was determined by X-ray crystallography.[23] The sorting signal which discriminates outer membrane proteins from periplasmic proteins is not well characterized. Recently, Struyvé *et al.* reported that many outer membrane proteins have a phenylalanine residue in the C-terminus and showed its importance by site-directed mutagenesis.[24] In the database of our analysis, 10 out of the 22 outer membrane proteins had a C-terminal phenylalanine while none of the 21 periplasmic proteins had one (6 of them were lysine residues). As Stuyvé *et al.* suggested, it seems possible that the outer membrane proteins of other C-terminal residues are located in special regions of the membrane. However, the current rule seems to be too simple to discriminate outer membrane proteins with sufficient predictability. According to our analysis, although the predicted secondary structure contents were not significantly different in the periplasmic and outer membrane proteins, their amino acid composition turned out to be different enough.[11] The discriminant function calculated from the amino acid composition was powerful enough to discriminate all proteins but one exception, OmpA. It is likely that the difference of amino acid composition reflects the difference of environments around the proteins because we could roughly guess the outer membrane-integrated region of OmpA protein when we applied the discriminant function to the local segment of that protein (Nakai and Kanehisa, unpublished result). As shown in Figure 2, the boundary of the discriminant score roughly corresponds to the structural boundary of Vogel and Jähnig's model.[25]

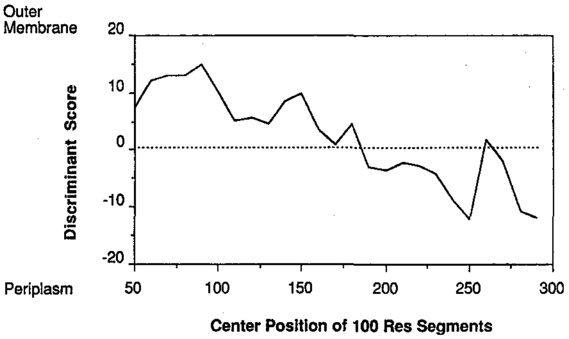

Fig. 2. Profile of the discriminant score along the sequence of OmpA precursor (Nakai and Kanehisa, unpublished result). Positive scores predict that the regions are integrated into the outer membrane, while negative scores mean that the regions are exposed to the periplasm. The horizontal axis represents the center position of segment of length 100. According to a current model (25), positions 22 to 198 are integrated into the membrane (1-21 is the signal sequence) in good agreement with our prediction.

Representing the above-mentioned results as 'if-then'-type rules, Nakai and Kanehisa constructed an expert system which can predict the protein localization sites of Gram-negative bacteria from their amino acid sequences.[11] An expert system is an artificial intelligence technique in which computers are equipped with domain specific knowledge. With this method, 83% of the 106 proteins were correctly classified into one of the four localization sites. Although the prediction accuracy when applied to unknown sequences has not been estimated, this system seems to be useful in characterizing ORFs found in the *E. coli* genome.

3. MEMBRANE TOPOLOGY OF PROTEINS

3. a. Recognition of transmembrane segments

In the original Varshavsky's definition, the topogenic signals of membrane proteins, *i.e.*, the signals which determine the membrane topology of proteins, are not included in targeting signals[4]. However, as described later, they are also a kind of sequence information which strongly affects the *in-vivo* fate of proteins especially in eukaryotic cells. In this section, the theoretical aspect of predicting protein membrane topology is reviewed.

In general, the recognition of probable transmembrane segments which are thought to be α-helices in membranes is not regarded as a difficult problem: in fact, as a theoretical method, it is an exceptionally widely used by experimental researchers. When one draws a plot of hydropathy using, say, the parameters of Kyte and Doolittle's,[26] transmembrane segments are usually visualized as hydrophobic peaks. Assessments of these predictions have shown positive results[27 28]. However, when one tries to predict the membrane topology of multiple membrane-spanning proteins, one must predict the exact number of these segments. For, if one misses a single transmembrane segment, the topology of the C-terminal region from that segment might be predicted totally in reverse. In this respect, in our recent analysis even the method of Klein *et al.*'s,[29] which has been ranked as one of the best methods for evaluation,[27] was totally insufficient.[30] One way for improvement is to adjust the cut-off value more precisely. For example, in our analysis, the effectiveness of using two cut-off values was implicated: when predicted to be a polytopic, *i.e.,* multiple membrane-spanning, protein, a less stringent value was employed for the prediction of more realistic number of transmembrane segments. It seems probable that once integrated into the membrane, less hydrophobic segments are also integrated into it. Another way may be to improve the parameter of hydrophobicity. Nakai *et al.* performed a cluster analysis of 222 parameters which represent various properties of amino acids.[31] Accordingly, hydrophobic values are classified into one of four major groups and various subclasses of hydrophobicity exist. Combinational use of these parameters may yield a more accurate prediction. In fact, it seems inappropriate to apply the present method to the prediction of membrane protein in some specific organelles.[30]

As described, current methods appear to be insufficient to predict transmembrane segments of polytopic proteins confidently. Nevertheless, quite a few membrane proteins are bitopic, *i.e.,* having a single transmembrane segment. In this case, rather accurate prediction was possible.

3. b. Prediction of membrane topology

The topology of membrane proteins has been classified by several authors.[32~34] According to Singer's latest classification (34; see Figure 3), bitopic proteins are classified into type I whose N-terminal part is exposed in the extracytoplasmic space (NexoCcyt configuration) and type II with an NcytCexo configuration. The former is further divided into type Ia and type Ib corresponding to the existence or lack of a cleavable signal sequence. Of the polytopic proteins, channel proteins are

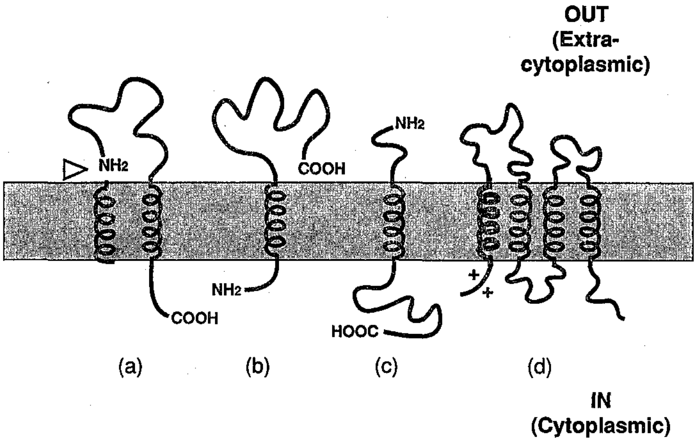

Fig. 3. Topology of membrane proteins (reproduced from 30). The classification is based on the definition by Singer (34). (a) Type Ia proteins. The N-terminal ER-transferon is cleaved off at the ER. (b) Type II proteins. (c) Type Ib proteins. (d) Type III proteins. The most N-terminal helix is shown in bold.

defined to type IV while others are to type III. In the database of known localization sites collected by us[30], most of the bitopic proteins were type Ia. This type of topology could be predicted by the recognition of N-terminal ER-transferon and an additional transmembrane segment (relatively) easily. One interesting finding in our analysis was that there seems to be a preference of membrane topology in each localization site[30]. For example, type Ib proteins are favored at the ER (endoplasmic reticulum) while type II tend towards the Golgi complex and the plasma membrane. However, more data must be analyzed to say anything definite.

While the existence of an ER-transferon seems to determine the membrane topology, the distinction between type Ib and type II should be made by some other signals. Both experimental and theoretical works have pointed out the importance of charged residues, especially positively charged ones, flanking the transmembrane segments, although the detailed consensus has not been made(33, 35, 36 and reviewed in 37). Of these, Hartmann *et al.*'s work seems most useful for prediction[35]. They claimed that the overall topology is determined from the net charge difference of both sides of 15 residues flanking the most N-terminal transmembrane segment and they showed that in their database including both bitopic and polytopic proteins, the NcytCexo and NexoCcyt configurations can be clearly distinguished by their method. One interesting point of their work is that this hypothesis gives us a simple and unified view of membrane protein topogenesis including the usual ER-transferons integrated in a type I fashion: as already mentioned, many ER-transferons have a few positively charged residues in the N-terminal region and the importance of the charge difference between N- and C- terminal regions has been suggested in bacterial sequences.[15]

Nakai and Kanehisa incorporated Hartmann *et al.*'s method in their scheme of predicting eukaryotic protein localization sites.[30] One problem was that the N-terminal transmembrane segments of type Ib proteins were often wrongly predicted to be cleaved off by von Heijne's method.[18] Therefore, we introduced the hypothesis that if the charge difference of the most N-terminal transmembrane segment is reversed to that of usual ER-transferons, it is not cleaved. By this hypothesis, the localization site of several ER membrane proteins were correctly predicted. However, since some cleavable ER-transferons had a reversed charge difference, we had to change the originally reported threshold value. In addition, it was still difficult to distinguish some type II proteins from type Ia proteins although transmembrane segments of many type II proteins reside apart from the N-terminus to some degree.

Recently, the existence of a protein-conducting channel in the ER was demonstrated with electrophysiologic techniques.[38] One important implication of this finding is that the topogenesis of membrane proteins is regulated by this channel[39]. Thus, the principle of membrane topogenesis must be understood in light of the nature of this protein-conducting channel.

4. SORTING OF EUKARYOTIC PROTEINS

4. a. Transferons in general

The sorting pathways of eukaryotic proteins are roughly divided into two categories: the pathway through the cytoplasm and the one through the ER. In the former, proteins must be translocated across the plane of membrane at least once in order to get to their target compartment.[40] Signals specifying such a fate are called transferons in Varshavsky's proposal[4]. Note that this definition includes the case such as the nuclear targeting in which proteins do not really translocate across the membrane. In the latter pathway, often called the default secretory pathway, proteins are first sorted to the ER by the signal of ER-transferons. The signals controlling the further localization sites are compartons which will be described later.

The ER-transferons are almost the same as bacterial signal sequences from the theoretical point of view. Some complications related to the distinction of internal start-transfer signals have already been mentioned above. Apart from these, they can be recognized by McGeoch's method[19] and von Heijne's method, which uses a different set of parameters specific for eukaryotes[18], with a reasonably high reliability including the cleavage sites.

4. b. Mitochondrial matrix transferons

Mitochondria, present in all eukaryotic cells, and chloroplasts, present in plant cells, are important organelles responsible for the energy conversion. Although both of them have their own genomes, the majority of their component proteins are synthesized in the cytoplasm and are transported into them. Thus, these proteins must have information for sorting. In mitochondria, many proteins are sorted through a 'conservative' pathway while others are sorted through 'nonconservative' pathways.[41,42] The proteins sorted through the former have mitochondrial matrix targeting signals (M-transferons) in their N-terminus. On the contrary, sequence features of protein sorting signals with 'nonconservative' pathways are hardly recognizable probably because of the lack of common sorting mechanisms. The M-transferons are from 10 to 70 residues in length, and rich in basic and hydroxylated residues.[41] In addition, from their comparative analysis, von Heijne *et al.* proposed the presence of a two-domain structure with different amphiphilic properties.[43] Nakai and Kanehisa developed a simple method to recognize M-transferons from this knowledge.[30] The method of discriminant analysis using the values of partial amino acid composition as variables was performed. For example, the arginine content turned out to be effective for prediction in good agreement with current knowledge. However, to our surprise, the values of hydrophobic moment,[44] which indicate the potential amphiphilicity, were not effective in our analysis. It seems that one reason is the lack of information on the cleavage site. Although some consensus sequence patterns are observed around cleavage sites,[45] they were not general enough to raise the prediction accuracy of our method. Nevertheless, most M-transferons were correctly discriminated in our training data.

4. c. Chloroplast stroma transferons

Proteins targeted to chloroplasts also have cleavable signals in the N-terminus (reviewed in 46). They are called S-transferons bacause they specify transport into the stroma. According to von Heijne *et al.*[43], they consist of three distinct regions when aligned on the cleavage sites. As in the case of M-transferons, however, to incorporate such knowledge fully into a prediction method was rather difficult. Again, the suggested consensus motif[47] was not so strongly conserved as to allow a reliable prediction. We performed a stepwise discriminant analysis of partial amino acid compositions (positions 3–10 and 1–30).[30] This time, the value of the hydrophobic moment of potential β-structure was also selected as an effective variable for discrimination. The result shows the abundance of alanine and serine residues in the N-terminal 30 residues in accordance with the previous analysis.[43] In addition, the observation that the second residue is often alanine[43] was also useful in our prediction. As a result, the discrimination of S-transferons, like that of others like M-transferons, turned out to be possible with reasonable accuracy.[30]

4. d. Intra-organelle sorting

Since mitochondria and chloroplasts have internal structures, there must be additional signals for further internal sorting pathways. Proteins targeted to the mitochondrial intermembrane space via the 'conservative' pathway, have an N-terminal signal of bipartite structure: its N-terminal half appears to be essentially an M-transferon and its C-terminal half is the signal for the translocation from the matrix to the intermembrane space (M/IMS-transferon)[41]. Similarly, proteins of chloroplast thylakoid lumen have a bipartite signal in their N-terminus. Its N-terminal half is essentially the same as an S-transferon and the C-terminal half is used for the translocation from the stroma to the thylakoid lumen (S/T-transferon).[46, 48] The equivalence of these N-terminal halves with usual M-, S-transferons was also suggested in our analysis.[30] That is, most of them were correctly discriminated as M-, S-transferons although they were not used for the derivation of discriminant functions. As for the C-terminal halves, they are rich in apolar residues and the resemblance with bacterial signal sequences has been pointed out in the context of their evolutionary origin[41,43,48]. In our analysis[30], since their hydrophobicity was sometimes weak, we had to develop a new method for the detection of apolar segments. Generally, it worked well for detecting the proposed apolar segments. For the detection of S/T-transferons, another clue, the weight matrix score around the cleavage sites[49], was also effectively used.

Other internal localization sites involve some kinds of membranes. For the mitochondrial outer membrane and the chloroplastic envelope (outer and inner membranes), only a few resident proteins have been sequenced and it seems too early to discuss their sorting mechanisms. Amino acid sequences of proteins localized at the mitochondrial inner membrane or the chloroplast thylakoid membrane are known in large quantity. However, many of them are likely to be peripheral membrane proteins which exist as members of large membrane complexes. In addition, their degree of hydrophobicity is relatively low compared with other membrane proteins, possibly reflecting the different nature of the membrane or the transportation mechanisms. Although we supposed that these membrane proteins are integrated into the membrane with the usual 'stop-transfer' mechanism, there is little evidence. Only recently, Gavel *et al.* claimed that their 'positive-inside rule' can also be applied to thylakoid membrane proteins based on currently available data.[50] Combined with the fact that there are many proteins sorted through 'nonconservative' pathways, the prediction result of these internal localization sites was not satisfactory at the present stage.[30]

4. e. Nuclear transferons

The most notable feature of nuclear targeting is that it does not accompany the real translocation process through the membrane: proteins are transported into the nucleus through nuclear pores (reviewed in 51). In other words, the internal space of the nucleus is topologically equivalent with the cytoplasm. It may even be possible that smaller proteins reach the nucleus through a simple diffusion only. However, no doubt there is specific machinery for nuclear targeting and some kinds of nuclear targeting signals (Nu-transferons) have been discovered.

Although it seems possible that a protein without its own Nu-transferon enters the nucleus via cotransport with a protein that has one[52], many nuclear proteins have their own Nu-transferons. Since there is little sequence homology between these signals, strict classification of them is impossible. Nevertheless, many of them identified so far share common features with the one first identified in SV40 large T antigen.[53] That is, they have one or more stretches of amino acid sequence rich in basic residues and often including prolines. Moreover, they are not cleaved off after localization and their positions in the entire sequence do not appear essential, which might be an universal feature of Nu-transferons.

Recently, another motif of Nu-transferon was discovered in *Xenopus* nucleoplasmin by Robbins *et al.*[54]. According to their report, this type of signal consists of two interdependent basic domains separated by about 10 'spacer' amino acids. The simple criterion which they proposed for the detection of these signals was quite effective in our prediction scheme.[30]

Some proteins exert their function as a form of ribonucleoprotein complex, *i.e.*, they bind with RNA molecules specifically. It has been shown that these RNA molecules can have the information for sorting to or from the nucleus and that some protein components of ribonucleoproteins alone cannot enter the nucleus[55 56]. The methylation state of the cap structure of RNAs turned out to be important. From the theoretical point of recognizing sorting signals in proteins, these kinds of signals are hard to deal with. However, the existence of a consensus motif characterizing the RNA binding sites[57] can be a clue for prediction.[30] However, it is apparent that this information alone is insufficient because knowledge of the nature of the bound RNAs is essential.

The regulation of gene expression is of primary importance for life systems. Therefore, there are many complicated but elegant regulation mechanisms. One such example exploits the possibility of utilizing protein sorting: the localization sites of some regulatory proteins are regulated in cells (reviewed in 58). Although the precise mechanism has not been clarified, the general scheme may be as follows. All of these regulatory proteins having Nu-transferons are usually anchored by specific cytoplasmic proteins. When the regulatory proteins are to be transported to the nucleus, either they or their anchoring proteins are modified, say, by phosphorylation resulting in their dissociation.

Combining the above-mentioned knowledge, Nakai and Kanehisa examined the feasibility of discriminating nuclear proteins.[30] In this analysis, ribosomal proteins were classified as nuclear proteins because they have Nu-transferons and are once transported into the nucleus.[59] The heuristic that a highly basic protein is likely to be a nuclear protein was also used. 62% of nuclear proteins in the testing data could be correctly predicted.

4. f. Peroxisomal transferons

Peroxisomes are organelles surrounded by a single membrane and are found in almost every eukaryotic cell. Functionally, they usually contain one or more oxidase and catalases that take part in various oxidative reactions. In some organisms, probable functional equivalents are called glyoxisomes or glycosomes, all of which are sometimes referred to as microbodies[60]. As one sorting signal (we call it P-transferon), the importance of the C-terminal three residues has been indicated: proteins are sorted to peroxisomes *in vitro* and *in vivo* if they have the sequence (S/A(/C))(K/R/H)L at the C-terminus[61 62]. This pathway seems to be conserved throughout eukaryotic cells.[63] The recognition of this P-transferon is thus easy and its presence at the C-terminus strongly suggests the protein localization at peroxisomes. However, many peroxisomal proteins do not have that motif at the appropriate position. Although there is no experimental evidence, the possibility that the motif at other sequence positions also works for the signal has been suggested.[61] In fact, all but one peroxisomal proteins by contrast with 45% of the cytoplasmic and nuclear proteins in our training data had at least one P-transferon motif.[30]

Some peroxisomal proteins have N-terminal presequences which are cleaved off after translocation. However, that does not necessarily mean that they are sorting signals because the cleavage process does not appear to be coupled with translocation[62]. According to our preliminary analysis, the amino acid composition of the N-terminal 20 residues were not very effective as variables of discriminant analysis.[30] Rather, the amino acid composition of the entire sequence seemed more useful for supplemental information for prediction. However, the testing data of peroxisomal proteins was poorly predicted although the data size was small. Clearly, more knowledge should be incorporated.

The sorting signal of peroxisomal membrane proteins is not known. Only a few membrane proteins have been sequenced. Recently, a membrane protein that restores the biogenesis of peroxisomes was cloned.[64] Thus, it is expected that subsequent studies on the membrane components of translocation machinery will increase our understanding of the sorting processes of both soluble and membrane proteins.

4. g. Compartons in general

Proteins with N-terminal ER-transferons are synthesized at the surface of rough ER. These ER-transferons also take part in the process of protein translocation through the ER membrane. If these signals are not cleaved off or the proteins have other stop-transfer signals, these proteins will be integrated into the membrane; otherwise, they pass through the membrane into the lumen of ER. In the ER, components of protein complexes are assembled[65]. In many proteins, however, the ER is not their final localization site. If so, they are further transported in a vesicle-mediated manner. One important finding is that there is a nonselective sorting pathway from the rough ER to the plasma membrane or the extracellular space, called the bulk flow (reviewed in 66). Thus, proteins without further sorting signals are transported to the cell surface or secreted by default. The signals controlling other fates are called compartons in Varshavsky's terminology[4]. Compartons specify a protein either for retention in a certain compartment within the bulk flow pathway or to leave from the pathway without crossing the plane of a membrane. The rest of this section is a review of recent findings on these signals and our attempt to use them for prediction though only a few compartons have been analyzed so far.

4. h. ER-compartons

The retention signal of ER luminal proteins is undoubtedly the most well known comparton. It is rather simple; the existence of the sequence motif KDEL in the C-terminus is usually essential[67]. In yeast and some plants, the consensus motif is HDEL. Several variant types of the motif are also tolerated as a signal possibly depending on the types of cells or organisms. Although the precise mechanism of retention has not been clarified, the signal is thought to be recognized by receptor(s) that continually 'salvages' ER-resident proteins which have 'flown out' from the ER[67,68]. Due to the simplicity of the signal, its recognition was very easy[30]: existence of the KDEL motif was necessary and sufficient for the discrimination in our data.

Compared with the KDEL motif, the ER-comparton identified in some membrane proteins seems less evident as a sequence motif: in one analysis using mutagenesis, two lysines positioned three and four or five residues from the C-terminus turned out to be important in some type Ia proteins.[69] This is one example of various comparton signals existing in cytoplasmic tails, which will be described later. Another group reported that there is a similarity between the *retention signal of a membrane protein and the microtubule binding sequence* identified so far.[70] Thus, the interaction with microtubules may be important for the retention of ER membrane proteins. Some type III proteins may also have this kind of signal. However, many ER membrane proteins do not have this kind of sequence motif. It is possible that some of them are associated with large membrane protein complexes that cannot enter transport vesicles.[71] Because of the lack of generality, the knowledge of the ER-comparton for membrane proteins was not so powerful in our analysis.[30] The preference of membrane topology was a rather useful clue.

4. i. Other compartons in cytoplasmic tails

As already exemplified above, many comparton signals in the membrane proteins have been found in cytoplasmic tails which are short terminal segments exposed to the cytoplasm in type Ia, Ib, and II proteins.

In relation to the default pathway of secretion, there is a pathway for protein internalization through coated pit-mediated endocytosis.[72] In this process, some ligand-bound receptors in the plasma membrane are internalized into the endosomes. After dissociating ligands in an acidic environment, these receptors are recycled to the plasma membrane. Two sequence motifs, NPXY and YXRF, have been identified as comparton signals for this rapid internalization process[73,74]. They exist in the cytoplasmic tails and their exact position in a sequence seems unimportant, provided that there is a spacer from the transmembrane region. A comprehensive computer pattern-matching search against a structural database was performed.[74] Many tetrapeptide analogs of YXRF favored tight-turn conformations. Interestingly, most NPXY related sequences were also found in similar structures. Thus, an exposed tight turn may be recognized as the structural motif *for high efficiency endocytosis. In our analysis, these sequence motifs were effectively* used as clues identifying plasma membrane proteins[30].

As described later, the sorting mechanism of lysosomal membrane proteins seems different from that of lysosomal luminal proteins. As an example of the former process, the importance of a tyrosine residue at a particular position in the cytoplasmic tail has been indicated[75]. In addition, the possibility that this signal may be recognized both in the *trans*-Golgi network and at the cell surface was suggested. Thus, there may be a close relation between this signal and internalization signals. For our predictive analysis[30], the existence of a GY motif within 17 residues from the membrane boundary in the cytoplasmic tails of type Ia proteins was used as a rule for discrimination. The three examples were successfully discriminated with the aid of discriminant score.

Other compartons are not well understood at present. For example, the existence of a general retention signal at the Golgi apparatus has not been experimentally proven. It was proposed that a consensus motif, (S/T)X(E/Q)(R/K), existed near the probable transmembrane domain of all Golgi-localized glycosyltransferases may be a comparton signal.[76] Although this motif seems unrelated to the active site, it is not clear whether the same motif also exists in other proteins with different functions.

In addition, a comparton signal which sorts out the proteins to secretory vesicles must exist: some proteins are secreted in a regulated manner in contrast to the constitutive secretion by default.[77] A motif consisting of leucines and a preceding serine, which are distributed in an amphipathic helix, was proposed as a comparton for propeptides and prohormones[78]. Although its generality should be tested experimentally, the knowledge may be useful for future improvement of our knowledge base.

Lastly, there exists a sorting signal that takes part in the formation of the polarized plasma membrane found in epithelial cells, for example[79]. Although the nature of the signal has not been clarified, it seems to be a comparton: at least, some membrane proteins are transported by the mechanism known as transcytosis.

4. j. Targeting to lysosomes and vacuoles

Lysosomes are acidic organelles that contain numerous hydrolytic enzymes capable of degrading most biological macromolecules. In yeast and plant cells, similar functions are recognized in vacuoles, which have diverse functions. It is not known whether there is a common machinery for protein sorting although it has been observed that a sorting signal of an yeast vacuolar protein is also functional in an animal cell.[80]

As already mentioned, there are at least two sorting pathways for lysosomal proteins.[81] For soluble proteins, the pathway which utilizes the post-translational modification of mannose 6-phosphate has been clarified. That is, two kinds of receptors which recognize mannose 6-phosphate modification sites and transport the proteins to lysosomes have been identified. From the theoretical point of view, sequence determinants that lead to this specific modification should be examined. Since no clear consensus patterns have been found except for the NX(S/T) pattern necessary for *N*-glycosylation, it has been suggested that the modification might be conformation-dependent. A recent experiment using chimeric enzymes supports this hypothesis: in cathepsin D, amino acids from different regions come together in three-dimensional space to form the recognition domain.[82] Since the prediction of protein conformation is very difficult, we used the discriminant score based on amino acid composition.[30] However, its prediction accuracy for unknown sequences was not tested because of the small data size.

The sorting mechanisms of yeast vacuolar proteins have been studied as a model system for understanding lysosomal or other vacuolar sorting mechanisms[83]. Indeed, it is likely that yeast and most plant cells share their sorting mechanism although there seem to exist divergent pathways also. Although some vacuolar proteins are sorted by mechanisms which do not follow the default secretory pathway[84 85], most of them have ER-transferons in their N-terminus. In addition, quite a few vacuolar proteins have pro regions that are cleaved off after translocation, which are thought to be comparton signals recognized at the Golgi complex.

Nevertheless, no common sequence features have been observed. We used the information of amino acid composition again for the discrimination[30]. The amino acid composition of lysosomal and vacuolar soluble proteins turned out to be totally defferent.

4. k. prediction of localization sites

Nakai and Kanehisa perfomed the prediction of protein localization sites in eukaryotic cells, making full use of the above-mentioned knowledge[30]. The knowledge on lipid anchors described later was also included. The number of localization sites was 14 and 17 in animal and plant cells respectively. The expert system previously constructed for the prediction of localization sites in Gram-negative bacteria[11] was expanded for dealing with eukaryotic proteins. Thus, 80 core rules were added to the knowledge base. The simplified reasoning tree is shown in Figure 4. Of the 295 proteins used for the tuning of our system, 66% were correctly discriminated. Moreover, of the 106 proteins selected randomly from the localization sites including more than 10 members for testing, 59% were correctly predicted. Many falsely predicted proteins seemed to be transported by specific pathways. However, the prediction accuracy will be certainly improved by incorporating the future accumulation of our knowledge. And the flexible nature of the knowledge base for modifying its contents will be undoubtedly useful for future improvement.

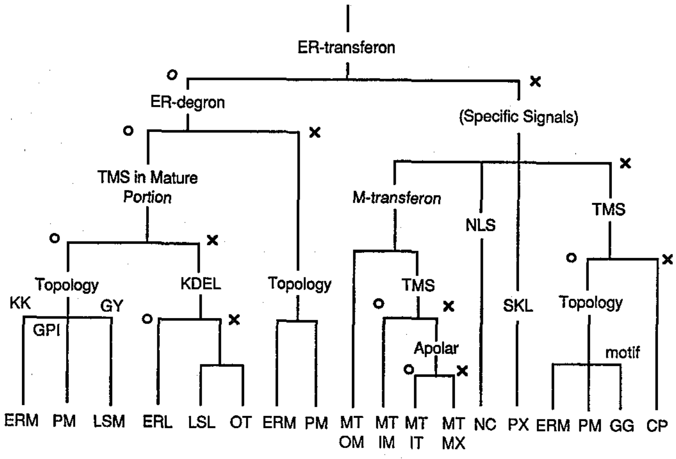

Fig. 4. Simplified reasoning tree for the prediction of localization sites of eukaryotic proteins (reproduced from 30). Abbreviations: TMS, transmembrane segment. KK, GY, KDEL, and SKL are amino acid sequence motifs. NLS, nuclear localization signal. CP, cytoplasm. ERL and ERM, lumen and membrane of ER. GG, Golgi apparatus. LSL and LSM, lumen and membrane of lysosome. MTIM, MTIT, MTMX and MTOM are inner membrane, intermembrane space, matrix and outer membrane of mitochondria. NC, nucleus. OT, outside space. PM, plasma membrane. PX, peroxisome.

5. DEGRONS

5. a. General aspects

In cells, most proteins undergo continuous degradation and synthesis. Turnover rates of proteins are not uniform but are regulated according to their physiological requirements. For example, some transcription factors are known to be degraded in extremely short half-lives. Thus the information of protein half-lives must be also encoded in their sequences. However, these molecular determinants, degrons, have not been so fully understood that we can predict the lifetime of an unknown protein, because there are so many distinct degradation pathways.[86] Another difficulty is that the half-lives of proteins are not the same in different cell species or even in the different states of cells. Several types of covalent modification reactions, such as oxidations, also affect the rates of protein turnover.[87] Nevertheless, the knowledge of a few well-studied degrons seems to enable us to predict the metabolic stability of proteins to some extent. Here, I will briefly review three of these examples. Other processing signals which are also important in understanding protein maturation are not treated because they are often specific (but see 88).

5. b. The N-end rule

In the cytoplasm, misfolded or abnormal proteins are rapidly degraded by the ubiquitin-mediated proteolytic pathway. Some selective normal proteins are also targeted to this pathway. There, ubiquitins, strongly conserved small proteins, are covalently linked to the target proteins forming a branched multiubiquitin chain. Afterwards, they are degraded by a large ATP-dependent protease. Studies of its molecular determinants have revealed that the type of N-terminal residue (N-end rule) and a specific internal lysine residue comprise the degron signal[89]. Although this N-end rule differs in some organisms, a protein which has a destabilizing residue in its N-terminus is degraded rapidly. From the theoretical point of view, the N-terminal residue in the cytoplasm must be predicted beforehand. However, it is determined not only from the trimming of initial methionine and subsequent residues but also from the end-addition of reactions mediated by aminoacyl-tRNA-protein transferases. Another complication factor exists in the case of multisubunit proteins. That is, a destabilizing N-terminal residue can be located on the different subunit than the one at which a specific lysine is located[90]. Therefore, the recognition of these N/aa-modons which form N-degrons seems difficult at the present stage.[91]

5. c. The PEST hypothesis

From the computer-based survey of amino acid sequences of 12 short-lived proteins and 35 long-lived proteins, Rogers *et al.* discovered that the most striking property that characterizes these short-lived proteins is the presence of region(s) rich in proline, glutamic acid, serine, and threonine.[92,93] These regions are called PEST regions from the one letter abbreviation of these amino acids. The importance of PEST regions has been confirmed experimentally. An algorithm to search for PEST regions was also proposed. In principle, it searches for a region of 10 or more residues long, which is enriched for P, E, S, and T and is flanked by basic residues. Then the score for the PEST propensity is calculated with the result of hydrophilicity analysis. Although its general predictability has not been ascertained, one prediction that caseins are rapidly degraded in eukaryotic cells turned out to be true. The intracellular pathway of degrading PEST proteins is not known. It is possible that it utilizes the ubiquitine-dependent pathway. Another possible explanation is that PEST regions or phosphorylated PEST regions may bind to calcium ion and stimulate calcium-dependent proteases (calpains).

5. d. The KFERQ motif

The third example is related to the lysosomal proteolysis. Among several pathways in which proteins are internalized by lysosomes, there is a pathway which responds to starvation and which takes up and degrades cytoplasmic proteins in a highly selective manner (reviewed in 94). Probing various microinjected fragments of ribonuclease A, the KFERQ peptide was indicated to be the molecular determinant of the enhanced degradation in response to serum withdrawal. Because such a motif is only found within the pancreatic ribonuclease A family, a more loose representation which can distinguish long-lived proteins that are degraded more rapidly during serum withdrawal was searched for extensively. It was (+, □, −, +/□) Q or Q (+, □, −, +/□) where + and − denote basic and acidic residues respectively and □ denotes F, I, L, and V. The location of this motif in the sequences did not appear to be crucial for its function. As an uptake mechanism for lysosomes, the role of a heat shock protein, prp 73, has been suggested.

6. MODONS

6. a. General aspects

There are a great many variations of protein modification reactions *in vivo*[95]. At present, many of their functions are unknown and even some of them may not have functional importance. However, some phosphorylation reactions, for example, play central roles in intracellular signal transductions. Furthermore, modification reactions are closely related to the intracellular sorting and degradation pathways. Some modification reactions specify the sorting site of proteins and, conversely, modification reactions are essentially dependent on the sorting pathway because of the localization of enzymes that catalyze these reactions. Therefore, future prediction systems must deal with both the localization and modification of proteins. So far, however, only comparative analyses of the sequences around the same kind of modification sites have been done. In the theoretical study of protein modification, the recognition of modon signals can be considered as the understanding of the sequence specificity of modification enzymes. Again, three typical examples are shown here.

6. b. Glycosylations

Glycosylation is one of the most abundant modifications in proteins. The linked carbohydrate groups not only confer important physical properties like conformational stability to proteins but they are also important in various biological recognition processes[96]. They are roughly divided into two categories:

N-glycosylation, in which carbohydrate groups are attached to asparagine residues co-translationally, and *O*-glycosylation, in which modifications occur mainly in serine and threonine residues post-translationally.

For the sequence determinants of *N*-glycosylation, the existence of a consensus sequence, NX(S/T), has well been established[97]. However, not a few positions satisfying this condition are not *N*-glycosylated. In order to better understand the determinants of this difference, Nakai and Kanehisa performed a further prediction of *N*-glycosylation sites from possible sites where the NX(S/T) requirement is already satisfied[98]. The method was essentially an elaboration of consensus sequence pattern-matching based on stepwise discriminant analysis. From our previous study of 222 amino acid properties[31], six representative parameters were chosen. The occurring residues near a potential modification site were represented by them. In order to introduce longer range effects, the length of the protein and the relative position of the site in the sequence were also used as potential variables. The prediction accuracy evaluated by a cross validation procedure was 60% for the further discrimination of potential *N*-glycosylation sites. From the inspection of selected variables, it was found, for example, that *N*-glycosylation sites appear more frequently near N-terminal regions than C-terminal regions. A similar observation was also reported in a more recent statistical analysis[99].

Several types of *O*-glycosylation are known to exist. One of the recent discoveries on *O*-glycosylation was that some cytoplasmic and nuclear proteins are extensively *O*-glycosylated[100]. In contrast to *N*-glycosylations, there are no clear sequence requirements. According to a recent compilation of 'mucin-type' *O*-glycosylation sites, some weak features are observed. However, they could not be used for the prediction of isolated *O*-glycosylation sites[101].

6. c. Phosphorylations

The importance of protein phosphorylation cannot be overestimated. In one review, it was suggested that the number of protein kinases, *i.e.*, enzymes mediating the phosphorylation reactions, may be as many as a thousand[102]. Thus, although the most promising approach for predicting phosphorylation sites is to collect knowledge of the sequence specificities of these kinases[103], that knowledge is insufficient for the prediction of unknown sequences for the time being. Since there is significant sequence homology between distinct protein kinase domains[104], there also may be some similarities in their sequence specificities. In this respect, Nakai and Kanehisa derived discriminant functions for the prediction of Ser/Thr kinases, *i.e.*, kinases that phosphorylate serine or threonine residues[98]. The method was the same as the one for the prediction of *N*-glycosylation sites. In this case, the prerequisite was the presence of a serine or threonine residue at the position to be predicted. With a cross-validation estimation, the prediction accuracy was about 80%. Its relatively high reliability seems to reflect the general preference that there are usually some basic residues scattered around the Ser/Thr phosphorylation sites. One obvious exception was the recognition sites of casein kinases, in which acidic residues frequently appear. However, the discrimination of three groups separating the casein kinase sites was not so effective in raising the predictability.

That was probably because that the presence of charged residues was sufficient for discriminating phosphorylation sites.

6. d. Lipid anchors

Amongst the forest of protein modifications, the reactions which bind lipid molecules to proteins are interesting because with these reactions the fate of modified proteins may be drastically changed. That is, a linked lipid moiety is sometimes thought to be integrated into various membranes and to anchor the bound protein. So far, several types of modifications have been clarified.[105] Typical examples are shown in Figure 5.

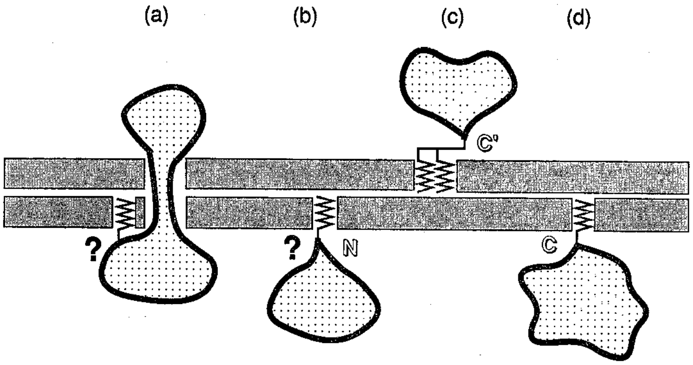

Fig. 5. Hypothetical forms of various membrane anchors (reproduced from 30). The space under the membrane depicts the cytoplasm. (a) Palmitoylation. (b) N-myristoylation. (c) Glycosyl phosphatidylinositol (GPI) anchor. (d) Isoprenylation.

Some kinds of fatty acids are found to be directly bound to proteins (acylations). In eukaryotic cells, palmitic acid and myristic acid are typical ones[105]. Palmitoylations occur at serine, threonine, and cysteine residues. Many of the modified proteins are membrane proteins which have hydrophobic transmembrane segments. The reactions occur post-translationally. In general, there are no apparent consensus sequences around the modification sites, although the cysteine residues located at the cytoplasmic side near the transmembrane region appear to be modified more often.[106] It is not known whether these palmitic acid moieties are really inserted into membranes. Myristoylations occur co-translationally often at the N-terminal glysine residue of proteins.[107] The yeast enzyme which mediates this reaction was purified and its substrate specificity has been extensively studied. The derived consensus sequence in the N-terminal 9 residues could be used for prediction. Formerly, N-myristoylated proteins were thought to be anchored to the membrane via their lipid moieties. However, recent studies suggest that many of them may not take part in the direct anchoring.[108] For example, the interaction of an N-myristoylated *src* protein with the plasma membrane is likely to be mediated by a specific receptor molecule.[109]

In contrast, all proteins linked to the glycosyl-phosphatidylinositol (GPI) molecules are thought to be anchored at the extracellular surface of the plasma membrane[110,111]. Thus, the prediction of this modification was very useful for the prediction of the localization site (plasma membrane) of the modified protein[30]. The signal and the mechanism of GPI-modification are not fully understood. However, all of the precursors seem to be type Ia membrane proteins and the cytoplasmic tails are thought to be cleaved off after the linkage with GPI. Although there are no strongly conserved sequence motifs near the cleavage sites, some preference of amino acids seems to exist. A more prominent feature is that the length of the predicted cytoplasmic tails is very short, if present at all. Although the existence of short tails has been shown to be insufficient for the reaction experimentally, the prediction could be performed rather accurately. Indeed, 75% of the data were correctly predicted and all failures were caused by the false prediction of membrane topology.[30]

Lastly, there is a lipid modification known as isoprenylation or farnesylation[112]. This modification requires a CaaX motif in the C-terminus, where 'a' denotes an aliphatic amino acid. Isoprenylated proteins have been found in the plasma membrane and the nuclear envelope. Thus, it is evident that the modification itself does not determine the localization site completely. As an example of the formar, it was shown that a polybasic domain or palmitoylation is required in addition to the motif to localize an oncogene product, *ras,* to the plasma membrane.[113] As an example of the latter, both the Nu-transferon and the CaaX motif of lamin A were required for its nuclear envelope assembly.[114]

7. CONCLUDING REMARKS

The main purpose of this review was to show how there are various types of targeting signals and how they are interconnected with each other. For example, there are signals like ER- or M-transferons that cannot be represented as a simple consensus sequence, while there are sequence motifs like KDEL that are used for the specific targeting of proteins. Some of these motifs are influenced by the membrane topology which is formed by hydrophobic segments and charged residues nearby. The sorting signal of lysosomal proteins are even more complicated. For soluble proteins, the modification of mannose 6-phosphate is important for the targeting and the determinant of that modification is thought to be conformation-dependent. Protein modification reactions occur only when proteins are sorted to the appropriate compartment. The examples seem to be unlimited. Therefore, various signals reviewed in this article must be understood in a total fashion. In other words, future prediction systems must give us the overall perspective of the *in-vivo* fate of proteins. Furthermore, in those systems, signals of various types should be treated in various ways; a simple method which covers all of them would not be useful because of their variety. In this respect, the knowledge-based approach seems to be the most promising way for future analysis. In my view, theoreticians should collect as much information on biological sequences as possible.

ACKNOWLEDGMENT

The author is deeply grateful to Dr. Minoru Kanehisa for helpful discussion

and to Paul Horton for reading the manuscript.

REFERENCES

(1) J. Garnier and J. M. Levin, *CABIOS*, **7**, 133 (1991).
(2) Y. Seto, Y. Ikeuchi. and M. Kanehisa, *PROTEINS*, **8**, 341 (1990).
(3) A. Aitken, "Identification of Protein Consensus Sequences", Ellis Horwood, Chichester (1990).
(4) A. Varshavsky, *Cell*, **64**, 13 (1991).
(5) S. J. Landry and L. M. Gierasch, *Trends Biochem. Sci.*, **16**, 159 (1991).
(6) T. R. Hirst and R. A. Welch, *Trends Biochem. Sci.*, **13**, 265 (1988).
(7) G. Kuwajima, I. Kawagishi, M. Homma, J. Asaka, E. Kondo, and R. M. Macnab, *Proc. Natl. Acad. Sci. USA*, **86**, 4953 (1989).
(8) B. E. Uhlin, M. Båga, M. Göransson, F. P. Lindberg, B. Lund, M. Norgren, and S. Normark, *Curr, Topics Microbiol. Immunol.*, **118**, 163 (1985).
(9) J. Beckwith and S. Ferro-Novick, *Curr. Topics Microbiol. Immunol.*, **125**, 5 (1986).
(10) P. Model and M. Russel, *Cell*, **61**, 739 (1990).
(11) K. Nakai and M. Kanehisa, *PROTEINS*, **11**, 95 (1991).
(12) G. von Heijne, *Eur. J. Biochem.*, **116**, 419 (1981).
(13) G. von Heijne, *Eur. J. Biochem.*, **133**, 17 (1983).
(14) G. von Heijne, *J. Mol. Biol.*, **184**, 99 (1985).
(15) G. von Heijne, *J. Mol. Biol.*, **192**, 287 (1986).
(16) G. von Heijne, *Protein Eng.*, **2**, 531 (1989).
(17) P. Bassford, J. Beckwith, K. Ito, C. Kumamoto, S. Miyazawa, D. Oliver, L. Randall, T. Silhavy, P. C. Tai, and B. Wickner, *Cell*, **65**, 367 (1991).
(18) G. von Heijne, *Nucl. Acids Res.*, **14**, 4683 (1986).
(19) D. J. McGeoch, *Virus Research*, **3**, 271 (1985).
(20) M. Sjöström, S. Wold, Å. Wieslander, and L. Rilfors, *EMBO J.*, **6**, 823 (1987).
(21) K. Yamaguchi, F. Yu, and M. Inoue, *Cell*, **53**, 423 (1988).
(22) K. Baker, N. Mackman, and B. Holland, *Prog. Biophys. Molec. Biol.*, **49**, 89 (1987).
(23) M. S. Weiss, A. Kreusch, E. Schiltz, U. Nestel, W. Welte, J. Weckesser and G. E. Schulz, *FEBS Lett.*, **280**, 379 (1991).
(24) M. Struyvé, M. Moons and J. Tommassen, *J. Mol. Biol.*, **218**, 141 (1991).
(25) H. Vogel and F. Jähnig, *J. Mol. Biol.*, **190**, 191 (1986).
(26) J. Kyte and R. F. Doolittle, *J. Mol. Biol.*, **157**, 105 (1982).
(27) G. D. Fasman and W. A. Gilbert, *Trends Biochem. Sci.*, **15**, 89 (1990).
(28) F. Jähnig, *Trends Biochem,. Sci.*, **15**, 93 (1990).
(29) P. Klein, M. Kanehisa and C. DeLisi, *Biochim. Biophys. Acta*, **815**, 468 (1985).
(30) K. Nakai and M. Kanehisa, submitted (1991).
(31) K. Nakai, A. Kidera and M. Kanehisa, *Protein Eng.*, **2**, 93 (1988).
(32) W. T. Wickner and H. F. Lodish, *Science*, **230**, 400 (1985).
(33) G. von Heijne and Y. Gaval, *Eur. J. Biochem.*, **174**, 671 (1988).
(34) S. J. Singer, *Ann. Rev. Cell Biol.*, **6**, 247 (1990).
(35) E. Hartmann, T. A. Rapoport and H.F. Lodish, *Proc. Natl. Acad. Sci. USA*, **86**, 5786 (1989).
(36) G. D. Parks and R. A. Lamb, *Cell*, **64**, 777 (1991).
(37) D. Boyd and J. Beckwith, *Cell*, **62**, 1031 (1990).
(38) S. M. Simon and G. Blobel, *Cell*, **65**, 371 (1991).
(39) V. R. Lingappa, *Cell*, **65**, 527 (1991).
(40) K. Verner and G. Schatz, *Science*, **241**, 1307 (1988).
(41) F.-U. Hartl and W. Neupert, *Science*, **247**, 930 (1990).
(42) K. P. Baker and G. Schatz, *Nature*, **349**, 205 (1991).
(43) G. von Heijne, J. Steppuhn, and R. G. Herrmann, *Eur. J. Biochem.*, **180**, 53 (1989).
(44) D. Eisenberg, *Ann. Rev. Biochem.*, **53**, 595 (1984).
(45) Y. Gavel and G. von Heijne, *Prot. Eng.*, **4**, 33 (1990).

(46) K. Keegstra, L. J. Olsen, and S. M. Theg, *Ann. Rev. Plant Physiol. Plant. Mol. Biol.*, **40**, 471 (1989).
(47) Y. Gavel and G. von Heijne, *FEBS lett.*, **261**, 455 (1990).
(48) J. M. Hand, L. J. Szabo, A. C. Vasconcelos, and A. R. Cashmore, *EMBO J.*, **8**,3195(1989).
(49) C. J. Howe and T. P. Wallace, *Nucl. Acids Res.*, **18**, 3417 (1990).
50) Y. Gavel, J. Steppuhn, R. Herrmann and G. von Heijne, *FEBS Lett.*, **282**, 41 (1991).
(51) P. A. Silver, *Cell*, **64**, 489 (1991).
(52) L.-J. Zhao and R. Padmanabhan, *Cell*, **64**, 13 (1988).
(53) D. Kalderon, B. L. Roberts, W. D. Richardson and A. E. Smith, *Cell* **39**, 499 (1984).
(54) J. Robbins, S. M. Dilworth, R. A. Laskey and C. Dingwall, *Cell*, **64**, 615 (1991).
(55) J. Hamm, E. Darzynkiewicz, S. Tahara and I. W. Mattaj, *Cell*, **62**, 569 (1990).
(56) J. Hamm and I. W. Mattaj, *Cell*, **63**, 109 (1990).
(57) C. Query, R. C. Bentley and J. D. Keene, *Cell*, **57**, 8 (1989).
(58) T. Hunt, *Cell*, **59**, 949 (1989).
(59) M. R. Underwood and H. M. Fried, *EMBO J.*, **9**, 91 (1990).
(60) P. Borst, *Biochim. Biophys. Acta*, **866**, 179 (1986).
(61) S. J. Gould, G.-A. Keller, N. Hosken, J. Wilkinson, and S. Subramani, *J. Cell Biol.*, **108**, 1657 (1989).
(62) T. Osumi and Y. Fujiki, *BioEssays*, **12**, 217 (1990).
(63) S. J. Gould, G.-A. Keller, M. Schneider, S. H. Howell, L. J. Garrard, J. M. Goodman, B. Distel, H. Takab and S. Subramani, *EMBO J.*, **9**, 85 (1990).
(64) T. Tsukamoto, S. Miura, and Y. Fujiki, *Nature*, **350**, 77 (1991).
(65) S. M. Hurtley and A. Helenius, *Annu. Rev. Cell Biol.*, **5**, 277 (1989).
(66) S. R. Pfeffer and J. E. Rothman, *Ann. Rev. Biochem.*, **56**, 829 (1987).
(67) H. R. B. Pelham, *Trends Biochem. Sci.*, **15**, 483 (1990).
(68) G. Warren, *Cell*, **62**, 1 (1990).
(69) M. R. Jackson, T. Nilsson. and P. A. Peterson, *EMBO J.*, **9**, 3153 (1990).
(70) B. Dahllöf, M. Wallin and S. Kvist, *J. Biol. Chem.*, **266**, 1804 (1991).
(71) H. R. B. Pelham, *Annu. Rev. Cell Biol.*, **5**, 1 (1989).
(72) B. M. F. Pearse and M. S. Robinson, *Annu. Rev. Cell Biol.*, **6**, 151 (1990).
(73) W.-J. Chen, J. L. Goldstein and M. S. Brown, *J. Biol. Chem.*, **265**, 3116 (1990).
(74) J. F. Collawn, M. Stangel, L. A. Kuhn, V. Esekogwu, S. Jing, I. S. Trowbridge, and J. A. Tainer, *Cell*, **63**, 1061 (1990).
(75) M. A. Williams and M. Fukuda, *J. Cell Biol.*, **111**, 955 (1990).
(76) B. Bendiak, *Biochem. Biophys. Res. Comm.*, **170**, 879 (1990).
(77) T. L. Burgess and R.B. Kelly, *Annu. Rev. Cell Biol.*, **3**, 243 (1987).
(78) J. S. Kizer and A. Tropsha, *Biochem. Biophys. Res. Comm.*, **174**, 586 (1991).
(79) K. Simons and A. Wandinger-Ness, *Cell*, **62**, 207 (1990).
(80) T. Roitsch and L. Lehle, *Eur. J. Biochem.*, **195**, 145 (1991).
(81) S. Kornfeld and I. Mellman, *Ann. Rev. Cell Biol.*, **5**, 473 (1989).
(82) T. J. Baranski, P. L. Faust and S. Kornfeld, *Cell*, **63**, 281 (1990).
(83) J. H. Rothman, C. T. Yamashiro, P. M. Kane, and T. H. Stevens, *Trends Biochem. Sci.*, **14**, 347 (1989).
(84) T. Yoshihisa and Y. Anraku, *J. Biol. Chem.* **265**, 22418 (1990).
(85) D. Klionsky and S. D. Emr, *J. Biol. Chem.* **265**, 5349 (1990).
(86) J. F. Dice, *FASEB J.*, **1**, 349 (1987).
(87) E. R. Stadtman, *Biochemistry*, **29**, 6323 (1990).
(88) L. Devi, *FEBS Lett.*, **280**, 189 (1991).
(89) A. Bachmair and A. Varshavsky, *Cell*, **56**, 1019 (1989).
(90) E. S. Johnson, D. K. Gonda and A. Varshavsky, *Nature*, **346**, 287 (1990).
(91) A. Ciechanover and A. L. Schwartz, *Trends Biochem. Sci.*, **14**, 483 (1989).
(92) S. Rogers, R. Wells, and M. Rechsteiner, *Science*, **234**, 364 (1986).
(93) M. Rechsteiner, S. Rogers, and K. Rote. *Trends Biochem. Sci.*, **12**, 390 (1987).
(94) J. F. Dice, *Trends Biochem. Sci.*, **15**, 305 (1990).

(95) F. Wold, *Ann. Rev. Biochem.*, **50**, 783 (1981).

(96) J. C. Paulson, *Trends Biochem. Sci.*, **14**, 212 (1989).

(97) P. V. Wagh and O. P. Bahl, *CRC Crit. Rev. Biochem.*, **10**, 307 (1981).

(98) K. Nakai and M. Kanehisa, *J. Biochem. (Tokyo)*, **104**, 693 (1988).

(99) Y. Gavel and G. von Heijne, *Protein Eng.*, **3**, 433 (1990).

(100) G.W. Hart, R.S. Haltiwanger, G.D. Holt, and W.G. Kelly, *Ann. Rev. Biochem.*, **58**, 841 (1989).

(101) I. B. H. Wilson, Y. Gavel and G. von Heijne, *Biochem. J.*, **275**, 529 (1991).

(102) T. Hunter, *Cell*, **50**, 823 (1987).

(103) B. E. Kemp and R. B. Pearson, *Trends Biochem. Sci.*, **15**, 342 (1990).

(104) S. K. Hanks, A. M. Quinn, and T. Hunter, *Science*, **241**, 42 (1988).

(105) A. M. Schultz, L. E. Henderson and S. Oroszlanm, *Ann. Rev. Cell Biol.*, **4**, 611 (1988).

(106) R. J. A. Grand, *Biochem. J.*, **258**, 625 (1989).

(107) D. A. Towler, J.I. Gordon, S. P. Adams, and L. Glaser, *Ann. Rev. Biochem.*, **57**, 69 (1988).

(108) R. A. J. McIlhinney, *Trends Biochem. Sci.*, **15**, 387 (1990).

(109) M. D. Resh, *Oncogene*, **5**, 1437 (1990).

(110) M. A. Ferguson and A. F. Williams, *Ann. Rev. Biochem.*, **57**, 285 (1988).

(111) G. A. M. Cross, *Ann. Rev. Cell Biol.*, **6**, 1 (1990).

(112) W. A. Maltese, *FASEB J.*, **4**, 3319 (1990).

(113) J. F. Hancock, H. Paterson and C. J. Marshall, *Cell*, **63**, 133 (1990).

(114) D. Holtz, R. A. Tanaka, J. Hartwig, and F. McKeon, *Cell*, **59**, 969 (1989).