PROTEINS: Structure, Function, and Genetics 11:95–110 (1991)

Expert System for Predicting Protein Localization Sites in Gram-Negative Bacteria

Kenta Nakai and Minoru Kanehisa
Institute for Chemical Research, Kyoto University, Uji, Kyoto 611, Japan

ABSTRACT **We have developed an expert system that makes use of various kinds of knowledge organized as "if-then" rules for predicting protein localization sites in Gram-negative bacteria, given the amino acid sequence information alone. We considered four localization sites: the cytoplasm, the inner (cytoplasmic) membrane, the periplasm, and the outer membrane. Most rules were derived from experimental observations. For example, the rule to recognize an inner membrane protein is the presence of either a hydrophobic stretch in the predicted mature protein or an uncleavable N-terminal signal sequence. Lipoproteins are first recognized by a consensus pattern and then assumed present at either the inner or outer membrane. These two possibilities are further discriminated by examining an acidic residue in the mature N-terminal portion. Furthermore, we found an empirical rule that periplasmic and outer membrane proteins were successfully discriminated by their different amino acid composition. Overall, our system could predict 83% of the localization sites of proteins in our database.**

Key words: expert system, prediction from amino acid sequences, sorting of proteins, gram-negative bacteria, genome analysis

INTRODUCTION

The recently started human and other genome projects are likely to change the situation of molecular biology. Comprehensive analyses of whole genomic sequences will enable us to understand general mechanisms of how protein and nucleic acid functions are encoded in the sequence data. In this respect, we have been developing computational methods to interpret genetic information. In the past, we used mostly multivariate analysis;[1,2] for example, a method for predicting in vivo modification sites of proteins from their amino acid sequences was presented.[3] Such a statistical method is useful to define consensus patterns and other sequence motifs that characterize specific functional sites. However, there is a great deal of additional experimental evidence that suggests important connections between sequence data and functional aspects. It is not easy to implement ways to detect and interpret these experimental features on computers since they are often fragmentary and context-dependent. Someone with an expert's knowledge would be required to use them in order to reason functional aspects of newly determined DNAs and proteins. With an expert system approach, we can perform similar processes on computers.

Expert systems have been developed in the field of artificial intelligence.[4] They are computer programs that manipulate experts' domain-specific knowledge and heuristics, organized in a knowledge base, to solve problems in a narrow and realistic problem area. Among various types of expert systems, one of the most classical and widely used is the production system, in which the knowledge base is realized as a collection of "if-then" rules or, what is called, production rules. We adopt this type of expert system and occasionally use the term expert system to mean production system. The knowledge base is one of the three essential components of an expert system; the other two components are the inference engine and the working memory. The working memory reflects changing states of inference, storing initial conditions, intermediate hypotheses, and so on. The inference engine selects rules that satisfy each state of the working memory by a pattern-matching mechanism. If there are competing rules, that is, if multiple rules are matched to the same situation, only one of them is selected by a "conflict resolution" procedure. The invoked rule will modify the working memory, which will cause another cycle of pattern-matching. Thus appropriate rules are invoked automatically and context-dependently in an expert system. This approach is advantageous to the types of problems in which algorithmic solutions are not easily obtainable. In the field of molecular biology, knowledge-based approaches have been tried, for example, for modeling of unknown protein structures from the knowledge of structurally homologous segments of proteins.[5]

Here we present an expert system for predicting

Received September 10, 1990; revision accepted January 28, 1991.
Address reprint requests to Dr. M. Kanehisa, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611, Japan.

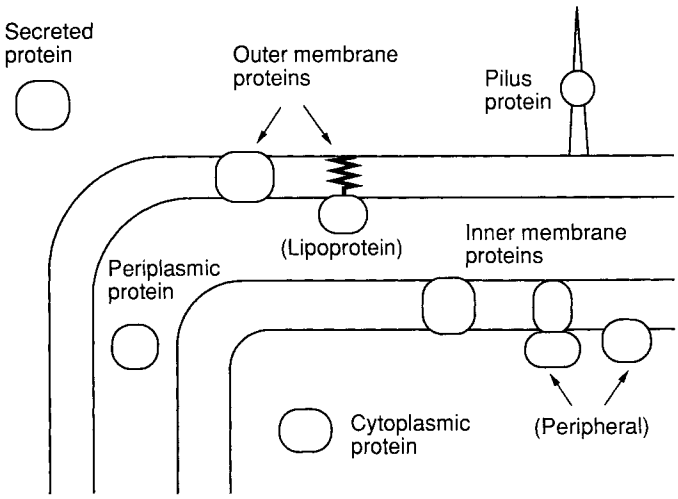

Fig. 1. Various protein localization sites in Gram-negative bacteria. Basically, proteins localized at four sites have been considered: cytoplasm, inner (cytoplasmic) membrane, periplasmic space, and outer membrane. Lipoproteins, which are localized at either membrane, have a lipid molecule covalently attached to their surface. Peripheral membrane proteins localize at the surface of a membrane and can be solubilized easily.

protein localization sites in cells from amino acid sequences. Our system diagnoses the destination of a given amino acid sequence like an experienced doctor, thus resembling classical examples of expert systems.[4] There are three main reasons why we chose this subject. First, the information about cellular translocation sites of newly synthesized proteins seems to be primarily encoded in the amino acid sequence and there is an extensive and growing body of experimental evidence for sorting signals. Second, because most of these signals cannot be recognized by simple consensus sequence patterns, an approach to combine different weak features in sequences seems to be promising. Third, the localization site of a protein is selected from many possibilities in actual cellular processes; thus it is necessary to consider many alternatives and evaluate the relative strength of each signal.

In this work, we focus our attention on the sorting of prokaryotic proteins, especially those of Gram-negative bacteria.[6] Although the problem of protein sorting is usually considered in eukaryotic cells, even in bacterial cells proteins must be localized in appropriate locations in order to exert specific functions. The framework of our system can deal with both prokaryotic and eukaryotic sequences, and the sorting of eukaryotic proteins will be dealt with in a forthcoming study (Nakai and Kanehisa, in preparation).

MATERIALS AND METHODS

Sequence Data

Amino acid sequence data of *Escherichia coli* and other Gram-negative bacteria were identified as entries of the NBRF-PIR database, release 22.0.[7] Only nascent forms (precursors) were used because processed forms may have lost a part of their sorting signals. In addition, complete sequences starting with methionine were collected, although some exceptions exist for cytoplasmic proteins. In order to avoid redundancy of data, a representative sequence was chosen from each protein family as classified by the PIR; there is no pair of representatives that has more than 50% identical residues. The total number of sequences used in the present study is 106.

Each sequence is categorized by its localization site (Fig. 1). The set of localization sites for Gram-negative bacteria includes: the cytoplasm, the inner (cytoplasmic) membrane, the periplasmic space, and the outer membrane. Secreted proteins and a few other categories of proteins, such as pili and flagella, were not considered (see Discussion). The localization sites were obtained partly from comments and keywords from the NBRF-PIR database, and mostly from many references. Among them, Watson's[8] and Sjöström et al.'s[9] compilations were especially useful for collecting Gram-negative bacterial sequences. The selection of cytoplasmic proteins was done rather arbitrarily because there were too many cytoplasmic proteins. The data set used for the construction of the expert system is summarized in Table I.

OPS83 Programming Language

The core part of our expert system was written in the programming language OPS83 (ver. 2.2),[10,11] which is especially suited for constructing a production system. Our knowledge base is organized as a set of if-then rules such as shown in Table II. In OPS83, the working memory is realized as a set of record structures called working memory elements. As for the inference engine, OPS83 users can freely

TABLE I. Proteins Used in This Analysis*

(a) Inner Membrane Proteins

Entry	McG	GvH(sig) scr	GvH(sig) pos	Lip	Clv	Klein score	Klein cnt	Length	Prd	Definition
DEECDL	−−	−	27	−	?	(+)	1	(571)	IM	D-Lactate dehydrogenase—*E. coli*
DEECS2	(−)	(−)	57	−	?	+++	3	(115)	IM	Succinate dehydrogenase 13K hydrophobic protein—*E. coli*
DEECS4	−	(−)	54	−	?	++	3	(129)	IM	Succinate dehydrogenase 14K hydrophobic protein—*E. coli*
DEECXA	−−	−	39	−	?	+	5	(502)	IM	$NAD(P)^+$ transhydrogenase (B-specific) α-chain—*E. coli*
DEECXB	+++	(−)	23	−	?	++	8	(462)	IM	$NAD(P)^+$ transhydrogenase (B-specific) β-chain—*E. coli*
RDECNG	+	(−)	18	−	?	++	4	(225)	IM	Nitrate reductase γ chain—*E. coli*
WQEC2G	(+)	(+)	49	−	?	++	9	(477)	IM	Phosphotransferase enzyme II, glucose-specific—*E. coli*
PSECL2	−−	−	16	−	?	+	1	(340)	IM	Lysophospholipase L2—*E. coli*
ZPECS	++	−	40	−	?	++	2	(323)	IM	Signal peptidase—*E. coli*
PWECAK	++	−	20	−	?	++	11	(557)	IM	Potassium-transporting ATPase, A-chain—*E. coli*
PWECBK	−	(+)	56	−	?	+++	6	(682)	IM	Potassium-transporting ATPase, B-chain—*E. coli*
PWECCK	+++	(−)	40	−	?	−	0	(190)	PP	Potassium-transporting ATPase, C-chain—*E. coli*
QRECBC	++	(−)	33	−	?	++	6	(292)	IM	Vitamin B12 transport protein *btuC*—*E. coli*
QRECBD	−−−	−	52	−	?	−	0	(249)	CP	Vitamin B12 transport protein *btuD*—*E. coli*
MMECMK	−−	−	14	−	?	−	0	(370)	CP	Inner membrane protein *malK*—*E. coli*
QREBPT	−−	(−)	31	−	?	−	0	(258)	CP	Histidine permease inner membrane receptor protein P—*S. typhimurium*
QREBOT	−	−	28	−	?	(+)	1	(335)	IM	Oligopeptide permease membrane protein *oppD*—*S. typhimurium*
MMECMF	+	(−)	38	−	?	++	6	(514)	IM	Inner membrane protein *malF*—*E. coli*
ZPECPA	+++	−	19	−	?	++	1	(850)	IM	Penicillin-binding protein 1A—*E. coli*
ZPECPB	−−−−	−−	13	−	?	++	1	(844)	IM	Penicillin-binding protein 1B—*E. coli*
GREC	++	−	24	−	?	++	11	(417)	IM	Lactose permease—*E. coli*
ZPECP3	−	(−)	36	29	40?	++	1	(588)	IM	Penicillin-binding protein 3 precursor—*E. coli*
ZPECP2	−	−	36	−	?	++	1	(633)	IM	Penicillin-binding protein 2—*E. coli*
ZPECP5	−	+	29	−	29	(−)	0	(403)	IM	Penicillin-binding protein 5 precursor—*E. coli*
JGECPP	+	(−)	29	−	?	++	9	(502)	IM	Proline carrier protein—*E. coli*
MMECMG	++	−	34	−	?	+++	6	(296)	IM	Maltose transport inner membrane protein *malG*—*E. coli*
LPEC28	++	(+)	31	23	23	−	0	(272)	IM	Lipoprotein-28 precursor—*E. coli*
BVECLA	−	−	21	−	?	−	0	(598)	CP	*lepA* protein—*E. coli*
A24400	(−)	−	38	−	?	−	0	(311)	IM	Nodulation protein I—*Rhizobium leguminosarum*
S01377	+++	−	26	−	?	++	2	(458)	IM	Chemotaxis protein *cpxA*—*E. coli*
B26871	−−	−	34	−	?	(−)	0	(300)	CP	Molybdenum transport protein *chlD*—*E. coli*
S01367	++	−	13	−	?	++	2	(450)	IM	Inner membrane protein *envZ*—*S. typhimurium*
B29333	++	−	32	−	?	++	6	(306)	IM	Periplasmic oligopeptide-binding protein *oppB*—*S. typhimurium*
C29333	−−−−	(−)	56	−	?	+++	5	(302)	IM	Periplasmic oligopeptide-binding protein *oppC*—*S. typhimurium*

(continued)

TABLE I. Proteins Used in This Analysis* (Continued)

(b) Periplasmic Proteins

Entry	McG	GvH(sig) scr	GvH(sig) pos	Lip	Clv	Klein score	Klein cnt	Length	Prd	Definition
QRECBE	−−	−−	38	−	?	(−)	0	(183)	CP	Vitamin B12 transport protein *btuE*—*E. coli*
PAECA	+	+	21	−	21	−	0	(471)	PP	Alkaline phosphatase precursor—*E. coli*
YXECUG	+	(+)	25	−	25	−	0	(550)	PP	UDP-glucose hydrolase precursor—*E. coli*
ESECPC	+	+	19	−	?	−	0	(646)	PP	2′-3′-Cyclic-nucleotide 2′-phosphodiesterase—*E. coli*
PNECP	+	+	23	−	23	(−)	0	(286)	PP	β-lactamase precursor—*E. coli* plasmids
QKEC	++	+	19	−	19	−	0	(377)	PP	β-lactamase precursor—*E. coli*
BYEC	+	++	19	−	19	−	0	(329)	PP	Sulfate-binding protein—*E. coli*
JHEBT	+	+	22	−	22	−−	0	(260)	PP	Histidine-binding protein J precursor—*S. typhimurium*
QRECFD	(+)	+	30	−	30?	+	1	(296)	IM	Ferrichrome-iron transport protein *fhuD* precursor—*E. coli*
BYECPR	+	+	25	−	25	(+)	1	(346)	IM	Phosphate-repressible phosphate-binding protein precursor—*E. coli*
BVECMM	++	(−)	20	−	22?	(−)	0	(306)	PP	*malM* protein—*E. coli*
JGECR	+	+	25	−	25	(−)	1	(296)	PP	D-Ribose-binding protein precursor—*E. coli*
JGECM	+	++	26	−	26	−	0	(396)	PP	Maltose-binding protein precursor—*E. coli*
S03711	+	++	23	−	?	−	0	(329)	PP	L-Arabinose-binding protein—*E. coli*
A23576	++	++	23	−	23	−	0	(367)	PP	LIV-binding protein precursor—*E. coli*
A25011	++	(+)	26	−	?	−	0	(542)	PP	Periplasmic oligopeptide-binding protein precursor—*S. typhimurium*
A26509	+	+	28	−	20?	−−	0	(230)	PP	Deoxyribonuclease precursor—*Vibrio cholerae*
B28380	−−	−	57	−	?	−	0	(518)	CP	Hydrogenase large chain—*Desulfovibrio baculatus*
A28380	(+)	−	32	−	32	(−)	0	(315)	IM	Hydrogenase small chain precursor—*Desulfovibrio baculatus*
B27492	−−	−−	44	−	?	−−	0	(560)	CP	Hydrogenase large chain—*Desulfovibrio gigas*
A27492	+	(+)	19	−	25	−	0	(291)	PP	Hydrogenase small chain precursor—*Desulfovibrio gigas*

(continued)

implement a conflict resolution strategy; we adopted a simplified version of the MEA strategy[10] modified to enable the use of certainty factors. A certainty factor represents the degree of certainty about the conclusion deduced from a rule. There are two basic ways in which rules can be used: one is called forward chaining and the other backward chaining. In the former, inference starts from an initial set of facts and continues until a certain conclusion is obtained; in the latter, inference starts from a certain hypothesis and attempts to prove it by examining its preconditions. Although OPS83 is primarily suited for forward chaining, one can simulate backward chaining using the working memory as a kind of stack. We mainly implemented our rules for backward chaining because it enabled us to examine all possible hypotheses extensively. In other words, not only the most probable hypothesis but also hypotheses of lower likelihood are fully examined. Our system provides alternative sites using certainty factors, which might enhance the usefulness and reliability of prediction.

System Architecture

The overall architecture of our system is shown schematically in Figure 2. Each rectangle represents a file containing one module. Lines connecting modules are drawn to show their relations. These

TABLE I. Proteins Used in This Analysis* (Continued)

(c) Outer Membrane Proteins

Entry	McG	GvH(Sig) scr	GvH(Sig) pos	Lip	Clv	Klein score	Klein cnt	Length	Prd	Definition
PSECA1	+	(−)	20	−	20	−	0	(289)	OM	Phospholipase A1 precursor (version 1)—*E. coli*
QRECFC	+	(+)	22	−	22	−	0	(745)	OM	Ferrienterochelin receptor precursor—*E. coli*
MMECOF	(−)	(+)	35	−	35	(−)	0	(812)	OM	Outer membrane *faeD* protein precursor—*E. coli*
MMECTC	+ +	+	22	−	22	−	0	(489)	OM	Outer membrane protein *tolC* precursor—*E. coli*
MMECF	+ +	+	22	−	22	−	0	(362)	OM	Outer membrane protein F precursor—*E. coli*
MMECA	+ +	+ +	21	−	21	−	0	(346)	PP	Outer membrane protein A precursor—*E. coli*
LPECPG	+	(+)	20	21	21?	−−	0	(173)	OM	Peptidoglycan-associated lipoprotein precursor—*E. coli*
LPECW	+ +	(+)	25	20	20	−−−	0	(78)	OM	Murein-lipoprotein precursor—*E. coli B*
QRECL	+	+ +	25	−	25	−−	0	(446)	OM	λ receptor protein precursor—*E. coli*
QRECBT	(+)	+	20	−	?	(−)	0	(614)	OM	Vitamin B12 receptor—*E. coli*
QRECFE	+	+	33	−	33	−	0	(747)	OM	Ferrichrome-iron receptor precursor—*E. coli*
BVECTJ	−	−	58	−	?	−	0	(229)	CP	*traJ* protein—*E. coli* plasmid F
MMBPL	+ +	(+)	21	−	?	−	0	(206)	OM	Possible membrane protein *lom*—Bacteriophage λ
S01042	+	(+)	23	−	?	−	0	(725)	OM	Cloacin receptor precursor—*E. coli*
S01751	+	(−)	20	−	?	−	0	(317)	OM	Gene *ompT* protein—*E. coli*
A25647	+	+ +	23	−	?	−−	0	(355)	OM	Outer membrane pore protein *nmpC* precursor—*E. coli*
S01757	+ +	(−)	28	21	?	(+)	1	(244)	OM	*traT* protein precursor—*E. coli* plasmid F
A29840	+	(+)	27	−	27	−	0	(1045)	OM	Extracellular serine protease precursor—*Serratia marcescens*
A27872	+	(+)	19	−	?	(−)	0	(257)	OM	Outer membrane protein *ompV* precursor—*Vibrio cholerae*
A28787	+ +	+	22	−	22	−	0	(459)	OM	Outer membrane protein P1—*Haemophilus influenzae* (type b)
A27558	+	(−)	16	19	?	−−	0	(153)	OM	Outer membrane protein P6 precursor—*Haemophilus influenzae*
B28543	+ +	(+)	18	18	?	(+)	1	(154)	OM	PAL cross-reacting lipoprotein precursor—*Haemophilus influenzae*

(continued)

modules were written following the example presented by Tanaka and Shimoi.[11] The module controlling the whole program is in PSORT.OPS. PSEQ.C is the sequence input routine, written in the C language, which can recognize several formats of typical amino acid sequence databases. Scientific names as well as code names can also be read from the NBRF-PIR database. Other kinds of user interfaces, such as the tracing of reasoning steps, are dealt with in the PUTIL.OPS section. In PEDF.OPS, types of working memory elements are defined and variables are initialized. The inference engine is realized in the COMMON.OPS module. PRULEM.OPS stores so-called metarules, which are "rules for rules" controlling the stream of reasoning.

The knowledge base is conventionally divided into three modules, corresponding to the order of reasoning steps. PRULE1.OPS includes rules for classifying scientific names into appropriate categories, making use of the powerful pattern-matching capability of OPS83. Although not reported in this study, our system is designed to invoke appropriate rules depending on the organism in question. The distinction between PRULE2.OPS and PRULE3.OPS is somewhat arbitrary; the former deals with common and basic knowledge and the latter deals with more specialized knowledge. In both modules, functions

TABLE I. Proteins Used in This Analysis* (Continued)

(d) Cytoplasmic Proteins

		GvH(sig)				Klein				
Entry	McG	scr	pos	Lip	Clv	score	cnt	Length	Prd	Definition
RDECD	---	(-)	19	-	?	-	0	(159)	CP	Dihydrofolate reductaase, type I—*E. coli*
DTECR	--	--	33	-	?	-	0	(153)	CP	Aspartate carbamoyltransferase R chain—*E. coli*
KIECFA	-	(-)	19	-	?	(+)	1	(320)	IM	6-Phosphofructokinase 1—*E. coli*
KIECA	(-)	(-)	17	-	?	(-)	0	(214)	PP	Adenylate kinase—*E. coli*
RNECA	---	-	24	-	?	-	0	(329)	CP	DNA-directed RNA polymerase α chain—*E. coli*
DJECI	-	(-)	58	-	?	(-)	0	(928)	CP	DNA-directed DNA polymerase I—*E. coli*
NCECX1	---	-	32	-	?	(-)	0	(466)	CP	Exodeoxyribonuclease I—*E. coli*
NDECPB	-	(-)	29	-	?	-	0	(276)	CP	Endonuclease EcoRI—*E. coli* plasmids pMB1 and pMB4
NRECH	--	-	58	-	?	-	0	(155)	CP	Ribonuclease H—*E. coli*
GBEC	--	-	39	-	?	(-)	0	(1023)	CP	β-galactosidase—*E. coli*
QYEC	-	-	47	-	?	(-)	0	(883)	CP	Phosphoenolpyruvate carboxylase—*E. coli*
NNSE2	---	-	13	-	?	-	0	(192)	CP	Anthranilate synthase component II—*Serratia marcescens*
WZEC	---	(-)	58	-	?	-	0	(471)	CP	Tryptophanase—*E. coli*
TSECA	--	-	47	-	?	(-)	0	(268)	CP	Tryptophan synthase α-chain—*E. coli*
TSECB	-	-	59	-	?	(-)	0	(396)	CP	Tryptophan synthase β-chain—*E. coli*
ISECTP	---	(-)	14	-	?	-	0	(864)	CP	DNA topoisomerase—*E. coli*
SYECYT	---	-	45	-	?	-	0	(423)	CP	Tyrosyl-tRNA synthetase—*E. coli*
AJECNA	--	-	34	-	?	(-)	0	(330)	CP	Asparagine synthetase—*E. coli*
DDEC	(+)	-	47	-	?	--	0	(178)	IM	Helix-destabilizing protein—*E. coli*
IQECAA	---	-	14	-	?	---	0	(99)	CP	Integration host factor (IHF), α-chain—*E. coli*
R3EC1	-	(-)	49	-	?	(+)	1	(557)	IM	30S ribosomal protein S1—*E. coli*
R5EC1	-	-	40	-	?	-	0	(233)	CP	50S ribosomal protein L1—*E. coli*
RPECL	---	-	34	-	?	(-)	0	(360)	CP	*lac* repressor—*E. coli*
EFECT	--	(+)	46	-	?	(-)	0	(394)	CP	Elongation factors Tu—*E. coli*
FIEC1	---	-	33	-	?	--	0	(71)	CP	Initiation factor IF-1—*E. coli*
TWECR	-	-	46	-	?	-	0	(419)	CP	Transcription termination factor ρ—*E. coli*
IQECDA	--	(-)	31	-	?	-	0	(467)	CP	*dnaA* protein—*E. coli*
RQECA	-	-	55	-	?	(-)	0	(352)	CP	*recA* protein—*E. coli*
QRECC	--	--	19	-	?	-	0	(210)	CP	cAMP receptor protein (CAP)—*E. coli*

*Entry: entry name from the NBRF-PIR database.
McG: discriminant score based on McGeoch's method. The magnitude of the score is represented by the number of plus or minus signs and by parentheses for marginal values.
GvH(sig): result of von Heijne's method for signal sequences: scr, von Heijne's score subtracted by 7.5; pos, predicted cleavage site.
Lip: result of von Heijne's method for lipoproteins.
Clv: experimentally determined cleavage site. A ? sign indicates *unknown*.
Klein: the result of Klein et al.'s method applied to the predicted mature sequence: score, maximum discriminant score; cnt, number of transmembrane segments.
Length: length of the precursor sequence.
Prd: predicted localization sites: IM, inner membrane; PP, periplasmic space; OM, outer membrane; and CP, cytoplasm.

that actually handle sequence data are called fromthe PSUBS.C section. Most rules in these two modules are for backward reasoning. Thus in the first reasoning step, all possible localization sites are stacked in the working memory. In the following steps, each site is activated one by one as a hypothesis and checked for its plausibility. Whenever a rule is found applicable to a hypothetical site, its certainty factor is updated. In order to avoid repetition of the same calculation, the results of calculations for earlier hypotheses are stored in the working memory for later use in other hypotheses. When the reasoning is terminated, that is, when no rules matching the current status of the working memory exist, the site with the highest certainty factor becomes the predicted localization site. The system can also list other possible sites with their certainty factors.

TABLE II. Examples of Rules in Knowledge Base*

```
—The 2nd Step: Hypothesis Selection

rule gvh1—G.von Heijne's prediction method for signal sequences
  {
        (if the C-function "gvh" has not been called yet);
  →
        (calculate the score and possible cleavage site by "gvh" from the sequence of N-terminal 100 residues at
          most);
        (store these values in the "calc" element in the Working Memory);
  };

rule mcg1—McGeoch's prediction method for signal sequences
  {
        (if the C-function "mcg" has not been called yet);
  →
        (calculate the length of UR, the peak value of UR, and the net charge of CR by "mcg" from the sequence
          of N-terminal 100 residues at most);
        (store these values in the "calc" element in the WM);
  };

rule sig1—Put the results for signal sequences together
  {
        (if the output of "lipop," the C-function for the recognition of signal sequences of lipoproteins, has
          already been called);
        (and if the outputs of "gvh" have already been stored in the WM);
        (and if the discriminant score from "mcg" has already been stored in the WM);
  →
        (store the value in another "calc" element in the WM);
        (if the result of "lipop" is positive) {
              —>>> May be a lipoprotein (cleavable signal)
            (store the probable N-terminus of the mature form with certainty factor in another "calc" element in
              the WM);
        } else if (both the scores of "mcg" and "gvh" exceed −2.5) {
              —>>> Seems to have a cleavable N-term signal seq.
            (store the probable N-terminus of the mature form with calculated certainty factor in another "calc"
              element);
        } else if (the score of "mcg" exceeds −2.5
            while that of "gvh" is less than −2.5) {
              —>>> Seems to have an uncleavable N-term signal seq.
            (make another "calc" element stating that it seems to have an uncleavable signal with calculated
              certainty factor);
        } else {
              — >>> Seems to have no N-terminal signal seq.
            (make another "calc" element stating that it does not seem to have a signal sequence with
              calculated certainty);
        };
  };

rule alom1—Klein et al's method for membrane-spanning regions
  {
        (if the C-function "alom" has not been called yet);
        (and if the presence or absence of N-terminal signal has been examined already);
  →
        (set the N-terminal position of the mature form according to the examination result of signal sequence);
        (calculate the number of occurrences of potential membrane-spanning regions and the maximum score
          among them by 'alom' from the deduced mature form);
        (transform the maximum score);
        (store these results in the "calc" element);
  };

rule iom1—Lipoproteins reside in either inner or outer membrane
{
        (if there is an active "goal" element representing either the inner membrane or outer membrane in the
          WM at step2);
        (and if there is also a "calc" element, which claims that it is a lipoprotein);
          [1.0];—Certainty Factor
  →
        (modify the certainty factor corresponding to the "goal" element into +0.7);
  };
```

(continued)

TABLE II. Examples of Rules in Knowledge Base* (Continued)

```
rule icm1—Inner Membrane
  {
        (if there is an active "goal" element representing the inner membrane in the WM at step2);
        (and if the result of "alom" shows that there are any hydrophobic stretches in the mature form);
         [1.0];
  →
        (modify the certainty factor corresponding to the "goal" element into the value calculated from the score
          of "alom");
  };
rule ompp2—Outer Membrane or Periplasm
  {
        (if there is an active "goal" element representing either the outer membrane or the periplasm in the WM
          at step2);
        (and if it seems to have a cleavable signal sequence);
        (and if it does not seem to be a lipoprotein);
        (and if the result of "alom" shows that there are no hydrophobic stretches in the mature form);
        (and if the amino acid composition of the mature protein has not been calculated yet);
         [1.0];
  →
        (set the N-terminal position of the mature form according to the examination result of the signal
          sequence);
        (calculate the amino acid composition of the mature protein);
        (calculate the discriminant score from the composition);
        (store the score in a "calc" element);
  };
```

—The 3rd Step: Hypothesis Verification

```
rule pps1—periplasmic space
  {
        (if there is an active "goal" element representing the periplasm in the WM at step3);
        (and if the discriminant score from the amino acid composition of the mature protein has already been
          calculated);
        (and if it does not seem to be a lipoprotein);
         [1.0];
  →
        if (the value of certainty factor corresponding to the periplasm is already positive) {
            (calculate the certainty, which should not exceed 0.7, from the discriminant score);
            (modify the certainty factor from the value calculated above);
        };
  };
rule imb2—Treatment of the exception (ZPECP5)
  {
        (if there is an active "goal" element representing the inner membrane in the WM at step3);
        (and if it does not seem to have a signal sequence but its certainty is sufficiently low);
        (and if the result of "alom" shows that there are no hydrophobic stretches in the mature form);
         [1.0];
  →
        (modify the certainty factor of the "goal" element to +0.1);
  };
rule imb3—Uncleavable Signal
  {
        (if there is an active "goal" element representing the inner membrane in the WM at step3);
        (and if it seems to have an uncleavable signal sequence);
        (and if the result of "alom" shows that there are no hydrophobic stretches in the mature form);
         [1.0];
  →
        (modify the certainty factor of the "goal" according to the certainty having an uncleavable signal);
  };
```

*Some key rules for reasoning are shown in a pseudo code. As described in the main text, rules are divided into LHS and RHS by the symbol "→". Otherwise, the"—" sign shows the beginning of a comment to the end of the line.

RESULTS

Recognition of Signal Sequence

In prokaryotic cells as well as eukaryotic ones, the signal sequence (also called the signal peptide) is thought to be recognized as the first sorting signal. Thus our system first examines its existence. The method used is mainly based on McGeoch's.[12] Although McGeoch's method was originally intended

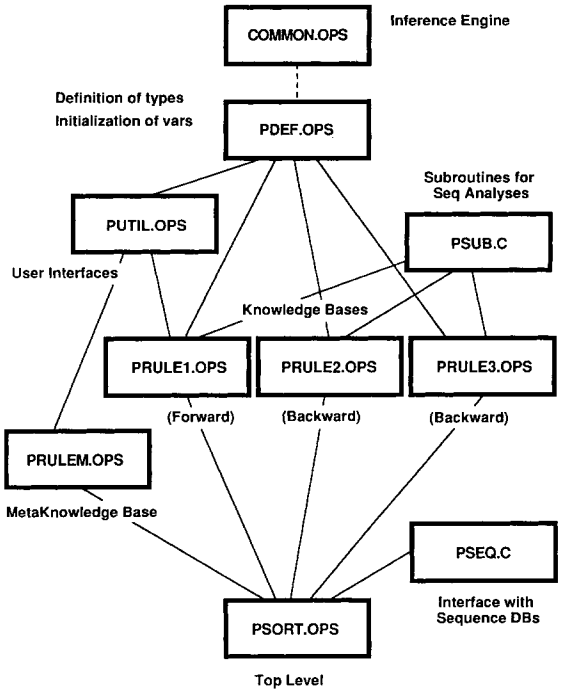

Fig. 2. Illustration of the system architecture. Each rectangle shows a module. Lines connecting modules roughly represent relationships where lower modules use upper ones.

for predictive recognition of signal sequences of eukaryotic types like HSVs, we have found that it is also useful for the recognition of prokaryotic ones. In his analysis, two regions were defined: the N-terminal region, which is often charged (abbreviated as CR), and the following uncharged (hydrophobic) stretch (abbreviated as UR). Then four parameters were examined as candidate predictors: namely, the length and the net charge of the CR, the length of the UR, and the degree of hydrophobicity of the 8-residue maximal hydrophobic region. He found that the latter two parameters were particularly effective when combined into a two-dimensional plot.

We refined his method by using discriminant analysis and by including the net charge of the CR as a third variable. The discriminant function, which is a linear combination of given variables, was obtained from a set of eukaryotic sequences with known signal peptides. The obtained formula is:

$$\begin{aligned}\text{Score} = {} & 0.305 * (\text{Length of UR}) \\ & + 6.991 * (\text{Peak Hydrophobicity}) \\ & + 0.668 * (\text{Net Charge of CR}) - 18.204\end{aligned}$$

A sequence is likely to contain a signal peptide if the score is nonnegative.

Although the formula was obtained from eukaryotic sequences, the existence of signal sequences turned out to be predicted satisfactorily in prokaryotic sequences as well. In Table I, the discrimination score is represented by a set of ± signs in the column McG. We can see, for example, almost all proteins, cleavage sites of which are experimentally determined and noted in Table I (column Clv), are predicted to have signal sequences. Similarly, all but one cytoplasmic protein are predicted not to have them. Moreover, it appears that the higher the discrimination score is, the more reliable the result of prediction becomes. Thus we used this score for the calculation of a certainty factor.

Note that this method does not use the sequence information around the cleavage site. Thus we can assume that McGeoch's method is also applicable to proteins with uncleavable signal sequences. For actual prediction, we incorporated cleavage information as described below. The result of discrimination scores in Table I suggests that most, if not all, of periplasmic and outer membrane proteins are likely to have N-terminal signal sequences, but that is not the case for inner membrane proteins where most of them may have internal signal sequences.

Signal Sequence of Lipoproteins

In prokaryotic cells, two kinds of signal peptidases are known to exist. One is signal peptidase I (Lep), which cleaves signal sequences of most proteins translocated through the inner membrane; the other is signal peptidase II (Lsp), which cleaves those of lipoproteins. Our system examines the existence of signal peptidase II cleavage sites before the existence of signal peptidase I cleavage sites because of the higher reliability of predicting the former. According to von Heijne,[13] signal sequences of lipoproteins are significantly different from other signal sequences only in the region close to the cleavage site. We incorporated his consensus search method to discriminate these two groups with one modification. The C-terminal end of the N-terminal charged region was defined from McGeoch's CR region.

We assume that all lipoproteins reside in either the inner or outer membrane through interaction with their lipid moiety (see Discussion). Presently, our database contains only a small number of lipoproteins: 1 inner membrane protein and 3 outer membrane proteins. When von Heijne's method was applied to the data set shown in Table I, all of them were correctly discriminated. These proteins were also positive with McGeoch's method supporting von Heijne's conclusion that both types of signal sequences are similar except for the cleavage site. Although there were two additional proteins with positive results by von Heijne's method, those proteins cannot be clearly classified as nonlipoproteins (see Discussion).

Cleavage of Signal Sequence

The signal sequence cleaved by signal peptidase I has a characteristic sequence pattern around its cleavage site, which is likely to represent the sub-

strate specificity of the enzyme. von Heijne proposed the "(-3,-1)-rule" for recognizing this pattern, which he later elaborated by a method based on a weight-matrix approach.[14,15] We used his weight-matrix for supplementing McGeoch's method to recognize the presence of signal sequences, for judging whether they are cleaved or not, and for deducing the mature product after cleavage. From the distribution of scores in Table I, we adopted the value 7.5 for the critical score in his weight-matrix approach for prokaryotes. In addition, we limited the range of search for the maximum score to up to 50 residues from the N-terminal.

The results in columns McG and GvH in Table I suggest that McGeoch's and von Heijne's methods are comparable in detecting signal sequences, but they provide additional clues when used in combination. The comparison of the two methods is very important for the analysis of inner membrane proteins because the existence of uncleavable N-terminal signal sequences causes localization at the inner membrane in our scheme. In fact, quite a few inner membrane proteins in Table I are positive in McGeoch's test but are negative in von Heijne's method. The cleavage sites reported in the literature are also listed in parentheses, and they are in good agreement with the prediction results.

Detection of Membrane Spanning Regions

Detection of membrane spanning regions by examining hydrophobicity of amino acid sequences has been done by several researchers.[16, 17] We used Klein et al.'s method for the recognition of inner membrane proteins.[18] However, as stated later, this method turned out to be not applicable to outer membrane proteins of Gram-negative bacteria. Klein et al.'s method is based on the maximum hydrophobicity of 17 residue segments with the parameters determined by discriminant analysis between integral and peripheral membrane proteins. It is also applicable to the discrimination between integral membrane proteins and soluble globular proteins with relatively high reliability. We did not update the discrimination parameters and used them as originally reported.

We used two variables to characterize a membrane protein: the discrimination score of the most hydrophobic segment and the number of predicted membrane-spanning segments (see Table I, column Klein). In many cases, the former turned out to be sufficient. However, for example, if a protein is predicted to have more than three membrane-spanning segments, we can assume rather convincingly that it is an inner membrane protein. As can be seen in Table I, most inner membrane proteins are predicted to have transmembrane segments. Some exceptions exist, but they seem to be peripheral rather than integral membrane proteins (see Discussion). However, some cytoplasmic proteins are falsely predicted to have transmembrane regions.

When applied to precursor sequences, this method reported many signal sequences as membrane spanning segments regardless of their cleavage (data not shown). Therefore, in order to classify inner membrane proteins, it is necessary to first determine the mature portion after cleavage and then to examine trans membrane segments by their hydrophobicity profile.

Further Discrimination of Lipoproteins

As described earlier, some lipoproteins are exclusively localized in the outer membrane and others in the inner membrane. Therefore, a certain factor for determining the final location of these proteins must exist. Here, the experiment of Yamaguchi et al. seems to provide a clue.[19] According to them a short sequence, especially the second residue from the N-terminal end, of a mature lipoprotein plays an essential role.

Based on their observation, we assumed that a lipoprotein is sorted to the inner membrane if it has a negatively charged residue at the second or third position of the mature sequence. This rule holds for all lipoproteins at inner and outer membranes in our small database. Obviously further experiments to clarify the sorting mechanism and/or expansion of the database with defined location sites are needed for the refinement of this rule.

The Problem of Outer Membrane Proteins

From the results with several tools described above, all outer membrane proteins seem to have cleavable signal sequences, but not to have any other detectable features such as hydrophobic stretches characteristic to inner membrane proteins. This was also true for periplasmic proteins. Therefore, there was no way to distinguish these two classes of proteins. Furthermore, neither the actual sorting mechanism of outer membrane proteins nor the encoded sorting signals have been determined experimentally.[20]

We found that there was a sufficient difference in the amino acid composition in the mature part of proteins between the two classes. First, we performed principal component analysis with the percentage values for the 20 amino acid compositions as variables. That is, the coefficients of linear combinations of variables were calculated to maximize the total variance of the population containing member sequences from the two groups. We then took the first two principal components and calculated the values for each member. Thus Figure 3 shows the distribution of members in the 20-dimensional space projected onto the plane. As can be seen in this plot, the two groups are separated quite well, especially by the first principal component. Note that throughout the calculation these two groups were not dis-

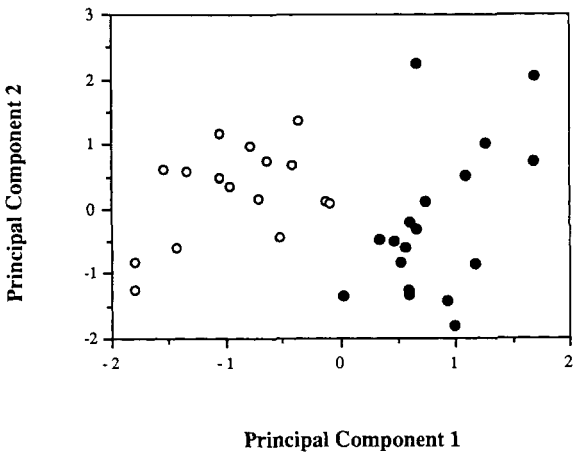

Fig. 3. Principal component analysis of the amino acid composition of the mature part of the periplasmic proteins (filled circles) and the outer membrane proteins (open circles). Only the proteins that are predicted to be cleaved by signal peptidase I were used. Note that information for distinguishing the two classes was not used in the analysis.

tinguished, i.e., it was *not* the result of setting parameters as to distinguish two groups maximally. Therefore, the difference in the amino acid composition is an intrinsic difference between the two groups. In addition, the amino acids that were most significantly different in their compositions could be identified by the coefficients of the first principal component. They were: Tyr (0.82), Pro (−0.74), Glu (−0.61), and Asn (0.56), where coefficients are given in parentheses and negative values mean favorable toward the periplasm class.

In order to actually distinguish between these two classes of proteins, we performed a stepwise discriminant analysis and obtained an optimized set of parameters for discrimination. The selected variables corresponded to the contents of Tyr, Ser, Pro(−), Lys(−), His, Gly, Asn, Thr, Phe and Cys(−), where the negative signs represent favorability toward the periplasm class. With the derived discriminant score, all but one of the proteins were correctly discriminated.

Organization of Knowledge

The reasoning process, which combines the results of the above-mentioned methods and infers a probable localization site, is implemented as rules in the knowledge base. The basic strategy is given in Figure 4. This treelike structure is intended to simulate the actual sorting pathways in cells as much as possible. However, the machine reasoning process does not necessarily follow this decision tree downward. Rather, each localization site is represented as a "goal" element, and all sites are stacked in the working memory in inactive forms before the start of reasoning. During the reasoning steps, they are activated one by one. Thus each one is checked to see whether it has any characteristics from the stacked repertoire. The order in the stack is not important except for the cytoplasm, which is characterized as having no features of any other localization sites. The working memory element for the cytoplasm must be activated last, checking for the results of other hypotheses stored in the working memory.

In Table II, some key rules dealing with Gram-negative bacterial proteins are listed in a pseudo code. The body of a rule is divided into two parts: the LHS (lefthand side) and the RHS (righthand side). The LHS, which is a conditional expression, is composed of parenthesized patterns that will match a state of the working memory, whereas the RHS is a procedural part to be executed when activated. In our system, most rules have a pattern matching to some of the goal elements, realizing a "hypothesis-driven" inference (for example, see the first pattern of rule "iom1"). For the prediction of Gram-negative bacterial proteins, basic calculation results are repeatedly used for checking hypotheses. Thus, they are stored as "calc" elements in an early stage (e.g., rule "mcg1"). Our rules can be classified into two categories. One category calls a specific (C−) function under certain conditions, which actually deals with a given amino acid sequence and stores the result in the working memory (like rule "ompp2"). The other category combines and interprets the results to generate interim hypotheses or modify certainty factors for goal elements ("sig1," for example).

In addition to more or less general rules, it is also possible to incorporate rules dealing with exceptional cases. We introduced one such rule, "imb2," which was for saving a specific protein otherwise falsely predicted. In spite of the danger of overadjustment to training data, such rules may be useful for representing an expert's heuristics and biological reality. In actual cells, there seem to be "exceptional" proteins, which are not sorted in general pathways. This is particularly the case for secreted proteins, i.e., proteins excreted outside the cell, in Gram-negative bacteria as discussed later.

Prediction Accuracy

The degree of overall prediction obtained by our expert system is summarized in Figure 5. For the complete listing of our database and predicted localization sites, see Table I. For the four groups of proteins known to be at specific sites, the ratios of predicted localization sites are represented as stacked columns. In all, 88 out of 106 proteins (83%) were successfully predicted. Although proteins of other localization sites such as pili and outside of the cell have not been considered at the present stage, they represent rather minor components in the whole Gram-negative bacterial genome. Thus it might be said that with our system, localization sites of most Gram-negative bacterial proteins could be predicted with high accuracy.

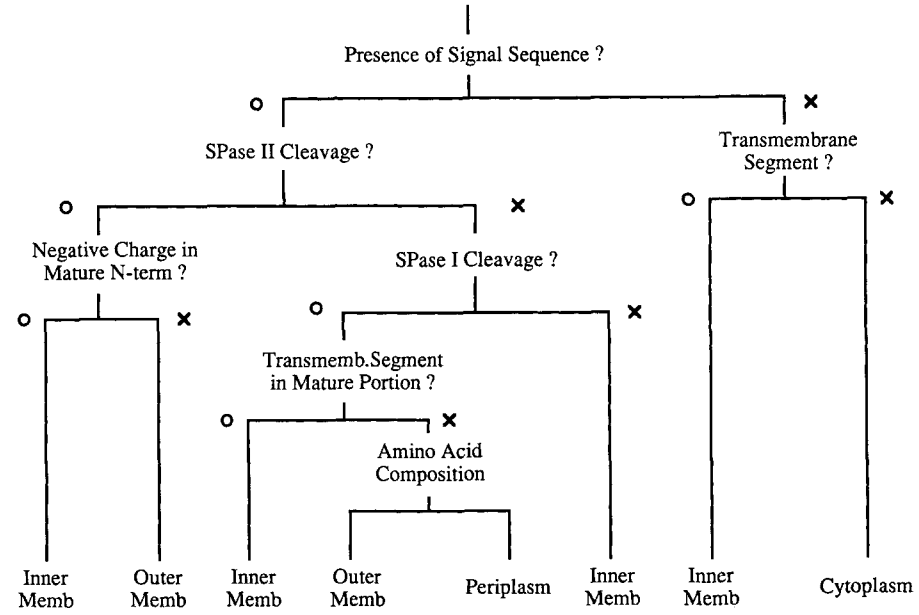

Fig. 4. Basic strategy for reasoning of protein localization sites. This is shown schematically in order to clarify the overall organization of rules. The actual path of reasoning does not always follow this tree exactly.

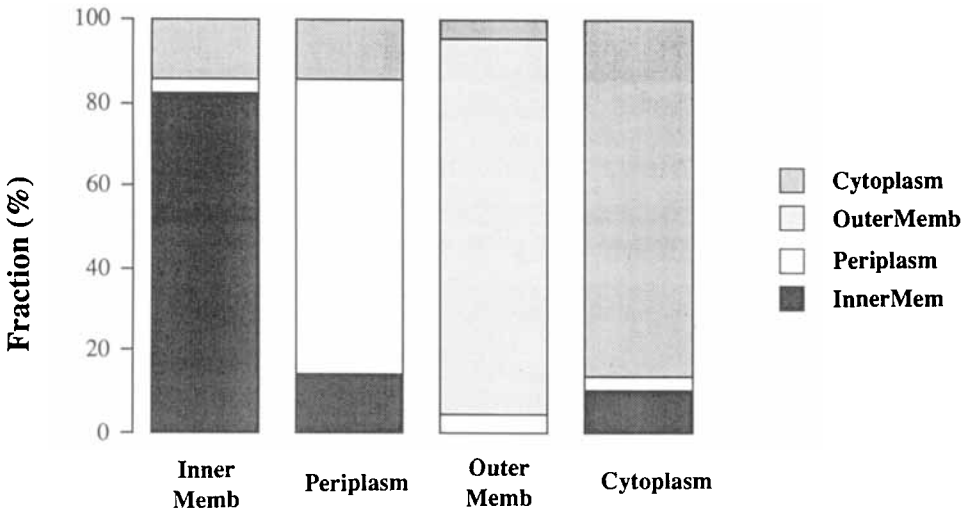

Fig. 5. Fraction of predicted localization sites in each category of proteins. The overall prediction accuracy was 83%.

As shown in Figure 5, most false cases of inner membrane proteins were predicted to be located at the cytoplasm. It turned out that they are peripheral membrane proteins belonging to the same superfamily (MalK). When peripheral membrane proteins were included in the cytoplasm class, the prediction accuracy rose to 87% (92 out of 106). Among the four groups of localization sites, outer membrane proteins and cytoplasmic proteins were predicted relatively well. One mistake (OmpA) in the outer membrane proteins was made from the discrimination by amino acid composition and another seems to be an exceptional protein (TraJ). In cytoplasmic proteins, most mistakes occurred at the detection of apparent hydrophobic segments. The group of periplasmic proteins was the most difficult to predict. Some of the falsely predicted ones did not have signal sequences and others had apparent membrane-spanning regions. In summary, some mistakes were made by simple failures of detecting needed features. Not a few mistakes seem to be due to exceptional cases of sorting by special pathways (see Discussion).

DISCUSSION

Comparison With Previous Works on Signal Sequences

Previous studies on signal sequences have shown that their basic characters are common both in prokaryotes and eukaryotes.[21, 22] Thus it may not be surprising that McGeoch's method, which was originally developed to recognize eukaryotic signal sequences, is also applicable to prokaryotic ones. We refined McGeoch's method by using discriminant analysis and by including the net charge of the CR as a third variable. The discriminant function enabled us to quantitatively evaluate the prediction result. The effect of the third variable in raising the prediction accuracy was not so large: only one more protein turned to be discriminated correctly. However, von Heijne observed that the net positive N-C charge imbalance was typically observed in prokaryotic cells,[23] which we consider is reflected in the third variable. The relative amounts of the contributions from the peak hydrophobicity, the length of the UR, and the net charge of the CR were 9:4:2 as estimated from our discriminant analysis.

Sjöström et al. reported a correlation of different amino acid patterns of signal sequences with different final localization sites using a technique called partial least squares discrimination.[9] However, the above mentioned universality of signal sequences, as well as the high prediction ability of our system, brings us to a hypothesis that the signal sequences play a role only in the translocation to the inner membrane. There is no evidence that they are used for later steps. Many signal peptide exchange experiments support this view.[24]

New Finding on Outer Membrane Proteins

In our attempt to construct an overall prediction scheme for sorting of prokaryotic proteins, it was most difficult to deal with outer membrane proteins. First, they do not have the long hydrophobic stretches characteristic of other membrane proteins. If there is no difference in the N-terminal signal sequences of periplasmic and outer membrane proteins, and if these signals are recognized and cleaved with the same machinery at the inner membrane, how are these two classes of proteins sorted? There might be further sorting signals in either class. Nikaido and Wu observed several regions of local homology in the major outer membrane proteins of *E. coli*.[25] However, there were no homologous regions common to all outer membrane proteins in our data set. Thus we had to give up incorporating knowledge about experimentally defined sorting signals and took a more practical approach.

From a structural point of view, outer membrane proteins seem to belong to a distinct class. Roughly speaking, they use β-sheet structures for membrane-spanning regions whereas other typical integral membrane proteins use α-helical structures, explaining the apparent lack of long hydrophobic segments. For selected proteins studied so far, experiments such as Raman spectroscopy and low resolution X-ray crystallography support this view.[26,27] Thus we tested if there were any differences in the contents of predicted secondary structures by Qian's and Sejnowski's method[28] between outer membrane proteins and periplasmic ones. No significant differences were observed (data not shown). Therefore, either the prediction accuracy is not sufficient or the secondary structure contents do not differ significantly.

However, our finding that the amino acid compositions of the two classes differ significantly still suggest an overall structural difference. A recent experiment suggests that outer membrane proteins are first secreted into the periplasm in a "water-soluble" form and then spontaneously inserted into the outer membrane possibly by affinity toward lipopolysaccharides.[29] It may be a global sequence feature rather than a local signal that is responsible for the driving force of insertion. In any event, our finding has some implications on the sorting mechanism to outer membranes.

The amino acids that were significantly different in their compositions between the two classes were identified by two methods (see above). The coefficients of the first principal component suggested that Tyr and Glu were favored in outer membranes, whereas Pro and Asn were favored in periplasmic proteins. Stepwise discriminant analysis suggested that Tyr and Ser were favored in outer membrane proteins, whereas Pro and Lys were favored in periplasmic proteins. Both methods suggest that Tyr is significant for the outer membrane class and Pro for the periplasm class. According to Chou and Fasman, Tyr is a weak former of β-sheet but a weak breaker of α-helix, whereas Pro is a strong breaker of α-helix and a weak breaker of β-sheet.[30] We checked against our amino acid index database[2] in order to see if there is any correlation with physicochemical and other properties of amino acids; no suitable indices were found. The amino acid preference is, at least, consistent with the observation that integral outer membrane proteins are β-rich. It is also consistent with a recent report that the double mutant ($Leu^{164} \rightarrow$ Pro and $Val^{166} \rightarrow$ Asp) of OmpA was not incorporated into the outer membrane.[31]

Some Comments on Falsely Predicted Proteins

In our prediction scheme, a nonlipoprotein inner membrane protein is assumed to have an uncleavable signal sequence in its N-terminus or hydrophobic transmembrane segments in the mature form. Obviously, this is not true for peripheral membrane proteins. It was not possible to discriminate them from cytoplasmic proteins. Perhaps they should be

classified into another class or should simply be included into the cytoplasm class.

Our assumption that all lipoproteins are integrated into either inner or outer membranes also turned out to have an exceptional case, pullulanase of *Klebsiella*. However, this protein is initially localized to the outer membrane and then secreted to the external medium; moreover, some specific factors are required for its localization.[32] By combining the results of McGeoch's method and von Heijne's method, we were able to discriminate all the lipoproteins in our database. One protein, outer membrane protein P6 precursor, which may actually be a lipoprotein,[33] was also discriminated. Another protein, penicillin-binding protein 3 precursor, which is positive with von Heijne's test, is predicted not to be a lipoprotein because it is negative with McGeoch's test. In fact, a small fraction (less than 15%) of it was observed to be modified with lipid.[34]

Proteins in the periplasm class and the outer membrane class, except lipoproteins, are postulated to have cleavable signal sequences at the N-terminus and have no hydrophobic segments in the mature part. However, some proteins apparently do not fulfil these requirements: BtuE protein of *E. coli*, hydrogenase large subunits of *Desulfovibrio baculkatus* and *D. gigas*, which belong to different families, do not seem to have normal signal sequences.[35] Prickril et al. reported that 34 N-terminal residues were cleaved in the small subunit of hydrogenase from *D. vulgaris*. Because the mature large subunit seemed to be intact from cleavage, they suggested the presence of an internal signal peptide among possible sorting mechanisms.[36]

Another exceptional case also exists in the outer membrane class: an N-terminal signal sequence of TraJ protein was not detected in our analysis, in agreement with the original report.[37] Because of its unique role played in conjugal transfer of DNA, it is possible that its mechanism of membrane binding is different from others. According to our discrimination rule, OmpA, a prominent outer membrane protein, does not show the amino acid composition characteristic of outer membrane proteins. Vogel and Jähnig reported that the C-terminal half of the molecule was exposed to the periplasmic space.[24] In fact, there was a difference in the amino acid composition between the N-terminal and the C-terminal halves (data not shown). Most of the other falsely predicted proteins had unexpected hydrophobic stretches. That seems, however, to be due to the imperfectness of Klein et al.'s method.

Sites Not Included in Our Repertoire

In the present study, not all possible localization sites were considered. The main groups not included in the prediction repertoire are proteins at pili, flagella, and the external medium, as many of them are considered to have specific sorting pathways. For example, flagellin of *E. coli* is exported by a flagellum-specific pathway, in which its uncleaved N-terminal portion seems to be essential.[38] The molecular mechanism of pilus-adhesion biogenesis is not well understood but it requires specific factors.[39] Furthermore, the extracellular proteins, studied in any detail so far, appear to be transported across bacterial envelopes by individual mechanisms.[40]

It is interesting, however, to see where these exceptional proteins are predicted to be located among the four categories in our system. For example, this may enable us to deduce their localization sites when their specific export machinery is missing or defective. In Table III, secreted proteins and pili proteins are listed with their predicted location sites. Of the secreted proteins, half are predicted to be at the inner membrane, including a major group of colicins. It is likely that this result implies their target site after secretion by a specific pathway and reflects their poreforming activity. Colicin M, which is predicted to be located at the cytoplasm, seems to work by a mechanism different from poreformation.[41] Except these colicins and hemolysin A of *E. coli*, other secreted proteins appear to have normal N-terminal signal peptides. Thus they are likely to be transported across the inner membrane as a first step. Most of the pilus proteins also seem to have normal N-terminal signal peptides.

In Table III, another class of proteins is listed: M13 coat proteins. One of them is known as an exceptional protein which does not need the *secA* or *secY* components of the translocation machinery through the inner membrane, although it has a cleavable N-terminal signal sequence.[42] Our method predicted, as expected, that both were inner membrane proteins with cleavable signal peptides and trans membrane segments. Thus the results of McGeoch's and von Heijne's methods do not seem to correlate with dependence on *sec* function for export.

Expert Systems for Interpreting Genetic Information

The main merit of adopting an expert system approach in extracting the biological meaning of sequence data lies in its flexibility. An expert system enables us to accumulate different types of knowledge, whether experimentally defined or computationally derived. The flexibility of inserting/deleting rules is also critical to future improvement by incorporating new insights. Although the present study has focused only on Gram-negative bacteria, our system can be easily expanded to be able to deal with sequences of other organisms. Most of the rules described in this work can also be used for eukaryotic proteins with minor modifications (Nakai and Kanehisa, in preparation). In addition, by regarding the external medium site of Gram-positive bacteria as the periplasm of Gram-negative ones, many of the rules described here can also be applied to predicting

TABLE III. Proteins Located at Other Sites

(a) Secreted Proteins Predicted Site	Entry	McG	Definition
Cytoplasm	IKECM	−−	Colicin M—*E. coli*
	A26569	−−	Exotoxin A regulatory protein—*Pseudomonas aeruginosa*
Inner Membrane	IKEC1	−	Colicin E1 protein—*E. coli* plasmid colicin E1
	IKECB	−−−−	Colicin Ib protein—*E. coli* plasmid ColIb
	IKECBB	−−−	Colicin B—*E. coli* plasmid ColBM-pF166
	S00867	−−−	Colicin N—*E. coli* plasmid pCHAP4
	LEECA	−−−−	Hemolysin A—*E. coli*
	S01892	+	Soluble hemolysin precursor—*Vibrio cholerae*
	A26879	+	Pullulanase precursor—*Klebsiella pneumoniae*
	QLECA	++	Heat-labile enterotoxin A chain precursor—*E. coli*
	B29831	++	Heat-labile enterotoxin IIa, B-chain precursor—*E. coli*
Outer Membrane	A29831	+++	Heat-labile enterotoxin IIa, A-chain precursor—*E. coli*
	A26039	+	IgA protease precursor—*Neisseria gonorrhoeae*
	A29840	+	Extracellular serine protease precursor—*Serratia marcescens*
Periplasm	QLECB	+	Heat-labile enterotoxin B-chain precursor—*E. coli*
	QHEC1	++	Heat-stable enterotoxin ST-I—*E. coli*
	A25158	+	Pectate lyase E—*Erwinia chrysanthemi*
	B25158	+	Pectate lyase B—*Erwinia chrysanthemi*
	A25976	+	Aerolysin precursor—*Aeromonas hydrophila*
(b) Pili Predicted Site	Entry	McG	Definition
Inner Membrane	YQECFX	−	Fimbrial protein precursor—*E. coli* plasmid F
Outer Membrane	YQECPH	+	*papH* fimbrial protein precursor—*E. coli*
	YQECT1	++	Type 1 fimbrial protein precursor—*E. coli*
Periplasm	YQECPE	+	*papE* fimbrial protein—*E. coli*
	YQECPF	+	*papF* fimbrial protein—*E. coli*
	YQECPP	++	*Pap* fimbrial protein precursor—*E. coli*
(c) M13 Coat Proteins Predicted Site	Entry	McG	Definition
Inner Membrane	VCBPFD	+	Coat protein B precursor—Bacteriophages fd, M13
	Z3BPFD	++	Coat protein A precursor—Bacteriophages fd, M13

the localization sites of Gram-positive bacteria. In fact, some gene-fusion experiments show that Gram-positive bacterial signal sequences are recognized by Gram-negative ones as export signals.[22] Thus the present system provides a basic skeleton for a system capable of making predictions for many different organisms.

One weak point of the expert system approach is the difficulty in objectively assessing prediction results. In our approach, the distinction of training set and test set is difficult. We used the parameters obtained by the original authors and data sets that sometimes contained the same data as ours. We obtained rules from the analysis of limited portions of our data set, such as the rule for discriminating periplasmic and outer membrane proteins. In spite of this difficulty, we expect our method will be useful in prediction because most rules are based on sequence characteristics that have been proved to be important experimentally. As a last note, the system will hopefully be useful in characterizing open reading frames when the entire genome of *E. coli* is sequenced.

ACKNOWLEDGMENTS

We greatly thank Dr. Koreaki Ito and Chiharu Ueguchi for giving us useful information and reading the manuscript; Dr. Katsumi Nitta and Paul Horton for critical reading of the manuscript; Dr. Gunnar von Heijne for the discussion on the OmpA protein. This work was partly supported by the Special Coordination Funds for Promoting Science and Technology from the Science and Technology Agency of Japan.

REFERENCES

1. Kanehisa, M. A multivariate analysis method for discriminating protein secondary structural segments. Protein Eng. 2:87–92, 1988.
2. Nakai, K., Kidera, A., Kanehisa, M. Cluster analysis of amino acid indices for prediction of protein structure and function. Protein Eng. 2:93–100, 1988.
3. Nakai, K., Kanehisa, M. Prediction of in-vivo modification sites of proteins from their primary structures. J. Biochem. (Tokyo) 104:693–699, 1988.
4. Waterman, D. A. A Guide to Expert Systems. Addison-Wesley, 1986.
5. Blundell, B. L., Sibanda, B. L., Sternberg, M. J. E., Thornton, J. M. Knowledge-based prediction of protein structures and the design of novel molecules. Nature (London) 326:347–352, 1987.

6. Model, P., Russel, M. Prokaryotic Secretion. Cell 61:739–741, 1990.
7. Barker, W.C., George D. G., Hunt, L. T. Protein sequence database. Methods Enzymol. 183:31–49, 1990.
8. Watson, M. E. E. Compilation of published signal sequences. Nucl. Acids Res. 12:5145–5164, 1984.
9. Sjöström, M., Wold, S., Wieslander, Å., Rilfors, L. Signal peptide amino acid sequences in Escherichia coli contain information related to final protein localization. A multivariate data analysis. EMBO J. 6:823–831, 1987.
10. Forgy, C. L. The OPS83 User's Manual System Version 2.2, Production System Technologies, 1986.
11. Tanaka, H., Shimoi, Y. Expert System Kochiku-no-hôhô (in Japanese), Personal Media, 1987.
12. McGeoch, D. J. On the predictive recognition of signal peptide sequences. Virus Research 3:271–286, 1985.
13. von Heijne, G. The structure of signal peptides from bacterial lipoproteins. Protein Eng. 2:531–534, 1989.
14. von Heijne, G. Patterns of amino acids near signal-sequence cleavage sites. Eur. J. Biochem. 133:17–21, 1983.
15. von Heijne, G. A new method for predicting signal sequence cleavage sites. Nucl. Acids Res. 14:4683–4690, 1986.
16. Eisenberg, D. Three-dimensional structure of membrane and surface proteins. Ann. Rev. Biochem. 53:595–623, 1984.
17. Fasman, G. D., Gilbert, W. A. The prediction of transmembrane protein sequences and their conformation: an evaluation. Trends Biochem. Sci. 15:89–92, 1990.
18. Klein, P., Kanehisa, M., DeLisi, C. The detection and classification of membrane-spanning proteins. Biochim. Biophys. Acta 815:468–476, 1985.
19. Yamaguchi, K., Yu, F., Inoue, M. A single amino acid determinant of the membrane localization of lipoproteins in *E. coli*. Cell 53:423–432, 1988.
20. Baker, K., Mackman, N., Holland, B. Genetics and biochemistry of proteins into the outer membrane of *E. coli*. Prog. Biophys. Molec. Biol. 49:89–115, 1987.
21. von Heijne, G. Signal sequences. The limits of variation. J. Mol. Biol. 184:99–105, 1985.
22. Briggs, M. S., Gierasch, L. M. Molecular mechanisms of protein secretion: The role of the signal sequence. Adv. Protein Chem. 38:109–180, 1986.
23. von Heijne, G. Net N-C charge imbalance may be important for signal sequence function in bacteria J. Mol. Biol. 192:287–290, 1986.
24. Beckwith, J., Ferro-Novick, S. Genetic studies on protein export in bacteria. Curr. Topics Microbiol. Immunol. 125: 5–27, 1986.
25. Nikaido, H., Wu, H. C. P. Amino acid sequence homology among the major outer membrane proteins of *Escherichia coli*. Proc. Natl. Acad. Sci. USA 81:1048–1052, 1984.
26. Vogel, H., Jähnig Models for the structure of outer-membrane proteins of *Escherichia coli* derived from Raman spectroscopy and prediction methods. J. Mol. Biol. 190: 191–199, 1986.
27. Weiss, M. S., Wacker, T., Woitzik, N. D., Weckesser, J., Kreutz, W., Welte, W., Schulz, G.E. The structure of porin from *Rhodobacter capsulatus* at 0.6 nm resolution. FEBS Lett. 256:143–146, 1989.
28. Qian, N., Sejnowski, T. J. Predicting the secondary structure of globular proteins using neural network models. J. Mol. Biol. 202:865–884, 1988.
29. Sen, K., Nikaido, H. *In vitro* trimerization of OmpF porin secreted by spheroplasts of *Excherichia coli*. Proc. Natl. Acad. Sci. USA 87:743–747, 1990.
30. Chou, P. Y., Fasman, G. D. Prediction of the secondary structure of proteins from their amino acid sequence. Adv. Enzym. 47:45–148, 1978.
31. Klose, M., Jähnig, F., Hindennach, I., Henning, U. Restoration of membrane incorporation of a *Escherichia coli* outer membrane protein (OmpA) defective in membrane insertion. J. Biol. Chem. 264:21842–21847, 1989.
32. d'Enfert, C., Ryter, A., Pugsley, A. P. Cloning and expression in *Escherichia coli* of the *Klebsiella pneumoniae* genes for production, surface localization and secretion of the lipoprotein pullulanase. EMBO J. 6:3531–3538, 1987.
33. Nelson, M. B., Apicella, M. A., Murphy, T. F., Vankeulen, H., Spotila, L. D., Rekosh, D. Cloning and sequencing of *Haemophilus influenzae* outer membrane protein P6. Infect. Immun. 56:128–134, 1988.
34. Hayashi, S., Hara, H., Suzuki, H., Hirota, Y. Lipid modification of *Escherichia coli* penicillin-binding protein 3. J. Bacteriol. 170:5392–5395, 1988.
35. Friedrich, M. J., DeVeaux, L. C., Kadner, R. J. Nucleotide sequence of the *btuCED* genes involved in vitamin B_{12} transport in *Escherichia coli* and homology with components of periplasmic-binding-protein-dependent transport systems. J. Bacteriol. 167:928–934, 1986.
36. Prickril, B. C., Czechowski, M. H. Przybyla, A. E., Peck, Jr., H. D., LeGall, J. Putative signal peptide on the small subunit of the periplasmic hydrogenase from *Desulfovibrio vulgaris*. J. Bacteriol. 167:722–725, 1986.
37. Fowler, T., Taylor, L., Thompson, R. The control region of the F plasmid transfer operon: DNA sequence of the *traJ* and *traY* genes and characterization of the *traY* → Z promoter. Gene. 26:79–89, 1983.
38. Kuwajima, G., Kawagishi, I., Homma, M., Asaka, J., Kondo, E., Macnab, R. M. Export of an N-terminal fragment of *Escherichia coli* flagellin by a flagellum-specific pathway. Proc. Natl. Acad. Sci. USA 86:4953–4957, 1989.
39. Uhlin, B. E., Båga, M., Göransson, M., Lindberg, F. P., Lund, B., Norgren, M., Normark, S. Genes determining adhesion formation in uropathogenic *Escherichia coli*. Curr. Topics Microbiol. Immunol. 118:163–178, 1985.
40. Hirst, T. R., Welch, R. A. Mechanisms for secretion of extracellular proteins by Gram-negative bacteria. Trends Biochem. Sci. 13:265–269, 1988.
41. Köck, J., Ölschläger, T., Kamp, R. M., Braun, V. Primary structure of colicin M, an inhibitor of murein biosynthesis. J. Biol. Chem. 169:3358–3361, 1987.
42. Wolfe, P. B., Rice, M., Wickner, W. Effects of two *sec* genes on protein assembly into the plasma membrane of *Escherichia coli*. J. Biol. Chem. 260:1836–1841, 1985.