PROTEINS: Structure, Function, and Genetics 34:137–153 (1999)

Prediction of Membrane Protein Types and Subcellular Locations

Kuo-Chen Chou* and David W. Elrod
Computer-Aided Drug Discovery, Pharmacia and Upjohn, Kalamazoo, Michigan

ABSTRACT **Membrane proteins are classified according to two different schemes. In scheme 1, they are discriminated among the following five types: (1) type I single-pass transmembrane, (2) type II single-pass transmembrane, (3) multipass transmembrane, (4) lipid chain-anchored membrane, and (5) GPI-anchored membrane proteins. In scheme 2, they are discriminated among the following nine locations: (1) chloroplast, (2) endoplasmic reticulum, (3) Golgi apparatus, (4) lysosome, (5) mitochondria, (6) nucleus, (7) peroxisome, (8) plasma, and (9) vacuole. An algorithm is formulated for predicting the type or location of a given membrane protein based on its amino acid composition. The overall rates of correct prediction thus obtained by both self-consistency and jackknife tests, as well as by an independent dataset test, were around 76–81% for the classification of five types, and 66–70% for the classification of nine cellular locations. Furthermore, classification and prediction were also conducted between inner and outer membrane proteins; the corresponding rates thus obtained were 88–91%. These results imply that the types of membrane proteins, as well as their cellular locations and other attributes, are closely correlated with their amino acid composition. It is anticipated that the classification schemes and prediction algorithm can expedite the functionality determination of new proteins. The concept and method can be also useful in the prioritization of genes and proteins identified by genomics efforts as potential molecular targets for drug design. Proteins 1999;34:137–153.**

Key words: organelles; transmembrane; anchored membrane; amino acid composition; component-coupled effect; bioinformatics

INTRODUCTION

Cell membranes are crucial to the life of a cell. A cell is enclosed by the plasma membrane (cell envelope), which defines its boundaries, and maintains the essential differences between the cytosol and the extracellular environment. Inside the cell there are various organelles such as the endoplasmic reticulum, Golgi apparatus, mitochondria, and other membrane-bound organelles. The characteristic differences between the contents of the cytosol and each of these organelles are maintained by their respective membranes (subcell envelopes). Although the basic structure of biological membranes is provided by the lipid bilayer, most of the specific functions are carried out by the membrane proteins.

Membrane proteins consist of transmembrane proteins and anchored membrane proteins. The former contains one or more hydrophobic segments, and hence is relatively easily discriminated from nonmembrane proteins.[1] The latter has a consensus sequence motif at either the N- or C-terminus,[2,3] and hence can be recognized to some extent. For example, anchored membrane proteins are usually either isoprenylated at the C-terminus with the consensus sequence motif of CAAX, or myristylated at the N-terminus with the motif of GXXXS/T, or palmitylated at a specific Cys-residue of the N-terminal region. Another type of anchored membrane protein is of GPI-anchored proteins which are modified through glycosylphosphatidylinositol (GPI) at the C-terminus with a unique sequence feature, such as a hydrophobic tail.

The way that a membrane-bound protein is associated with the lipid bilayer usually reflects the function of the protein. For example, only transmembrane proteins can function on both sides of the bilayer or transport molecules across it. By contrast, proteins that function on only one side of the lipid bilayer are often associated exclusively with either the lipid monolayer or a protein domain on that side.

Also, associated with different locations, membrane proteins usually have different biological functions. Proteins associated with the cell plasma membrane act as sensors of external signals, transferring information across the membrane and allowing the cell to change its behavior in response to environmental cues. The ion gradients across membranes, which can be used to synthesize ATP, to drive the transmembrane movement of selected solutes, or to produce and transmit electrical signals in nerve and muscle cells, are established by the activities of specialized membrane proteins.

Therefore, the determination of function for new membrane proteins can be expedited significantly if we can find an effective scheme and algorithm to predict their types and subcellular locations. Especially nowadays, the number of protein sequences entering into public data banks is rapidly increasing; it would be both time-consuming and costly to rely on completely experiments for the solution of

*Correspondence to: Kuo-Chen Chou, Computer-Aided Drug Discovery, Pharmacia and Upjohn, Kalamazoo, MI 49007-4940.
Received 2 June 1998; Accepted 4 September 1998

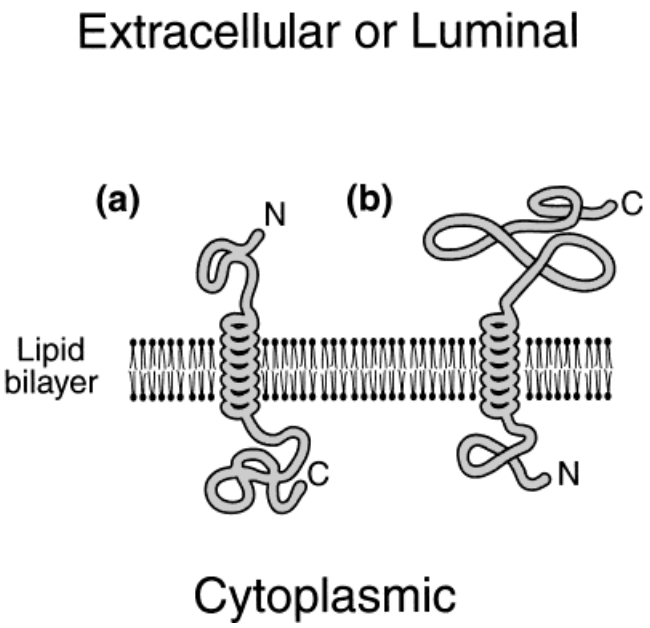

Fig. 1. Schematic drawing showing **(a)** type I transmembrane protein, and **(b)** type II transmembrane protein. Type I and II membrane proteins are of single-pass transmembrane. However, type I has a cytoplasmic C-terminus and an extracellular or luminal N-terminus for plasma membrane or organelle membrane, respectively, while the arrangement of N- and C-termini in type II membrane proteins is just the reverse.

these problems. Furthermore, the establishment of such an algorithm can also help prioritize genes and proteins to be identified by genomics efforts as potential molecular targets for drug design. The issue is, however, given the sequence of a membrane protein, can we predict its inherent attributes in a cell? In other words, is it associated with the cell membrane (i.e., plasma membrane) or with the membrane of a specific organelle inside the cell? How is it embedded in, or bound to, a membrane? Is it an inner or outer membrane protein? The present study was devoted to these problems.

CLASSIFICATION SCHEMES

Membrane proteins, also called membrane-bound proteins or membrane-associated proteins, were classified according to two different schemes: one based on their interaction modes with membranes, and the other on their cellular locations.

Types of Membrane Proteins

In the literature, the definitions for the category of membrane proteins and their types are not unique. In this article, the membrane proteins are categorized into six types.

1. *Type I membrane protein:* This single-pass transmembrane protein has an extracellular (or luminal) N-terminus and cytoplasmic C-terminus for a cell (or organelle) membrane (Fig. 1a).
2. *Type II membrane protein:* This single-pass transmembrane protein has an extracellular (or luminal) C-terminus and cytoplasmic N-terminus for a cell (or organelle) membrane (Fig. 1b).
3. *Multipass transmembrane proteins:* In type I and II membrane proteins, the polypeptide crosses the lipid bilayer only once (Fig. 1), whereas in multipass membrane proteins, the polypeptide crosses the lipid bilayer multiple times (Fig. 2a). Most of the membrane-spanning segments of polypeptide chains are thought to have an α-helical conformation. This is because in a lipid environment the hydrogen bonding between peptide bonds would be maximized if the polypeptide chain were to form a regular α-helix.
4. *Lipid chain-anchored membrane proteins:* See Figure 2b and further explanation below.
5. *GPI-anchored membrane proteins:* See Figure 2c and further explanation below.

   Both lipid chain- and GPI-anchored membrane protein are also called membrane-anchored proteins. However, the former is associated with the bilayer only by means of one or more covalently attached fatty acid chains or other types of lipid chains called prenyl groups, whereas the latter is bound to the membrane by a glycosylphosphatidylinositol (GPI) anchor.
6. *Peripheral membrane proteins:* (See Figure 2d.) Proteins of this type are actually bound to the membrane indirectly by noncovalent interactions with other membrane proteins.

The peripheral membrane proteins can be released from the membrane by relatively gentle extraction procedures without affecting the intactness of the lipid bilayer. By contrast, membrane proteins of the other five types cannot be released using these procedures and are therefore called integral membrane proteins. For clarity, Figure 3 presents a categorization chart for the six types of membrane proteins. In this study, however, peripheral membrane proteins were left out for further consideration because, unlike integral membrane proteins, they do not have a unique sequence feature that can be used to discriminate them from non-membrane proteins. Also, so far the available peripheral membrane protein sequences are too few to be statistically significant.

The classification was based on release 35.0 of SWISS-PROT.[4] In order to obtain a high-quality, well-defined training set, the data were screened strictly according to the following three procedures. The first procedure included only those sequences with clear descriptions, those without labels to indicate explicitly one of the above five types (i.e., type I, type II, multipass, lipid-chain anchored, and GPI-anchored), or those with ambiguous annotations, such as "probable," "potential," and "by similarity," were totally left out. Note that in the SWISS-PROT data bank, the annotation "integral membrane protein" represents only multipass transmembrane proteins. In the second procedure, for protein sequences having the same name,

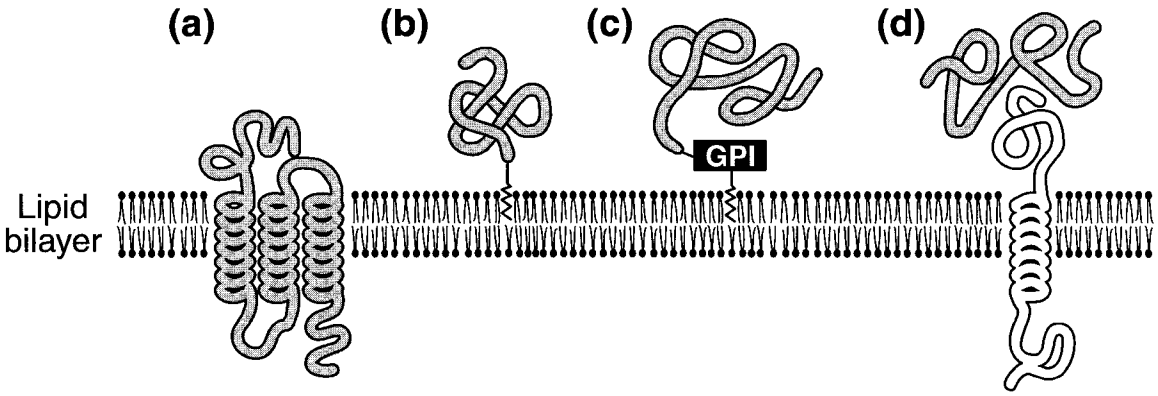

Fig. 2. Schematic drawing showing **(a)** multipass transmembrane, **(b)** lipid chain-anchored membrane, **(c)** GPI-anchored membrane, and **(d)** peripheral membrane proteins.

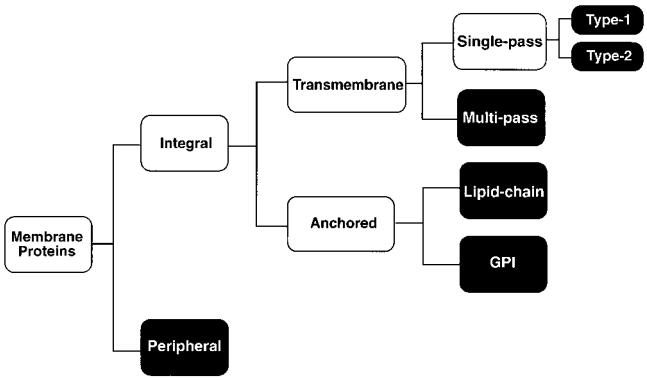

Fig. 3. Categorization chart to show the relationship of the six membrane protein types illustrated in Figs. 1 and 2.

but from different species, only one was included. Third, sequences whose type is described by two or more types were not included because of lack of uniqueness. After the above screening procedures, we obtained a dataset of 2,059 protein sequences, of which 435 are type I transmembrane proteins, 152 type II transmembrane proteins, 1,311 multipass transmembrane proteins, 51 lipid chain-anchored membrane proteins, and 110 GPI-anchored membrane proteins. Table I presents the 2,059 membrane proteins. This dataset was used as a training dataset for predicting the membrane protein types.

Locations of Membrane Proteins

On the basis of their cellular locations, membrane proteins may be classified according to nine discriminative categories: (1) chloroplast, (2) endoplasmic reticulum, (3) Golgi apparatus, (4) lysosome, (5) mitochondria, (6) nucleus, (7) peroxisome, (8) plasma membrane, and (9) vacuole (Fig. 4). Such a classification covers almost all the organelles in an animal or plant cell that have a lipid bilayer for their membrane structure.[5,6] Note that the vacuole and chloroplast exist only in a plant cell. Some transmembrane proteins are also accompanied with the label "inner membrane," "outer membrane," or "thylakoid membrane." This is because some organelles (e.g., mitochondria) contain both inner and outer membrane structures (Fig. 5a), and some organelle (chloroplast) contains a thylakoid membrane structure as well as inner and outer membrane structures (Fig. 5b).

The classification data were based on release 35.0 of SWISS-PROT.[4] In order to get a high-quality well-defined training set, the data were screened strictly according to the following procedures. In the first procedure, only those sequences with clear locational descriptions were included; those with ambiguous or uncertain words such as "probable," "potential," and "by similarity" were totally left out. In the second procedure, sequences labeled with "peripheral membrane protein" were excluded. In the third procedure, for protein sequences with the same name, but from different species, only one was included. In the fourth procedure, sequences whose location is described by two or more organelles were not included because of lack of uniqueness. After these four screening procedures, we obtained a dataset of 2,105 protein sequences, of which 55 are chloroplast membrane proteins, 64 endoplasmic reticulum membrane proteins, 44 Golgi membrane proteins, 21 lysosome membrane proteins, 154 mitochondria membrane proteins, 26 nucleus membrane proteins, 37 peroxisome membrane proteins, 1,680 plasma membrane proteins, and 24 vacuole membrane proteins. For the subsets of lysosome, nucleus, and peroxisome, the constraint in the third procedure was relaxed; otherwise, the numbers of proteins in these subsets would be too few to be statistically significant. This might lead to a somewhat optimistic error estimate for the rates of correct prediction for these classes. However, the impact on the overall rate of correct prediction is trivial because of the small size of these subsets themselves. Furthermore, as more and more protein sequences belonging to these subsets are accumulated in future releases of SWISS-PROT, such a problem will automatically vanish. The names of the 2,105 membrane proteins are given in Table II. This dataset was used as a training dataset for predicting the cellular locations of membrane proteins.

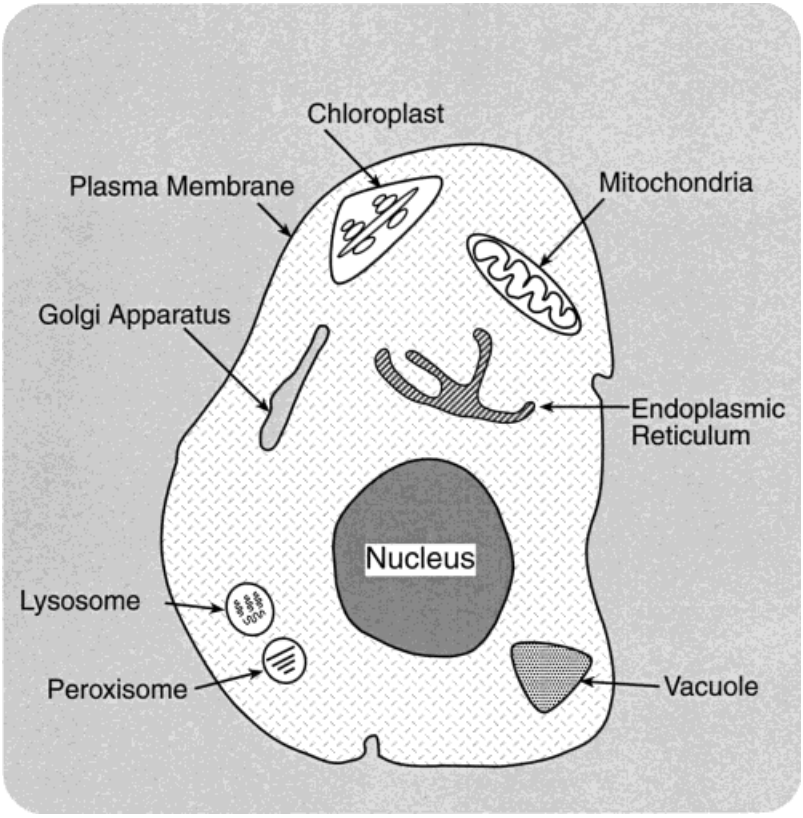

Fig. 4. Schematic drawing showing different cellular locations of membrane proteins: (1) chloroplast, (2) endoplasmic reticulum, (3) Golgi apparatus, (4) lysosome, (5) mitochondria, (6) nucleus, (7) peroxisome, (8) plasma, and (9) vacuole. Shown are only those organelles whose lipid bilayer envelopes provide a matrix for membrane proteins. Note that the vacuole and chloroplast exist only in a plant cell.

PREDICTION ALGORITHM

For brevity, let us use numerical indexes to represent the respective categories. For the case of membrane protein types, we use 1, 2, 3, 4, and 5 to represent the type I transmembrane, type II transmembrane, multipass transmembrane, lipid chain-anchored membrane, and GPI-anchored membrane proteins, respectively. Thus, S_1 represents the subset consisting of only type I transmembrane proteins, S_2 represents the subset consisting of only type II transmembrane proteins, and so forth. For the case of membrane protein locations, we use 1, 2, 3, 4, 5, 6, 7, 8, and 9 to represent the chloroplast, endoplasmic, Golgi, lysosome, mitochondria, nucleus, peroxisome, plasma, and vacuole membrane proteins, respectively. Thus, S_1 represents the chloroplast subset consisting of only chloroplast membrane proteins, S_2 represents the endoplasmic subset consisting of only endoplasmic membrane proteins, and so forth.

Suppose there are N proteins forming a set S, which is the union of m subsets; i.e.,

$$S = S_1 \cup S_2 \cup S_3 \cup S_4 \cup \cdots \cup S_m \qquad (1)$$

The size of each subset is given by N_ξ ($\xi = 1, 2, 3, \ldots, m$), where N_ξ represents the number of proteins in the subset S_ξ. Obviously, $N = \Sigma_{\xi=1}^{m} N_\xi$. For example, for the dataset of Table I, we have $m = 5$, $N_1 = 435$, $N_2 = 152$, $N_3 = 1{,}311$, $N_4 = 51$, $N_5 = 110$, and $N = 2{,}059$. For the dataset of Table II, we have $m = 9$, $N_1 = 55$, $N_2 = 64, \ldots,$ $N_8 = 1{,}680$, $N_9 = 24$, and $N = 2{,}105$.

The prediction algorithm is established based on the correlation between the cellular location of a membrane protein and its amino acid composition. It has been demonstrated in the studies of protein structural class prediction[7,8] that the incorporation of coupling effects among different amino-acid-components through the covariance matrices can significantly improve the prediction quality. However, the covariant discriminant algorithm formulated in those studies is valid only when the subset sizes in a training dataset are the same, or approximately the same. However, for the current case, the subset sizes are very different. For example, the subset size for the multipass transmembrane proteins is much bigger than that of the lipid chain-anchored membrane proteins; the subset size for the plasma membrane is much bigger than those of the other eight organelles. Therefore, it is necessary to introduce a more general covariant discriminant algorithm that will not be subject to such a limitation of "same subset size." This algorithm is formulated as follows.

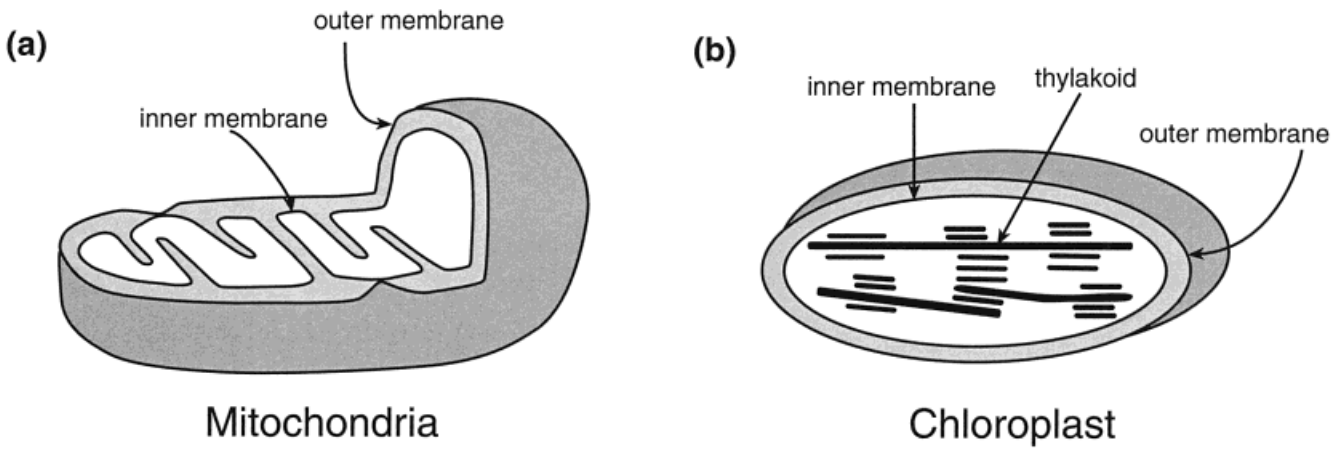

Fig. 5. Schematic drawing showing **(a)** mitochondria and **(b)** chloroplast. The former has inner and outer membrane structures, while the latter has thylakoid as well as inner and outer membrane structures. Adapted from Alberts et al.[5]

Suppose any protein in the set S corresponds to a vector or a point in the 20-dimensional (20-D) space; i.e., it can be described by[8]

$$\mathbf{X}_k^\xi = \begin{bmatrix} x_{k,1}^\xi \\ x_{k,2}^\xi \\ \vdots \\ x_{k,20}^\xi \end{bmatrix}, \quad (k = 1, 2, \ldots, N_\xi; \quad \xi = 1, 2, 3, \ldots, m) \quad (2)$$

where $x_{k,1}^\xi$, $x_{k,2}^\xi$, . . . , $x_{k,20}^\xi$ are the normalized occurrence frequencies of the 20 amino acids in the kth protein $\mathbf{X}_k^\xi$ of the subset S^ξ. The *standard vector* for the subset S^ξ is defined by

$$\mathbf{X}^\xi = \begin{bmatrix} x_1^\xi \\ x_2^\xi \\ \vdots \\ x_{20}^\xi \end{bmatrix}, \quad (\xi = 1, 2, 3, \ldots, m) \quad (3)$$

where

$$x_i^\xi = \frac{1}{N_\xi} \sum_{k=1}^{N_\xi} x_{k,i}, \quad (i = 1, 2, \ldots, 20) \quad (4)$$

Suppose $\mathbf{X}$ is a membrane protein whose type or cellular location is to be predicted. It can be either one of the N proteins in the set S, or a protein outside of it. It also corresponds to a point $(x_1, x_2, \ldots, x_{20})$ in the 20-D space, where x_i has the same meaning as $x_{k,i}^\xi$ but is associated with protein $\mathbf{X}$ instead of $\mathbf{X}_k^\xi$. Thus, the current algorithm can be formulated as follows.

The similarity between the standard vector $\mathbf{X}^\xi$ and the protein $\mathbf{X}$ can be characterized by the covariant discriminant function, as defined by Johnson and Wichem.[9]

$$F(\mathbf{X}, \mathbf{X}^\xi) = \Lambda \ln (2\pi) - 2 \ln \Psi_\xi + (\mathbf{X} - \mathbf{X}^\xi)^{\mathrm{T}} \mathrm{C}_\xi^{-1} (\mathbf{X} - \mathbf{X}^\xi) + \ln (\lambda_1^\xi \lambda_2^\xi \lambda_3^\xi \ldots \lambda_{19}^\xi) \quad (5)$$

where Λ is the dimension of the amino acid composition space, and hence the first term can be ignored because it is a constant. Ψ_ξ is the prior probability of the subset S^ξ. Because the prior probabilities Ψ_ξ ($\xi = 1, 2, 3, \ldots, m$) are unknown, a common practice is to assume that they are equal. Thus, the second term in equation (5) can also be ignored. Accordingly, equation (5) can be reduced to

$$F(\mathbf{X}, \mathbf{X}^\xi) = (\mathbf{X} - \mathbf{X}^\xi)^{\mathrm{T}} \mathbf{C}_\xi^{-1} (\mathbf{X} - \mathbf{X}^\xi) + \ln (\Pi_{i=1}^{19} \lambda_i^\xi) \quad (6)$$

where the superscript T is the transposition operator; $\mathbf{C}_\xi$ is the covariance matrix for subset S^ξ defined by

$$\mathbf{C}_\xi = \begin{bmatrix} c_{1,1}^\xi & c_{1,2}^\xi & \cdots & c_{1,20}^\xi \\ c_{2,1}^\xi & c_{2,2}^\xi & \cdots & c_{2,20}^\xi \\ \vdots & \vdots & \ddots & \vdots \\ c_{20,1}^\xi & c_{20,2}^\xi & \cdots & c_{20,20}^\xi \end{bmatrix} \quad (7)$$

with the matrix elements $c_{i,j}^\xi$ given by

$$c_{i,j}^\xi = \frac{1}{N_\xi - 1} \sum_{k=1}^{N_\xi} [x_{k,i}^\xi - x_i^\xi][x_{k,j}^\xi - x_j^\xi], \quad (i, j = 1, 2, \ldots, 20) \quad (8)$$

TABLE I. List of 2,059 Protein Sequences Used as Training Data for Predicting Membrane Protein Types[†]

(1) 435 type1 I transmembrane proteins

41BB_HUMAN	A33_HUMAN	A4_DROME	ACET_HUMAN	ACE_HUMAN	AMA1_PLACH	AMFR_HUMAN	ANPA_HUMAN	ANPB_ANGJA	ANPC_BOVIN
APP1_HUMAN	APP2_RAT	APX1_CAEEL	ARP4_STRPY	ASA1_ENTFA	AVR2_BOVIN	AVRB_HUMAN	BAG_STRAG	BASI_CHICK	BCA_STRAG
BFR2_HUMAN	BGP1_HUMAN	BLVR_BOVIN	BTC_HUMAN	BUTY_BOVIN	C114_MOUSE	C166_BRARE	C22A_HUMAN	C22B_HUMAN	C79A_BOVIN
C79B_HUMAN	C8B1_HUMAN	C8B2_HUMAN	CAD1_CHICK	CAD2_CHICK	CAD3_HUMAN	CAD4_CHICK	CAD5_MOUSE	CAD6_HUMAN	CAD8_HUMAN
CADB_HUMAN	CADC_HUMAN	CADF_HUMAN	CADL_RAT	CADN_XENLA	CADO_XENLA	CALG_MOUSE	CALX_CANFA	CAML_HUMAN	CD11_MOUSE
CD12_MOUSE	CD19_HUMAN	CD1A_HUMAN	CD1B_HUMAN	CD1C_HUMAN	CD1D_HUMAN	CD1E_HUMAN	CD27_HUMAN	CD28_BOVIN	CD2_HORSE
CD30_HUMAN	CD33_HUMAN	CD34_CANFA	CD36_BOVIN	CD3D_HUMAN	CD3E_CANFA	CD3G_HUMAN	CD3H_MOUSE	CD3Z_HUMAN	CD40_HUMAN
CD44_BOVIN	CD45_HUMAN	CD4_CANFA	CD5_BOVIN	CD6_HUMAN	CD7_HUMAN	CD80_HUMAN	CD83_HUMAN	CD86_HUMAN	CD8A_BOVIN
CD8B_MOUSE	CEK2_CHICK	CEK3_CHICK	CGM1_HUMAN	CINB_HUMAN	CINC_RAT	CR1_HUMAN	CR2_HUMAN	CRB_DROME	CRF4_HUMAN
CTL4_HUMAN	CYGD_BOVIN	CYGE_MOUSE	CYGF_HUMAN	CYGR_ARBPU	CYGS_STRPU	CYGX_RAT	CYRB_HUMAN	CYRG_BOVIN	DAF1_CAEEL
DAF4_CAEEL	DCC_HUMAN	DEXT_STRDO	DLK_HUMAN	DLL1_MOUSE	DL_DROME	DSC1_BOVIN	DSC2_HUMAN	DSC3_BOVIN	DSG1_BOVIN
DSG3_HUMAN	E310_ADE02	E3GL_ADE02	ECTO_RAT	EDD1_HUMAN	EFB1_HUMAN	EFB2_HUMAN	EFB3_HUMAN	EG15_CAEEL	EGFR_HUMAN
EGF_HUMAN	EGLN_MOUSE	EM24_YEAST	EPA1_HUMAN	EPA2_HUMAN	EPA3_CHICK	EPA4_CHICK	EPA5_CHICK	EPA6_MOUSE	EPA7_HUMAN
EPA8_MOUSE	EPB1_HUMAN	EPB2_CHICK	EPB3_HUMAN	EPB4_HUMAN	EPB5_CHICK	EPOR_HUMAN	ER53_HUMAN	ERB2_HUMAN	EV2A_HUMAN
EV2B_HUMAN	FAS2_SCHAM	FAS3_DROME	FASA_BOVIN	FAT_DROME	FCE1_RAT	FCEA_HUMAN	FCEG_CAVPO	FCG0_HUMAN	FCG1_HUMAN
FCG2_BOVIN	FCGA_HUMAN	FCGB_HUMAN	FCGC_HUMAN	FET3_YEAST	FGR1_CHICK	FGR2_DROME	FGR3_HUMAN	FGR4_HUMAN	FL3L_HUMAN
FLT3_HUMAN	FM1_ACTVI	FM2_ACTNA	FNBA_STAAU	FS21_DROME	G25L_CANFA	G49A_MOUSE	G49B_MOUSE	G731_HUMAN	G732_HUMAN
GARP_HUMAN	GCSR_HUMAN	GHRH_MOUSE	GHR_BOVIN	GLP1_CAEEL	GLPA_HUMAN	GLPB_HUMAN	GLPC_HUMAN	GLPE_HUMAN	GLP_HORSE
GP10_DICDI	GP38_CANFA	GP70_MOUSE	GPBA_HUMAN	GPBB_HUMAN	GPIX_HUMAN	GPV_HUMAN	GRK_DROME	HEMA_RACVI	HSER_CAVPO
I12R_HUMAN	I131_HUMAN	I132_HUMAN	ICA1_BOVIN	ICA2_HUMAN	ICA3_BOVIN	ICCR_DROME	IDD_HUMAN	IG1R_HUMAN	IL1R_HUMAN
IL1S_HUMAN	IL2A_BOVIN	IL2B_HUMAN	IL3A_MOUSE	IL3B_MOUSE	IL3R_HUMAN	IL4R_HUMAN	IL5R_HUMAN	IL6A_HUMAN	IL6B_HUMAN
IL7R_MOUSE	INGR_HUMAN	INGS_HUMAN	INLA_LISMO	INR1_BOVIN	INR2_HUMAN	INSR_DROME	IRE1_YEAST	IRR_CAVPO	ITA1_DROME
ITA2_DROME	ITA3_CRISP	ITA4_HUMAN	ITA5_HUMAN	ITA6_CHICK	ITA8_CHICK	ITA9_HUMAN	ITAB_HUMAN	ITAE_HUMAN	ITAL_HUMAN
ITAM_HUMAN	ITAV_CHICK	ITAX_HUMAN	ITB0_XENLA	ITB1_CHICK	ITB2_BOVIN	ITB3_HUMAN	ITB4_HUMAN	ITB5_HUMAN	ITB6_HUMAN
ITB7_HUMAN	ITB8_HUMAN	ITBX_DROME	KAPP_ARATH	KEX2_YEAST	KFMS_FELCA	KIR1_BOVIN	KIR2_HUMAN	KIR3_HUMAN	KIR4_HUMAN
KIR5_HUMAN	KIR6_CHICK	KKIT_BOVIN	KLTK_HUMAN	KPR0_MAIZE	KROS_HUMAN	LAG2_CAEEL	LAG3_HUMAN	LAGC_DICDI	LAR_DROME
LDL1_XENLA	LDL2_XENLA	LDLR_CRIGR	LDVR_CHICK	LEM1_BOVIN	LEM2_BOVIN	LEM3_BOVIN	LEPR_HUMAN	LEUK_HUMAN	LI12_CAEEL
LIN3_CAEEL	LPH_HUMAN	LRP1_CHICK	LRP_CAEEL	LT23_CAEEL	LU_HUMAN	LY9_MOUSE	M21_STRPY	M22_STRPY	M24_STRPY
M49_STRPY	M5_STRPY	M6_STRPY	MAGL_MOUSE	MAGS_MOUSE	MAG_HUMAN	MANR_HUMAN	MCP_HUMAN	MEPA_HUMAN	MEPB_HUMAN
MET_HUMAN	MINK_HUMAN	MPRD_BOVIN	MPRI_BOVIN	MRP4_STRPY	MRP_STRSU	MS2_HUMAN	MU18_HUMAN	MUC1_HUMAN	MX_STRPY
MYP0_BOVIN	NCA1_BOVIN	NCA2_RAT	NEU_RAT	NGCA_CHICK	NGFR_CHICK	NK10_HUMAN	NKR0_HUMAN	NKR1_HUMAN	NKR2_HUMAN
NKR3_HUMAN	NKR4_HUMAN	NKR5_HUMAN	NKR6_HUMAN	NKR7_HUMAN	NKR9_HUMAN	NOTC_BRARE	NRCA_CHICK	NRG_DROME	NRP_CHICK
NTC1_MOUSE	NTC4_MOUSE	OST4_CANFA	OSTA_YEAST	OSTB_YEAST	OX2G_RAT	OX40_HUMAN	P1P_LACLC	P2P_LACLA	P3P_LACLC
PA2R_BOVIN	PAC_STRMU	PCP2_HUMAN	PD1_HUMAN	PEC1_BOVIN	PEP1_YEAST	PGDR_HUMAN	PGDS_HUMAN	PGG2_RAT	PHLX_RABIT
PIGR_HUMAN	PK66_PLAKN	PRLR_BOVIN	PTP1_DROME	PTP6_DROME	PTP9_DROME	PTPA_HUMAN	PTPB_HUMAN	PTPD_HUMAN	PTPE_HUMAN
PTPF_HUMAN	PTPJ_HUMAN	PTPK_HUMAN	PTPM_HUMAN	PTPN_HUMAN	PTPO_RAT	PTPZ_HUMAN	PVDA_PLAKN	PVDB_PLAKN	PVDG_PLAKN
PVDR_PLAVI	PVR_MOUSE	RAGE_BOVIN	RET_HUMAN	RIB1_HUMAN	RIB2_HUMAN	RON_HUMAN	SCF_CANFA	SDC1_CRIGR	SDC2_HUMAN
SDC3_CHICK	SDC4_CHICK	SDC_DROME	SEPL_HUMAN	SERR_DROME	SHAK_DROME	SL17_ENTHI	SPA1_STAAU	SPA2_STAAU	SPAA_STRDO
SPAP_STRMU	SPER_STRPU	SPG1_STRSP	SPG2_STRSP	SPH_STRPY	SPIT_DROME	SRK6_BRAOL	SSRA_CANFA	SSRB_CANFA	SSRD_HUMAN
STRH_STRPN	TACT_HUMAN	TEE6_STRPY	TF_BOVIN	TGFA_HUMAN	TGR2_HUMAN	TIE1_BOVIN	TIE2_BOVIN	TMK1_ARATH	TML1_ARATH
TNR1_HUMAN	TNR2_HUMAN	TNRC_HUMAN	TOLL_DROME	TOP_DROME	TOR_DROME	TPOR_HUMAN	TRBM_HUMAN	TRK3_HUMAN	TRKA_HUMAN
TRKB_HUMAN	TRKC_HUMAN	TSA4_GIALA	TYO3_HUMAN	TYR2_HUMAN	TYRO_CHICK	UFO_HUMAN	UPK3_BOVIN	VCA1_HUMAN	VEGR_RAT
VGL2_CVH22	VGLI_HSVEB	VGP_EBOV	VGR1_HUMAN	VGR2_COTJA	VGR3_HUMAN	VP36_CANFA	WAPA_STRMU	XMRK_XIPMA	ZAN_PIG
ZIPP_DROME	ZP2_FELCA	ZP3A_CALSQ	ZP3_BOVIN	ZPB_FELCA					

(2) 152 type II transmembrane proteins

41BL_HUMAN	4F2_HUMAN	A15_HUMAN	ALG5_YEAST	AMPE_HUMAN	AMPN_HUMAN	ASPH_BOVIN	ATHB_CANFA	ATNB_ANGAN	ATNC_BOVIN
ATND_BUFMA	ATNG_BOVIN	BAGT_LYMST	BGAT_HUMAN	BGIB_HUMAN	BST2_HUMAN	CAG1_CHICK	CAG2_HUMAN	CAG4_CHICK	CAG6_HUMAN
CAGB_MOUSE	CAGC_HUMAN	CD2L_HUMAN	CD37_HUMAN	CD38_HUMAN	CD3L_HUMAN	CD4L_HUMAN	CD53_HUMAN	CD63_HUMAN	CD69_HUMAN
CD72_HUMAN	CD81_HUMAN	CD82_HUMAN	CD94_HUMAN	CD9_BOVIN	CME1_BACSU	CO02_HUMAN	CYAG_DICDI	DAP1_YEAST	DAP2_YEAST
DPP4_HUMAN	ECE1_BOVIN	ECE2_BOVIN	EXBD_ECOLI	FCE2_HUMAN	FTSL_ECOLI	FTSN_ECOLI	FTSQ_ECOLI	FUT1_HUMAN	FUT2_HUMAN
FUT3_BOVIN	FUT4_HUMAN	FUT5_HUMAN	FUT6_HUMAN	FUT7_HUMAN	G6NT_BOVIN	GATR_BOVIN	GCS1_HUMAN	GDA1_YEAST	GM12_SCHPO
GNT1_HUMAN	GNT2_HUMAN	GNT3_HUMAN	GNT5_HUMAN	HEPS_HUMAN	ILT4_HUMAN	IM23_SCHHA	KRE2_YEAST	KRE6_CANAL	KTR1_YEAST
KTR2_YEAST	KUCR_MOUSE	LAFU_VIBPA	LECH_HUMAN	LECI_HUMAN	LEPS_BACSU	LEP_ECOLI	LHA1_RHOAC	LHA2_ECTHL	LHA3_RHOAC
LHA4_RHOAC	LHA5_RHOPA	LHA6_RHOAC	LHA7_RHOAC	LHA_CHLAU	LHB1_RHOAC	LHB2_ECTHL	LHB3_RHOAC	LHB4_RHOAC	LHB5_RHOAC
LHB6_RHOAC	LHB7_RHOAC	LHB_CHLAU	LPG1_LEIDO	LY4A_MOUSE	LY4B_MOUSE	LY4D_MOUSE	LY4E_MOUSE	LY4F_MOUSE	LY4G_MOUSE
LYAH_MOUSE	LYII_HUMAN	LYTR_BACSU	M121_DROME	M122_DROME	MA12_HUMAN	MAN2_MOUSE	MANX_MOUSE	MMGL_MOUSE	MNS1_YEAST
MOTB_BACSU	MSRE_BOVIN	N121_RAT	4)galactosylNEP_HUMAN		NK11_MOUSE	NK12_MOUSE	NK13_RAT	NK14_MOUSE	NKGA_HUMAN
NKGC_HUMAN	NKGD_HUMAN	NKGE_HUMAN	NRT_DROME	OCH1_YEAST	OM22_NEUCR	OX4L_HUMAN	P152_YEAST	PAGT_BOVIN	PBPB_BACSU
PC1_HUMAN	PSM_HUMAN	SC20_YEAST	SC66_YEAST	SED4_YEAST	SKN1_CANAL	SPC2_CANFA	SPC3_CANFA	SPC4_CANFA	STUB_DROME
SUIS_HUMAN	SYB1_HUMAN	SYB2_HUMAN	SYB_APLCA	TAL6_HUMAN	TLPA_BRAJA	TNFA_BOVIN	TOLA_ECOLI	TOLQ_ECOLI	TRSR_HUMAN
UPKA_BOVIN	UPKB_BOVIN								

(3) 1311 multi-pass transmembrane proteins

5H1A_HUMAN	5H1B_CRIGR	5H1D_CANFA	5H1E_HUMAN	5H1F_HUMAN	5H2A_CRIGR	5H2B_HUMAN	5H2C_HUMAN	5H4_RAT	5H5A_HUMAN
5H5B_MOUSE	5H6_HUMAN	5H7_CAVPO	5HT1_APLCA	5HT2_APLCA	5HT3_HUMAN	5HTA_DROME	5HTB_DROME	5HT_BOMMO	A1AA_HUMAN
A1AB_HUMAN	A1AC_BOVIN	A2AA_CAVPO	A2AB_CAVPO	A2AC_CAVPO	A2AD_HUMAN	A2AR_CARAU	A3_VIGUN	A4P_HUMAN	AA1R_BOVIN
AA2A_CANFA	AA2B_HUMAN	AA3R_CANFA	AAAT_MOUSE	AAS_ECOLI	AC22_STRCO	AC45_BOVIN	ACAT_HUMAN	ACH1_CAEEL	ACH2_CAEEL
ACH3_BOVIN	ACH4_CAEEL	ACH5_CAEEL	ACH6_CAEEL	ACH7_BOVIN	ACH9_RAT	ACHA_BOVIN	ACHB_BOVIN	ACHD_BOVIN	ACHE_BOVIN
ACHG_BOVIN	ACHN_CHICK	ACHO_CARAU	ACHP_CARAU	ACM1_DROME	ACM2_CHICK	ACM3_BOVIN	ACM4_CHICK	ACM5_HUMAN	ACRB_ECOLI
ACRF_ECOLI	ACSA_ACEXY	ACTR_BOVIN	ADT1_BOVIN	ADT2_ARATH	ADT3_BOVIN	ADT_ANOGA	AFQ2_STRCO	AG22_HUMAN	AG2R_BOVIN
AG2S_HUMAN	ALCP_BACP3	ALG8_YEAST	ALKB_PSEOL	ALP1_YEAST	ALST_BACSU	AMPE_ECOLI	AMSL_ERWAM	AMT_CORGL	ANSP_ECOLI
APJ_HUMAN	APRD_PSEAE	AQP1_BOVIN	AQP2_HUMAN	AQP3_HUMAN	AQP4_HUMAN	AQP5_HUMAN	AQPA_RANES	AQPL_HUMAN	AQPZ_ECOLI
AQUA_ATRCA	AR11_YEAST	ARAE_ECOLI	ARAH_ECOLI	ARE1_YEAST	ARE2_YEAST	AROP_CORGL	AT7A_HUMAN	AT7B_HUMAN	ATA1_SYNY3
ATC1_DICDI	ATC2_YEAST	ATC3_HUMAN	ATC4_YEAST	ATC5_YEAST	ATC8_YEAST	ATC9_YEAST	ATCA_RABIT	ATCB_CHICK	ATCE_HUMAN
ATCF_HUMAN	ATCL_MYCGE	ATCP_HUMAN	ATCQ_HUMAN	ATCR_HUMAN	ATCS_SYNP7	ATCX_SCHPO	ATC_ARTSF	ATHA_CANFA	ATHL_HUMAN
ATKA_ECOLI	ATKB_ECOLI	ATM1_YEAST	ATMA_ECOLI	ATMB_SALTY	ATN1_BUFMA	ATN2_CHICK	ATN3_CHICK	ATNA_ANGAN	ATP6_ALBCO
ATPI_ANTSP	ATPY_YEAST	ATR1_YEAST	ATSY_SYNP7	ATU1_YEAST	ATU2_YEAST	ATXA_LEIDO	ATXB_LEIDO	AVP3_ARATH	B1AR_CANFA
B2AR_CANFA	B3A2_HUMAN	B3A3_HUMAN	B3AR_BOVIN	B3AT_CHICK	B4AR_MELGA	BAC1_HALS1	BAC2_HALS2	BAC3_HALVA	BACA_RHIME
BACH_HALHP	BACR_HALAR	BACS_HALHA	BACT_HALSA	BCR_ECOLI	BCSA_ACEXY	BENE_ACICA	BETP_CORGL	BETT_ECOLI	BFR1_SCHPO
BIB_DROME	BIOX_BACSH	BLR1_HUMAN	BMR1_BACSU	BMR2_BACSU	BMRP_CANAL	BNA1_HUMAN	BOFA_BACSU	BRAB_PSEAE	BRAD_PSEAE
BRAE_PSEAE	BRAZ_PSEAE	BRB1_HUMAN	BRB2_HUMAN	BRNQ_ECOLI	BROW_DROME	BRS3_CAVPO	BRS4_BOMOR	C24B_HUMAN	C550_BACSU
C560_BOVIN	C561_HUMAN	C5AR_CANFA	CADA_BACFI	CADB_ECOLI	CADD_STAAU	CAFA_YERPE	CAIT_ECOLI	CAKB_BOVIN	CALR_HUMAN

$\mathbf{C}_\xi^{-1}$ is the inverse matrix of $\mathbf{C}_\xi$; Π is the symbol for operating the product of multiple factors (e.g., $\Pi_{i=1}^{19}\ \lambda_i^\xi$ represents the product of λ_i^ξ for i from 1 to 19), and λ_i^ξ is the ith eigenvalue of the matrix $\mathbf{C}_\xi$. It can be proved that the covariance matrix $\mathbf{C}_\xi$ as defined by equation (8) has no negative eigenvalues; it has one, and only one, eigenvalue equal to zero.[8] Such a null eigenvalue is represented here by λ_{20}^ξ and excluded from equation (6). Actually, the

TABLE I. (Continued)

CAMG_HUMAN	CAN1_CANAL	CAR1_DICDI	CAR2_DICDI	CAR3_DICDI	CARR_MYXXA	CASR_BOVIN	CB11_RABIT	CB12_RABIT	CB1A_FUGRU
CB1B_FUGRU	CB1R_HUMAN	CB21_RABIT	CB22_RABIT	CB2R_HUMAN	CBIN_SALTY	CBIQ_SALTY	CCKR_CAVPO	CCP1_RAT	CCR3_HUMAN
CCR4_BOVIN	CCT1_RAT	CD20_HUMAN	CD2R_HUMAN	CD47_HUMAN	CD97_HUMAN	CDSA_ECOLI	CFTR_BOVIN	CGRR_HUMAN	CHAA_ECOLI
CHL1_ARATH	CHS2_NEUCR	CHS3_NEUCR	CHS4_NEUCR	CHS_SAPMO	CIC1_CYPCA	CIC2_HUMAN	CIC5_HUMAN	CICB_RAT	CICC_RABIT
CICG_HUMAN	CICH_TORCA	CICK_HUMAN	CICL_HUMAN	CICP_BOVIN	CIK1_DROME	CIK2_DROME	CIK3_HUMAN	CIK4_BOVIN	CIK5_HUMAN
CIK6_HUMAN	CIKA_RAT	CIKB_DROME	CIKD_HUMAN	CIKE_DROME	CIKF_RAT	CIKG_RAT	CIKL_DROME	CIKW_DROME	CIN1_LOLBL
CIN2_RAT	CIN3_RAT	CIN4_HUMAN	CIN5_RAT	CIN6_HUMAN	CINA_DROME	CIT1_ECOLI	CITN_KLEPN	CKR1_HUMAN	CKR2_HUMAN
CKR3_HUMAN	CKR4_HUMAN	CKR5_HUMAN	CKR6_HUMAN	CKR7_HUMAN	CKRV_MOUSE	CLC1_HUMAN	CLC2_HUMAN	CLC3_HUMAN	CLC4_HUMAN
CLC5_HUMAN	CLC6_HUMAN	CLC7_RAT	CLD1_ECOLI	CLD2_ECOLI	CLD_SALTY	CMCT_NOCLA	CMLA_PSEAE	CMLR_STRLI	CMST_CRIGR
CNG1_BOVIN	CNG2_BOVIN	CNG3_BOVIN	CNG4_BOVIN	CNGX_RAT	COA0_HELPY	COA1_HELPY	COA2_HELPY	COA3_HELPY	CODB_ECOLI
COMA_STRPN	COMP_BACSU	COQ2_YEAST	COX1_ALBCO	COX2_ACHDO	COX3_BACFI	COX4_BACFI	COXM_BRAJA	COXN_BRAJA	COXX_BACFI
COXY_YEAST	CPSD_STRAG	CPT1_YEAST	CPXA_ECOLI	CRED_ECOLI	CRF2_HUMAN	CRFR_CHICK	CRNA_EMENI	CSCB_ECOLI	CSG2_YEAST
CTK1_RABIT	CTPA_MYCLE	CTPB_MYCLE	CTR1_HUMAN	CTR2_HUMAN	CVAB_ECOLI	CX1A_PARDE	CX1B_PARDE	CX32_ARATH	CX33_MICUN
CX35_RAJER	CX41_XENLA	CX43_BRARE	CX56_CHICK	CXA1_BOVIN	CXA2_XENLA	CXA3_BOVIN	CXA4_HUMAN	CXA5_CANFA	CXA6_CANFA
CXA7_RAT	CXA8_CHICK	CXB1_HUMAN	CXB2_HUMAN	CXB3_MOUSE	CXB4_MOUSE	CXB5_MOUSE	CXB6_MOUSE	CY14_NEUCR	CYA1_BOVIN
CYA2_RAT	CYA3_RAT	CYA4_RAT	CYA5_CANFA	CYA6_CANFA	CYA7_BOVIN	CYA8_HUMAN	CYA9_MOUSE	CYAA_ANACY	CYAB_BORPE
CYBH_ALCEU	CYB_SULAC	CYCM_BRAJA	CYDA_AZOVI	CYDB_ECOLI	CYHR_CANMA	CYOA_ECOLI	CYOB_ECOLI	CYOC_ECOLI	CYOD_ECOLI
CYOE_ECOLI	CYPR_CALVI	CYST_SYNP7	D1DR_CARAU	D2D1_XENLA	D2DR_BOVIN	D3DR_CERAE	D4DR_HUMAN	D5DR_FUGRU	DADR_DIDMA
DAGA_ALTHA	DAL4_YEAST	DAL5_YEAST	DBDR_HUMAN	DCDR_XENLA	DCOB_KLEPN	DCOG_KLEPN	DCTA_ECOLI	DCTB_RHILE	DCTS_RHOCA
DEG1_CAEEL	DEGW_CAEEL	DEGX_CAEEL	DEL1_CAEEL	DHAQ_ACEPO	DHG_ECOLI	DHSC_COXBU	DHSD_COXBU	DIHR_ACHDO	DIP5_YEAST
DIP_ANTMA	DIVJ_CAUCR	DMSC_ECOLI	DOP1_DROME	DOP2_DROME	DPPB_ECOLI	DPPC_ECOLI	DSBB_ECOLI	DSBD_ECOLI	DTD_HUMAN
DTPT_LACHE	DUFF_HUMAN	DUR3_YEAST	EAT1_BOVIN	EAT2_HUMAN	EAT3_BOVIN	EAT4_HUMAN	EAT_CAEEL	EDG1_HUMAN	EDG2_BOVIN
EDG3_HUMAN	EMP1_HUMAN	EMP2_HUMAN	EMP3_HUMAN	EMR1_HUMAN	EMRB_ECOLI	ENVZ_ECOLI	EPT1_YEAST	ER21_CAEEL	ER22_CAEEL
ERD1_KLULA	ERD2_ARATH	ERS1_YEAST	ET1R_BOVIN	ET3R_XENLA	ETBR_BOVIN	EXBB_ECOLI	EXOQ_RHIME	EXOY_RHIME	EXOZ_RHIME
EXUT_ECOLI	F480_MOUSE	FADL_ECOLI	FANA_HELAS	FASD_ECOLI	FATP_MOUSE	FCEB_HUMAN	FCY2_YEAST	FDFT_HUMAN	FDNH_ECOLI
FDNI_ECOLI	FDOH_ECOLI	FDOI_ECOLI	FDXH_HAEIN	FEOB_ECOLI	FEPD_ECOLI	FEPG_ECOLI	FET4_YEAST	FEUB_BACSU	FEUC_BACSU
FHUA_ECOLI	FIXG_RHIME	FIXI_RHIME	FIXL_AZOCA	FLHA_ECOLI	FLIP_ECOLI	FLIQ_ECOLI	FLIR_ECOLI	FLX1_YEAST	FML1_HUMAN
FML2_HUMAN	FMLR_HUMAN	FPS1_YEAST	FRDC_ECOLI	FRDD_ECOLI	FRE1_YEAST	FRIZ_DROME	FRP1_SCHPO	FSHR_BOVIN	FTH1_HAEIN
FTH2_SYNY3	FTH3_SYNY3	FTH4_SYNY3	FTSH_BACSU	FTSW_ECOLI	FUR4_SCHPO	G10D_MOUSE	G6PT_HUMAN	GAA1_BOVIN	GAA2_BOVIN
GAA3_BOVIN	GAA4_BOVIN	GAA5_HUMAN	GAA6_HUMAN	GAB1_BOVIN	GAB2_HUMAN	GAB3_CHICK	GAB4_CHICK	GABP_BACSU	GAB_DROME
GAC1_RAT	GAC2_BOVIN	GAC3_MOUSE	GAC4_CHICK	GAD_MOUSE	GAL2_YEAST	GALP_ECOLI	GALR_HUMAN	GAP1_YEAST	GAR1_HUMAN
GAR2_HUMAN	GAR3_RAT	GAS1_HUMAN	GASR_CANFA	GC96_HUMAN	GCRC_MOUSE	GCRT_CHICK	GCY6_HUMAN	GEF1_YEAST	GHSR_HUMAN
GIPR_HUMAN	GLCP_SYNY3	GLF_ZYMMO	GLHR_ANTEL	GLNP_ECOLI	GLPF_BACSU	GLPR_HUMAN	GLPT_BACSU	GLR1_HUMAN	GLR2_HUMAN
GLR3_HUMAN	GLR4_HUMAN	GLR5_HUMAN	GLR6_HUMAN	GLR7_HUMAN	GLRK_CHICK	GLR_HUMAN	GLTP_BACSU	GLTS_ECOLI	GLTT_BACCA
GMS1_SCHPO	GNP1_YEAST	GNS1_YEAST	GNTP_BACLI	GP21_RAT	GPCR_LYMST	GPR1_HUMAN	GPR2_HUMAN	GPR3_HUMAN	GPR4_HUMAN
GPR5_HUMAN	GPR6_HUMAN	GPR7_HUMAN	GPR8_HUMAN	GPRA_HUMAN	GPRC_HUMAN	GPRD_HUMAN	GPRE_RAT	GPRF_HUMAN	GPRH_HUMAN
GPRJ_HUMAN	GPRK_HUMAN	GPRL_HUMAN	GPRM_HUMAN	GPRN_HUMAN	GPRO_HUMAN	GPT_CRIGR	GRA1_HUMAN	GRA2_BACSU	GRA3_RAT
GRB2_BACSU	GRB_HUMAN	GRFR_HUMAN	GRHR_BOVIN	GRPR_HUMAN	GTR1_BOVIN	GTR2_HUMAN	GTR3_CANFA	GTR4_BOVIN	GTR5_HUMAN
GTR7_RAT	GTRL_DROME	GU27_RAT	GUDT_BACSU	GUFA_MYXXA	GUSB_BOVIN	H218_RAT	HAK1_SCHOC	HEX6_RICCO	HGT1_KLULA
HH1R_BOVIN	HH2R_CANFA	HIP1_YEAST	HISM_ECOLI	HISQ_ECOLI	HLY2_ECOLI	HLYB_ACTAC	HM74_HUMAN	HMC2_DESVH	HMC3_DESVH
HMC4_DESVH	HMC5_DESVH	HMD1_ARATH	HMD2_ARATH	HMDH_BLAGE	HMT1_SCHPO	HMUU_YERPE	HN36_HUMAN	HNM1_YEAST	HOXN_ALCEU
HS30_YEAST	HST6_CANAL	HUP1_CHLKE	HXT1_YEAST	HXT2_YEAST	HXT3_YEAST	HXT4_YEAST	HXT5_YEAST	HXT6_YEAST	HXT7_YEAST
HXTC_YEAST	HXTD_YEAST	HXTE_YEAST	HXTG_YEAST	HYBB_ECOLI	HYCC_ECOLI	HYCD_ECOLI	HYFC_ECOLI	IL8A_GORGO	IL8B_BOVIN
IM17_YEAST	IM22_YEAST	IM23_YEAST	IMMA_CITFR	IMMB_ECOLI	IMP1_YEAST	IMP2_YEAST	INA1_TRIHA	IP3R_DROME	IP3S_HUMAN
IPAB_SHIDY	IPYR_PHAAU	IRK1_HUMAN	IRK2_CAVPO	IRK3_CHICK	IRK4_HUMAN	IRK5_HUMAN	IRK6_HUMAN	IRK8_HUMAN	IRK9_HUMAN
IRKA_HUMAN	IRKB_HUMAN	IRKC_HUMAN	IRKE_HUMAN	IRKI_HUMAN	IRKX_RAT	ITR1_SCHPO	ITR2_YEAST	KBAA_BACSU	KDGL_ECOLI
KDGT_BACSU	KEFB_ECOLI	KEFC_ECOLI	KGTP_ECOLI	KHT2_KLULA	KINB_BACSU	KINC_BACSU	KUP_ECOLI	LACF_AGRRD	LACG_AGRRD
LACP_KLULA	LACY_CITFR	LAFT_VIBPA	LAMB_ECOLI	LCN3_LACLA	LCNC_LACLA	LCRD_YEREN	LEP3_PSEAE	LGT_ECOLI	LIMA_PSEGL
LIVH_ECOLI	LIVM_ECOLI	LLDP_ECOLI	LM1P_BOVIN	LMRA_STRLN	LNT_ECOLI	LPLB_BACSU	LPLC_BACSU	LSHR_BOVIN	LSPA_BACSU
LYP1_YEAST	LYSI_CORGL	LYSP_ECOLI	M2OM_BOVIN	M6A_HUMAN	M6B_MOUSE	MA3T_YEAST	MA6T_YEAST	MAE1_SCHPO	MALC_STRPN
MALD_STRPN	MALF_ECOLI	MALG_ECOLI	MAL_HUMAN	MAM2_SCHPO	MAP3_SCHPO	MAS_HUMAN	MC3R_HUMAN	MC4R_HUMAN	MC5R_HUMAN
MCAT_RAT	MCBE_ECOLI	MCP1_ECOLI	MCP2_ECOLI	MCP3_ECOLI	MCP4_ECOLI	MCPA_CAUCR	MCPC_SALTY	MCPD_ENTAE	MCPS_ENTAE
MD10_YEAST	MDOH_ECOLI	MDR1_CAEEL	MDR2_CRIGR	MDR3_CAEEL	MDR4_DROME	MDR5_DROME	MDR_LEITA	ME10_CAEEL	MEC4_CAEBR
MELB_ECOLI	MEP1_YEAST	MEP2_YEAST	MEP3_YEAST	MERT_PSEAE	MESD_LEUME	MEXB_PSEAE	MGLC_ECOLI	MGR1_HUMAN	MGR2_HUMAN
MGR3_HUMAN	MGR4_HUMAN	MGR5_HUMAN	MGR6_RAT	MGR7_HUMAN	MGR8_HUMAN	MIP_BOVIN	ML1A_CHICK	ML1B_HUMAN	ML1C_CHICK
ML1X_HUMAN	MMM1_YEAST	MMR_BACSU	MOG_BOVIN	MOT1_CRILO	MOTA_BACSU	MPCP_BOVIN	MRED_BACSU	MRG_HUMAN	MRP1_HUMAN
MRS3_YEAST	MRS4_YEAST	MSBB_ECOLI	MSCL_CLOPE	MSHR_BOVIN	MSP1_YEAST	MTRC_METTH	MTRD_METTH	MTRE_METTH	MTR_ECOLI
MYP1_XENLA	MYP2_XENLA	MYPR_BOVIN	NAC1_BOVIN	NAC2_RAT	NAC3_RAT	NAG1_HUMAN	NAG2_HUMAN	NAG3_PIG	NAH1_BOVIN
NAH2_RABIT	NAH3_DIDMA	NAH4_RAT	NAH_SCHPO	NAMI_BOVIN	NANT_ECOLI	NANU_RABIT	NAPA_ENTHR	NARI_BACSU	NARK_BACSU
NARV_ECOLI	NASA_BACSU	NASU_RAT	NDHF_BACSU	NDVB_RHIME	NFRA_ECOLI	NFRB_ECOLI	NHAA_ECOLI	NHAB_ECOLI	NHAC_BACFI
NI9M_BOVIN	NISK_LACLA	NIST_LACLA	NK1R_CAVPO	NK2R_BOVIN	NK3R_HUMAN	NKC1_HUMAN	NKC2_MOUSE	NKCL_MANSE	NMBR_HUMAN
NME1_MOUSE	NME2_MOUSE	NME3_HUMAN	NME4_MOUSE	NMPC_ECOLI	NMZ1_HUMAN	NO26_SOYBN	NORA_STAAU	NPT1_HUMAN	NPT2_HUMAN
NQO7_PARDE	NQO8_PARDE	NQOA_PARDE	NQOB_PARDE	NQOC_PARDE	NQOD_PARDE	NQOE_PARDE	NSR_LACLA	NTBE_CANFA	NTCH_RAT
NTCI_CRIGR	NTCP_HUMAN	NTCR_HUMAN	NTDO_BOVIN	NTG1_HUMAN	NTG2_MOUSE	NTG3_HUMAN	NTGL_BOVIN	NTNO_BOVIN	NTPI_ENTHR
NTPJ_ENTHR	NTPR_RAT	NTR1_HUMAN	NTR2_MOUSE	NTRY_AZOCA	NTS1_RAT	NTS2_RAT	NTSE_DROME	NTT4_RAT	NTT7_RAT
NTTA_CANFA	NU2C_MAIZE	NU2M_ALBCO	NU5C_SYNP2	NUOA_ECOLI	NUOH_ECOLI	NUOJ_ECOLI	NUOK_ECOLI	NUOL_ECOLI	NUOM_ECOLI
NUON_ECOLI	NUPC_BACSU	NUPG_ECOLI	NURM_NEUCR	NY1R_HUMAN	NY2R_BOVIN	NY4R_HUMAN	NY5R_HUMAN	NY6R_MOUSE	NYR_DROME
OAR_BOMMO	OL1E_HUMAN	OLF0_RAT	OLF1_CANFA	OLF2_CANFA	OLF3_CANFA	OLF4_CANFA	OLF5_CHICK	OLF6_CHICK	OLF7_RAT
OLF8_RAT	OLF9_RAT	OLFD_CANFA	OLFE_HUMAN	OLFI_HUMAN	OLFJ_HUMAN	OM06_YEAST	OM07_YEAST	OM11_HAEIN	OM12_HAEIN
OM20_NEUCR	OM21_HAEIN	OM22_HAEIN	OM23_HAEIN	OM24_HAEIN	OM25_HAEIN	OM32_COMAC	OM37_YEAST	OM40_NEUCR	OM51_HAEIN
OM52_HAEIN	OM53_HAEIN	OM6B_CHLTR	OM6C_CHLTR	OM6E_CHLTR	OM6L_CHLTR	OM6_CHLPN	OM70_NEUCR	OMA1_NEIGO	OMA2_NEIME
OMB1_NEIGO	OMB2_NEIGO	OMB3_NEIME	OMB4_NEIME	OMB_NEILA	OMP1_CHLPN	OMP2_CHLPS	OMP3_CHLPS	OMP4_NEIME	OMPA_BORAV
OMPB_CHLTR	OMPC_CHLTR	OMPE_CHLTR	OMPF_CHLTR	OMPH_CHLTR	OMPL_CHLTR	OMPM_CHLTR	OMPN_CHLTR	OMPX_ECOLI	OMP_BORPE
OPPB_ECOLI	OPPC_ECOLI	OPRD_HUMAN	OPRK_CAVPO	OPRM_HUMAN	OPRX_CAVPO	OPS1_CALVI	OPS2_DROME	OPS3_DROME	OPS4_DROME
OPSB_ANOCA	OPSD_ALLMI	OPSG_ASTFA	OPSH_ASTFA	OPSI_ASTFA	OPSP_CHICK	OPSR_ANOCA	OPSU_BRARE	OPSV_CHICK	OPUB_BACSU
OPUD_BACSU	OUSA_ERWCH	OXYR_HUMAN	P2UR_HUMAN	P2X1_HUMAN	P2X2_RAT	P2X3_RAT	P2X4_HUMAN	P2X5_HUMAN	P2X6_RAT
P2X7_HUMAN	P2Y3_CHICK	P2Y4_HUMAN	P2Y6_HUMAN	P2Y7_HUMAN	P2Y8_XENLA	P2YR_BOVIN	P47A_CANBO	P47B_CANBO	PACR_BOVIN
PAFR_CAVPO	PAGC_SALTY	PANF_ECOLI	PAPC_ECOLI	PAR2_HUMAN	PATC_DROME	PBP4_NOCLA	PBUX_BACSU	PD2R_HUMAN	PDR5_YEAST
PDUF_SALTY	PECM_ERWCH	PEDD_PEDAC	PER1_HUMAN	PER2_HUMAN	PER3_BOVIN	PER4_HUMAN	PET1_HUMAN	PET2_HUMAN	PEX2_CRIGR
PEX3_PICAN	PEXC_PICPA	PEXD_PICPA	PF2R_BOVIN	PGPA_ECOLI	PGSA_BACSU	PGTB_SALTY	PGTP_SALTY	PH84_YEAST	PHEP_ECOLI
PHOE_CITFR	PHOR_ECOLI	PI2R_HUMAN	PIGA_HUMAN	PIGF_HUMAN	PIP_LACLA	PKBS_BOVIN	PKN6_MYXXA	PLLP_RAT	PM1_HUMAN
PM22_HUMAN	PMA1_AJECA	PMA2_ARATH	PMA3_ARATH	PMA4_NICPL	PMIP_NICAL	PMP2_MOUSE	PMP7_HUMAN	PMT1_YEAST	PNTA_ECOLI
PNTB_ECOLI	PNUC_ECOLI	PONA_DICDI	PORD_PSEAE	PORF_PSEAE	PORI_RHOBL	PORO_PSEAE	PORP_PSEAE	POTE_ECOLI	PPA1_YEAST
PRA1_USTHO	PRA2_USTMA	PRO1_LEIEN	PROP_ECOLI	PROW_BACSU	PROY_ECOLI	PRTD_ERWCH	PSAA_ANAVA	PSAB_ANAVA	PSAL_SYNEN
PSB1_ANASP	PSB2_ANASP	PSBA_AMAHY	PSBD_CHLRE	PSN1_HUMAN	PSN2_HUMAN	PSS1_CRILO	PSS_BACSU	PSTA_ECOLI	PSTC_ECOLI
PSY_NEUCR	PT2A_ARATH	PT2B_ARATH	PTAA_ECOLI	PTBA_BACSU	PTCC_ECOLI	PTDA_ECOLI	PTFB_ECOLI	PTFC_BACSU	PTFD_BACSU
PTGA_BACSU	PTGB_ECOLI	PTHB_ECOLI	PTKC_ECOLI	PTLB_LACCA	PTMA_BACSU	PTMB_BACST	PTNC_ECOLI	PTND_ECOLI	PTOA_ECOLI
PTR2_CANAL	PTRR_DIDMA	PTSA_PEDPE	PTSB_BACSU	PTTB_BACSU	PUR8_STRLP	PUT4_YEAST	PUTP_ECOLI	PUTX_EMENI	PXA1_YEAST
PXA2_YEAST	P_HUMAN	QACA_STAAU	QAY_NEUCR	QOX1_ACEAC	QOX2_ACEAC	QOX3_BACSU	QOX4_BACSU	QOXM_SULAC	QUTD_EMENI
RAFB_ECOLI	RAFP_PEDPE	RAG1_KLULA	RBSC_BACSU	RCEL_CHLAU	RCEM_CHLAU	RCO3_NEUCR	RDC1_CANFA	RDS_BOVIN	RDXA_RHOSH
REIS_TODPA	RFAL_ECOLI	RFBP_SALTY	RFE_ECOLI	RGR_BOVIN	RGT2_YEAST	RHAG_HUMAN	RHAT_ECOLI	RHCE_HUMAN	RHD_HUMAN
RHLA_PANTR	RHLC_GORGO	RHLD_GORGO	RHLF_PANTR	RHLR_PANTR	RHL_HYLPI	RHOM_DROME	ROCE_BACSU	RODA_ECOLI	ROM1_BOVIN

TABLE I. (Continued)

RT1B_ACTPL	RT3B_ACTPL	RTA_RAT	S61A_CANFA	SANA_ECOLI	SAT1_RAT	SATT_HUMAN	SBMA_ECOLI	SC62_YARLI	SCAA_BOVIN
SCAB_HUMAN	SCAD_HUMAN	SCAG_HUMAN	SCRC_HUMAN	SCRT_DROME	SCRY_KLEPN	SDH3_YEAST	SDH4_YEAST	SE12_CAEEL	SECD_ECOLI
SECE_ECOLI	SECF_ECOLI	SECG_ECOLI	SECY_ANTSP	SENR_RAT	SEO1_YEAST	SFUB_SERMA	SHIA_ECOLI	SLY4_YEAST	SNF3_YEAST
SNQ2_YEAST	SOT1_SPIOL	SP2E_BACSU	SP5E_BACSU	SPAK_BACSU	SPAT_BACSU	SPC1_YEAST	SPE4_CAEEL	SSR1_HUMAN	SSR2_BOVIN
SSR3_HUMAN	SSR4_HUMAN	SSR5_HUMAN	STE2_SACKL	STE3_YEAST	STE6_YEAST	STL1_YEAST	STP1_ARATH	STT3_CAEEL	STV1_YEAST
SUL1_YEAST	SUR_CRICR	SYNP_RAT	SYPH_BOVIN	SYRD_PSESY	SYT3_RAT	SYT4_MOUSE	SYT5_HUMAN	TA2R_BOVIN	TAP1_HUMAN
TAP2_HUMAN	TAT2_YEAST	TCMA_STRGA	TCR1_ECOLI	TCR2_BACSU	TCR3_ECOLI	TCR4_SALOR	TCR5_ECOLI	TCR7_VIBAN	TCR8_PASMU
TCRB_BACSU	TCR_BACST	TEHA_ECOLI	TERC_ALCSP	TH11_TRYBB	TH12_TRYBB	TH2A_TRYBB	THAS_HUMAN	THI7_YEAST	THIX_YEAST
THIY_YEAST	THRR_CRILO	TIP1_TOBAC	TIP2_TOBAC	TIPA_ARATH	TIPE_DROME	TIPG_ARATH	TIPR_ARATH	TIPW_LYCES	TJ6_MOUSE
TLR1_DROME	TLR2_DROME	TNAB_ECOLI	TOK1_YEAST	TRA2_CAEBR	TRAN_ECOLI	TRBA_ECOLI	TRBE_ECOLI	TRBI_ECOLI	TRD1_ECOLI
TRD2_ECOLI	TRFR_HUMAN	TRG1_ECOLI	TRK1_SACUV	TRK2_YEAST	TRKG_ECOLI	TRKH_ECOLI	TRK_SCHPO	TSAB_RICTS	TSAG_RICTS
TSAK_RICTS	TSAR_RICTS	TSAS_RICTS	TSAT_RICTS	TSAW_RICTS	TSCC_HUMAN	TSHR_BOVIN	TSX_ECOLI	TUTB_ERWHE	TXKR_HUMAN
TXTP_BOVIN	TYRP_ECOLI	UAPA_EMENI	UAPC_EMENI	UBIA_ECOLI	UCP1_HUMAN	UCP2_HUMAN	UGA4_YEAST	UGAT_HUMAN	UHPB_ECOLI
UHPC_ECOLI	UHPT_ECOLI	UL33_HCMVA	UN17_CAEEL	UN36_CAEEL	UNC7_CAEEL	UNC8_CAEEL	URAA_ECOLI	US27_HCMVA	US28_HCMVA
UT1_HUMAN	UT2_HUMAN	V1AR_HUMAN	V1BR_HUMAN	V2R_BOVIN	VAL1_YEAST	VATL_ARATH	VC03_SPVKA	VG74_HSVSA	VGLB_HSVA1
VGLM_INSV	VIPR_CARAU	VIPS_HUMAN	VK02_SPVKA	VLOM_LAMBD	VLYS_LAMBD	VM11_YEAST	VM21_YEAST	VMT1_HUMAN	VMT2_BOVIN
VPH1_YEAST	VPP1_RAT	VQ3L_CAPVK	VU51_HSV6U	WC1A_ARATH	WC1B_ARATH	WC1C_ARATH	WC2A_ARATH	WC2B_ARATH	WC2C_ARATH
WHIT_ANOAL	XAPB_ECOLI	XYLE_ECOLI	Y4JF_RHISN	Y4MJ_RHISN	Y736_HAEIN	Y889_HELPY	YADS_ECOLI	YAF3_YEAST	YBE2_YEAST
YCEE_ECOLI	YCKJ_BACSU	YD19_METJA	YEP0_YEAST	YG90_HAEIN	YGT6_YEAST	YKH3_CAEEL	YMN2_CAEEL	YNZ3_CAEEL	YOPB_YEREN
YOPD_YEREN	YOR1_YEAST	YQGH_BACSU	YQGI_BACSU	YQJV_BACSU	YRO2_YEAST	YTP1_YEAST	YWOE_BACSU	YXEN_BACSU	YXX5_CAEEL
YZN4_CAEEL									

(4) 51 lipid-chain-anchored membrane proteins

BLC_CITFR	CHB_VIBHA	CUTF_ECOLI	CYCR_RHOVI	EST2_CAEEL	GLPQ_HAEIN	H81_NEIGO	H82_NEIGO	HBPA_HAEIN	HCY_NATPH
HEL_HAEIN	LPPB_HAEIN	MP17_FRATU	MULI_ECOLI	NLPA_ECOLI	NLPB_ECOLI	OM3B_CHLTR	OM3L_CHLTR	OM3_CHLPS	OMLA_ACTPL
OPUC_BACSU	OSA1_BORBU	OSA2_BORBU	OSA3_BORBU	OSA4_BORBU	OSA5_BORBU	OSA6_BORBU	OSA7_BORBU	OSB1_BORBU	OSB2_BORBU
P37_MYCGE	PAL_ECOLI	PCP_HAEIN	PMEB_ERWCH	PULS_KLEPN	SLP_ECOLI	TCPC_VIBCH	TMPA_TREPA	TRAV_ECOLI	TRT1_ECOLI
TRT2_ECOLI	TRT3_ECOLI	TRT4_ECOLI	VACJ_SHIFL	VM03_BORHE	VM07_BORHE	VM17_BORHE	VM21_BORHE	VM24_BORHE	VM25_BORHE
YSCJ_YERPS									

(5) 110 GPI-anchored membrane proteins

5NTD_BOVIN	ACES_ANOST	AMPM_HELVI	AXO1_CHICK	BCM1_HUMAN	BST1_HUMAN	CADD_CHICK	CAH4_HUMAN	CCEM_HUMAN	CD14_HUMAN
CD24_HUMAN	CD52_HUMAN	CD59_HUMAN	CEPU_CHICK	CGM6_HUMAN	CNTR_HUMAN	CONN_DROME	CONT_CHICK	CSA_DICDI	DAF1_MOUSE
DAF_CAVPO	E48A_HUMAN	EFA1_HUMAN	EFA3_HUMAN	EFA4_HUMAN	FAS1_DROME	FCG3_HUMAN	FOL1_HUMAN	FOL2_HUMAN	FOL3_HUMAN
FS22_DROME	G13A_DICDI	G13B_DICDI	G156_PARPR	G168_PARPR	GAS1_YEAST	GDNR_RAT	GLYP_HUMAN	GP2_CANFA	GP42_RAT
GP46_LEIAM	GP63_LEICH	GP85_TRYCR	GPC2_RAT	GPC3_RAT	GPCK_MOUSE	HYA1_CAVPO	LAMP_RAT	LAZA_SCHAM	LIPL_CAVPO
LY6A_MOUSE	LY6C_MOUSE	MDP1_HUMAN	MKC7_YEAST	NAR3_HUMAN	NARG_HUMAN	NART_MOUSE	NCA2_HUMAN	NCA3_MOUSE	NCA_HUMAN
NRT1_RAT	NRT2_RAT	NTRI_RAT	OMGP_HUMAN	OPCM_RAT	PAG1_TRYBB	PAR1_TRYBB	PARA_TRYBB	PARB_TRYBB	PARC_TRYBB
PARX_TRYBB	PONA_DICDI	PPB1_HUMAN	PPB2_HUMAN	PPB3_HUMAN	PPBE_MOUSE	PPBI_BOVIN	PPBJ_RAT	PPBN_HUMAN	PPBT_BOVIN
PRIO_ATEPA	PRP1_TRAST	PRP2_BOVIN	PSA_DICDI	SP63_STRPU	THY1_HUMAN	THYB_MOUSE	TREA_RABIT	TRFM_HUMAN	UPAR_BOVIN
UROM_BOVIN	VSA1_TRYBB	VSG2_TRYEQ	VSI1_TRYBB	VSI2_TRYBB	VSI3_TRYBB	VSI4_TRYBB	VSI5_TRYBB	VSI6_TRYBB	VSIB_TRYBB
VSM1_TRYBB	VSM2_TRYBB	VSM4_TRYBB	VSM5_TRYBB	VSM6_TRYBB	VSWA_TRYBR	VSWB_TRYBR	VSY1_TRYCO	VSY3_TRYCO	YAP3_YEAST

†As classified under Classification Schemes. Codes are according to the SWISS-PROT data bank.

first term in equation (6) is the squared Mahalanobis distance[10,11] between $\mathbf{X}^{\xi}$ and $\mathbf{X}$, while the second term reflects the difference of covariance matrices for different subsets. Incorporation of the second term into the discriminant function is very important, especially when the subset sizes in the training dataset are much different.[12] It is because of this term that the covariant discriminant function F as defined by equation (6) is no longer a distance because it does not satisfy the condition of $F(\mathbf{X}, \mathbf{X}^{\xi}) = 0$ when $\mathbf{X} \equiv \mathbf{X}^{\xi}$; also, it may have a negative value, obviously in conflict with the classical definition that a distance must satisfy positivity, symmetry, and the triangular inequality.

Thus, the prediction rule is formulated by

$$F(\mathbf{X}, \mathbf{X}^{\chi}) = \mathbf{Min}\left\{F(\mathbf{X}, \mathbf{X}^{1}), \quad F(\mathbf{X}, \mathbf{X}^{2}), F(\mathbf{X}, \mathbf{X}^{3}), \ldots, F(\mathbf{X}, \mathbf{X}^{m})\right\} \quad (9)$$

where χ can be 1, 2, 3, . . . , or m, and the operator **Min** means taking the least one among those in the parentheses; the superscript χ of equation (9) is the predicted type or cellular location for the membrane protein $\mathbf{X}$. If there is a tie, ξ is not uniquely determined, but in practice, a tie is rarely observed for real data.

RESULTS AND DISCUSSION

Predictions were performed for both the membrane protein types and cellular locations. The prediction quality was examined by two approaches. One is based on the self-consistency test, and the other the jackknife test. The former is for testing the self-consistency of a prediction method, while the latter is for testing its extrapolating effectiveness by cross-validation. When the self-consistency test is performed for the current study, the type or location for each of the proteins in a given dataset is predicted using the rules derived from the same dataset, the so-called development dataset or training dataset. According to such an operation, the parameters derived from the training dataset include the information from a protein that is later plugged back into the test. This will certainly give a somewhat optimistic error estimate because of the memorization effect; i.e., the same proteins are used to derive the rule parameters and to test themselves. Nevertheless, this kind of test is absolutely necessary because it reflects the self-consistency of a prediction method, especially for its algorithmic part. A prediction algorithm certainly cannot be deemed a good one if its self-consistency is poor. In other words, the self-consistency test is necessary, but it is not sufficient for evaluating a prediction method. As a complement, a cross-validation examination is needed because it can reflect the extrapolating effectiveness of a prediction method. It is known that the single independent dataset test, subsampling test and jackknife test are the three methods often used for cross-validation. It is also known that, of the three test methods, the most objective and effective is the jackknife test, also

TABLE II. List of 2,105 Protein Sequences Used as Training Data for Predicting Cellular Locations of Membrane Proteins†

(1) 55 chloroplast membrane proteins

ATP1_ARATH	ATP2_ARATH	ATPA_ANTSP	ATPB_AEGCO	ATPD_ANTSP	ATPE_ANTFO	ATPF_ANTSP	ATPG_CHLRE	ATPH_ANTSP	ATPI_ANTSP
ATPX_ANTSP	CAB4_ARATH	CB11_LYCES	CB12_LYCES	CB13_LYCES	CB21_ARATH	CB22_ARATH	CB23_HORVU	CB24_LYCES	CB25_NICPL
CB26_PETSP	CB27_TOBAC	CB28_PEA	CB29_MAIZE	CB2A_PINSY	CB2B_LYCES	CB2G_LYCES	CB2_CHLMO	CB48_MAIZE	DS22_CRAPL
FENR_CHLRE	FENS_ORYSA	HS7E_SPIOL	IN37_SPIOL	L181_CHLEU	NU2C_MAIZE	PLAS_ARATH	PLAT_POPNI	PORI_MAIZE	PSAA_CHLRE
PSAB_ANTMA	PSAE_CHLRE	PSBA_AMAHY	PSBB_CHLRE	PSBD_CHLRE	PSBO_ARATH	PSBP_CHLRE	PSBQ_CHLRE	PSBR_ARATH	PSBT_CHLRE
PSBX_ARATH	PSE1_NICSY	PSE2_NICSY	SECA_PEA	SOT1_SPIOL					

(2) 64 endoplasmic reticulum membrane proteins

ACAT_HUMAN	ALG5_YEAST	ALG8_YEAST	ARE1_YEAST	ARE2_YEAST	ASPH_BOVIN	ATC3_YEAST	ATCE_HUMAN	CALG_MOUSE	CALX_CANFA
CYB5_BOVIN	DHA4_HUMAN	E310_ADE02	E3GL_ADE02	EF11_CRIGR	EM24_YEAST	ER25_HUMAN	ER53_HUMAN	ES22_MOUSE	EST1_MESAU
EST2_RABIT	EST3_RAT	EST4_RAT	EST5_RAT	ESTM_MOUSE	ESTN_MOUSE	FDFT_HUMAN	G25L_CANFA	G6PT_HUMAN	GAA1_YEAST
GCS1_HUMAN	GPT_CRIGR	HMD1_ARATH	HMD2_ARATH	HMDH_BLAGE	IP3R_DROME	IP3S_HUMAN	IRE1_YEAST	LPG1_LEIDO	LYSH_CHICK
MAL_HUMAN	NC5R_BOVIN	NCPR_CANMA	NPL1_YEAST	OST4_CANFA	OSTA_YEAST	OSTB_YEAST	PIGA_HUMAN	PMT1_YEAST	RIB1_HUMAN
RIB2_HUMAN	S61A_CANFA	SC62_YARLI	SC66_YEAST	SHR3_YEAST	SRPB_MOUSE	SRPR_CANFA	SSRA_CANFA	SSRB_CANFA	SSRD_HUMAN
SSRG_RAT	TRAM_CANFA	UBC6_YEAST	VM21_YEAST						

(3) 44 Golgi apparatus membrane proteins

APP1_HUMAN	ATC1_YEAST	BAGT_LYMST	BGAT_HUMAN	BGIB_HUMAN	CAG1_CHICK	CAG2_HUMAN	CAG4_CHICK	CAG6_HUMAN	CAGB_MOUSE
CAGC_HUMAN	ER53_HUMAN	FURL_DROME	FUT1_HUMAN	FUT2_HUMAN	FUT3_BOVIN	FUT4_HUMAN	FUT5_HUMAN	FUT6_HUMAN	FUT7_HUMAN
G6NT_BOVIN	GATR_BOVIN	GDA1_YEAST	GM12_SCHPO	GM13_RAT	GNT1_HUMAN	GNT2_HUMAN	GNT3_HUMAN	GNT5_HUMAN	KEX2_YEAST
KRE2_YEAST	KRE6_CANAL	M121_DROME	M122_DROME	MA12_HUMAN	MAN2_MOUSE	4)galactosylOCH1_YEAST		PAGT_BOVIN	PEP1_YEAST
RGP2_HUMAN	SC14_YEAST	SEC7_YEAST	VP36_CANFA						

(4) 21 lysosome membrane proteins

CD63_HUMAN	LMP1_CHICK	LMP2_CHICK	LYII_HUMAN	MPRD_BOVIN	MPRI_BOVIN	CD63_MOUSE	CD63_RABIT	LMP1_CRIGR	LMP1_HUMAN
LMP1_MOUSE	LMP1_RAT	LMP2_CRIGR	LMP2_HUMAN	LMP2_MOUSE	LMP2_RAT	LYII_RAT	MPRD_HUMAN	MPRD_MOUSE	MPRI_HUMAN
MPRI_MOUSE									

(5) 154 mitochondrial membrane proteins

ACDV_BOVIN	ADT1_BOVIN	ADT2_ARATH	ADT3_BOVIN	ADT_ANOGA	AOFA_BOVIN	AOFB_HUMAN	AOF_ONCMY	AR11_YEAST	ATM1_YEAST
ATP0_BOVIN	ATPA_HUMAN	ATPL_BOVIN	ATPM_BOVIN	ATPN_HUMAN	ATPY_YEAST	C560_BOVIN	CACM_YEAST	CBP3_YEAST	CBP4_YEAST
CCHL_CANAL	COQ1_YEAST	COQ2_YEAST	COX1_ALBCO	COX2_ACHDO	COX4_DICDI	COX5_DICDI	COX6_DICDI	COX7_YEAST	COX9_YEAST
COXA_BOVIN	COXB_BOVIN	COXC_HORVU	COXD_BOVIN	COXE_BOVIN	COXH_BOVIN	COXI_RAT	COXJ_BOVIN	COXK_BOVIN	COXX_YEAST
COXY_YEAST	COXZ_YEAST	CPT1_HUMAN	CPT2_HUMAN	CPTM_HUMAN	CY11_SOLTU	CY1_BOVIN	CYB2_HANAN	CYT2_YEAST	DHSA_BOVIN
DHSB_CHOCR	DHSD_PORPU	DHSX_YEAST	DPS1_YEAST	ETFD_HUMAN	FLX1_YEAST	GATM_HUMAN	HEMZ_HUMAN	HXK1_BOVIN	HXK2_HUMAN
IM17_YEAST	IM22_YEAST	IM23_YEAST	IM44_YEAST	IMP1_YEAST	IMP2_YEAST	KAD2_BOVIN	KCRS_HUMAN	KCRU_CHICK	M2OM_BOVIN
MBA1_YEAST	MCAT_RAT	MCR1_YEAST	MD10_YEAST	MMM1_YEAST	MPCP_BOVIN	MPP1_SOLTU	MRS3_YEAST	MRS4_YEAST	MSP1_YEAST
N4AM_BOVIN	N4BM_BOVIN	NB2M_BOVIN	NB4M_BOVIN	NB5M_BOVIN	NB7M_BOVIN	NB8M_BOVIN	NDI1_YEAST	NDKM_DICDI	NI2M_BOVIN
NI8M_BOVIN	NI9M_BOVIN	NIAM_BOVIN	NIDM_BOVIN	NIGM_BOVIN	NIKM_BOVIN	NIMM_BOVIN	NINM_BOVIN	NIPM_BOVIN	NISM_BOVIN
NU2M_ALBCO	NUBM_ASPNG	NUC1_YEAST	NUCM_BOVIN	NUFM_BOVIN	NUGM_ARATH	NUIM_NEUCR	NUML_BOVIN	NUMM_BOVIN	NUOM_BOVIN
NURM_NEUCR	NUYM_BOVIN	OM06_YEAST	OM07_YEAST	OM20_NEUCR	OM22_NEUCR	OM37_YEAST	OM40_NEUCR	OM45_YEAST	OM70_NEUCR
P18_LEITA	PKBS_HUMAN	POR1_WHEAT	POR2_HUMAN	POR4_SOLTU	POR6_SOLTU	PORI_DICDI	PPOX_HUMAN	PT22_YEAST	PT54_YEAST
PT94_YEAST	PYRD_ARATH	SCO1_YEAST	SDH3_YEAST	SDH4_YEAST	SUOX_CHICK	TXTP_BOVIN	UCP1_HUMAN	UCP2_HUMAN	UCR1_BOVIN
UCR2_BOVIN	UCR3_TOBAC	UCR4_TOBAC	UCR5_TOBAC	UCR6_BOVIN	UCR7_YEAST	UCR9_EUGGR	UCRH_BOVIN	UCRI_BOVIN	UCRQ_BOVIN
UCRX_BOVIN	UCRY_BOVIN	VDHA_CHICK	YM19_WHEAT						

(6) 26 nuclear membrane proteins

BPI_HUMAN	CC48_ARATH	E311_ADE02	GP21_RAT	HN36_HUMAN	LAM1_CHICK	LAM2_CHICK	LAM3_MOUSE	LAM4_XENLA	LBR_CHICK
N121_RAT	NPL1_YEAST	OTE_DROME	P152_YEAST	S160_YEAST	SAD1_SCHPO	SNF1_YEAST	E311_ADE05	HN36_MOUSE	LAM1_HUMAN
LAM1_MOUSE	LAM1_XENLA	LAM2_MOUSE	LAM2_XENLA	LAM3_XENLA	LBR_HUMAN				

(7) 37 peroxisome membrane proteins

P47A_CANBO	P47B_CANBO	PEX2_CRIGR	PEX3_PICAN	PEX5_HUMAN	PEX6_RAT	PEXB_YEAST	PEXC_PICPA	PEXD_PICPA	PEXE_PICAN
PEXG_YARLI	PEXH_YARLI	PMP2_MOUSE	PMP7_HUMAN	PMPA_CANBO	PMPB_CANBO	PXA1_YEAST	PXA2_YEAST	PXBA_CANBO	PXBB_CANBO
PEX2_HUMAN	PEX2_MOUSE	PEX2_PICPA	PEX2_PODAN	PEX2_RAT	PEX3_PICPA	PEX3_YEAST	PEX5_MOUSE	PEX5_PICAN	PEX5_PICPA
PEX5_YARLI	PEX5_YEAST	PEXD_YEAST	PEXE_YEAST	PMP2_RAT	PMP7_MOUSE	PMP7_RAT			

(8) 1680 plasma membrane proteins

5H1A_HUMAN	5H1B_CRIGR	5H1D_CANFA	5H1E_HUMAN	5H1F_HUMAN	5H2A_CRIGR	5H2B_HUMAN	5H2C_HUMAN	5H4_RAT	5H5A_HUMAN
5H5B_MOUSE	5H6_HUMAN	5H7_CAVPO	5HT1_APLCA	5HT2_APLCA	5HT3_HUMAN	5HTA_DROME	5HTB_DROME	5HT_BOMMO	A1AA_HUMAN
A1AB_HUMAN	A1AC_BOVIN	A2AA_CAVPO	A2AB_CAVPO	A2AC_CAVPO	A2AD_HUMAN	A2AR_CARAU	A3_VIGUN	AA1R_BOVIN	AA2A_CANFA
AA2B_HUMAN	AA3R_CANFA	AAAT_MOUSE	AC22_STRCO	ACH1_CAEEL	ACH2_CAEEL	ACH3_BOVIN	ACH4_CAEEL	ACH5_CAEEL	ACH6_CAEEL
ACH7_BOVIN	ACH9_RAT	ACHA_BOVIN	ACHB_BOVIN	ACHD_BOVIN	ACHE_BOVIN	ACHG_BOVIN	ACHN_CHICK	ACHO_CARAU	ACHP_CARAU
ACM1_DROME	ACM2_CHICK	ACM3_BOVIN	ACM4_CHICK	ACM5_HUMAN	ACTR_BOVIN	ADT_RICPR	AFQ2_STRCO	AG22_HUMAN	AG2R_BOVIN
AG2S_HUMAN	ALCP_BACP3	ALKB_PSEOL	ALP1_YEAST	ALST_BACSU	AMSL_ERWAM	AMT_CORGL	APJ_HUMAN	APRD_PSEAE	AQP1_BOVIN
AQP2_HUMAN	AQP3_HUMAN	AQP4_HUMAN	AQP5_HUMAN	AQPA_RANES	AQPL_HUMAN	AQUA_ATRCA	AROP_CORGL	AT7A_MOUSE	AT7B_HUMAN
ATA1_SYNY3	ATC1_DUNBI	ATC3_HUMAN	ATC4_YEAST	ATC5_YEAST	ATC8_YEAST	ATC9_YEAST	ATCF_HUMAN	ATCL_MYCGE	ATCP_HUMAN
ATCQ_HUMAN	ATCR_HUMAN	ATCS_SYNP7	ATCX_SCHPO	ATC_PLAFK	ATHA_CANFA	ATHL_HUMAN	ATKA_ENTFA	ATKB_ENTFA	ATMA_ECOLI
ATMB_SALTY	ATN1_BUFMA	ATN2_CHICK	ATN3_CHICK	ATNA_ANGAN	ATP6_ALBCO	ATR1_YEAST	ATSY_SYNP7	ATU1_YEAST	ATU2_YEAST
ATXA_LEIDO	ATXB_LEIDO	B1AR_CANFA	B2AR_CANFA	B3A2_HUMAN	B3A3_HUMAN	B3AR_BOVIN	B3AT_CHICK	B4AR_MELGA	BAC1_HALS1
BAC2_HALS2	BAC3_HALVA	BACH_HALHP	BACR_HALAR	BACS_HALHA	BACT_HALSA	BENE_ACICA	BETP_CORGL	BFR1_SCHPO	BIB_DROME
BIOX_BACSH	BLR1_HUMAN	BMR1_BACSU	BMR2_BACSU	BMRP_CANAL	BNA1_HUMAN	BOFA_BACSU	BRB1_HUMAN	BRB2_HUMAN	BRNQ_LACDL
BROW_DROME	BRS3_CAVPO	BRS4_BOMOR	C24B_HUMAN	C550_BACSU	C561_HUMAN	C5AR_CANFA	CADA_STAAU	CADD_STAAU	CAKB_BOVIN
CALR_HUMAN	CAMG_HUMAN	CAN1_CANAL	CAR1_DICDI	CAR2_DICDI	CAR3_DICDI	CASR_BOVIN	CB11_RABIT	CB12_RABIT	CB1A_FUGRU
CB1B_FUGRU	CB1R_HUMAN	CB21_RABIT	CB22_RABIT	CB2R_HUMAN	CBIN_SALTY	CBIQ_SALTY	CCKR_CAVPO	CCP1_RAT	CCR3_HUMAN
CCR4_BOVIN	CCT1_RAT	CD20_HUMAN	CD2R_HUMAN	CD47_HUMAN	CD97_HUMAN	CFTR_BOVIN	CGRR_HUMAN	CHAA_ECOLI	CHS2_YEAST
CHS3_YEAST	CIC1_CYPCA	CIC2_HUMAN	CIC5_HUMAN	CICB_RAT	CICC_RABIT	CICG_HUMAN	CICH_TORCA	CICK_HUMAN	CICL_HUMAN
CIK1_DROME	CIK2_DROME	CIK3_HUMAN	CIK4_BOVIN	CIK5_HUMAN	CIK6_HUMAN	CIKA_RAT	CIKB_DROME	CIKD_HUMAN	CIKE_DROME
CIKF_RAT	CIKG_RAT	CIKL_DROME	CIKW_DROME	CIN1_LOLBL	CIN2_RAT	CIN3_RAT	CIN4_HUMAN	CIN5_RAT	CIN6_HUMAN

called the leave-one-out test.[13] In the jackknife test each protein in a given dataset is singled out in turn as a *test protein*, and all the rule parameters are determined from the remaining $N - 1$ proteins. Hence, the memorization effects as included in the self-consistency tests can be completely excluded. During the process of jackknife analysis, both the training and testing datasets are actually open; in turn, a protein will move from each to the other.

TABLE II. (Continued)

CINA_DROME	CITN_KLEPN	CKR1_HUMAN	CKR2_HUMAN	CKR3_HUMAN	CKR4_HUMAN	CKR5_HUMAN	CKR6_HUMAN	CKR7_HUMAN	CKRV_MOUSE
CLC1_HUMAN	CLC2_HUMAN	CLC3_HUMAN	CLC4_HUMAN	CLC5_HUMAN	CLC6_HUMAN	CLC7_RAT	CMLR_STRLI	CMST_CRIGR	CNG1_BOVIN
CNG2_BOVIN	CNG3_BOVIN	CNG4_BOVIN	CNGX_RAT	COA0_HELPY	COA1_HELPY	COA2_HELPY	COA3_HELPY	COMA_STRPN	COMP_BACSU
COX2_BACFI	COX3_BACP3	COX4_BACP3	COXM_BRAJA	COXX_BACFI	CPSD_STRAG	CRF2_HUMAN	CRFR_CHICK	CRNA_EMENI	CSG2_YEAST
CTK1_RABIT	CTPA_MYCLE	CTPB_MYCLE	CTR1_HUMAN	CTR2_HUMAN	CVAB_ECOLI	CX32_ARATH	CX33_MICUN	CX35_RAJER	CX41_XENLA
CX43_BRARE	CX56_CHICK	CXA1_BOVIN	CXA2_XENLA	CXA3_BOVIN	CXA4_HUMAN	CXA5_CANFA	CXA6_CANFA	CXA7_RAT	CXA8_CHICK
CXB1_HUMAN	CXB2_HUMAN	CXB3_MOUSE	CXB4_MOUSE	CXB5_MOUSE	CXB6_MOUSE	CY14_NEUCR	CYA1_BOVIN	CYA2_RAT	CYA3_RAT
CYA4_RAT	CYA5_CANFA	CYA6_CANFA	CYA7_BOVIN	CYA8_HUMAN	CYA9_MOUSE	CYAB_BORPE	CYBH_ALCEU	CYB_SULAC	CYCM_BRAJA
CYHR_CANMA	CYPR_CALVI	D1DR_CARAU	D2D1_XENLA	D2DR_BOVIN	D3DR_CERAE	D4DR_HUMAN	D5DR_FUGRU	DADR_DIDMA	DAGA_ALTHA
DAL4_YEAST	DAL5_YEAST	DBDR_HUMAN	DCDR_XENLA	DCOB_KLEPN	DCOG_KLEPN	DEG1_CAEEL	DEGW_CAEEL	DEGX_CAEEL	DEL1_CAEEL
DIHR_ACHDO	DIP5_YEAST	DIP_ANTMA	DIVJ_CAUCR	DOP1_DROME	DOP2_DROME	DTD_HUMAN	DTPT_LACHE	DUFF_HUMAN	DUR3_YEAST
EAT1_BOVIN	EAT2_HUMAN	EAT3_BOVIN	EAT4_HUMAN	EAT_CAEEL	EDG1_HUMAN	EDG2_BOVIN	EDG3_HUMAN	EMP1_HUMAN	EMP2_HUMAN
EMP3_HUMAN	EMR1_HUMAN	ER21_CAEEL	ER22_CAEEL	ERD1_KLULA	ERD2_ARATH	ERS1_YEAST	ET1R_BOVIN	ET3R_XENLA	ETBR_BOVIN
EXOQ_RHIME	EXOY_RHIME	EXUT_ECOLI	F480_MOUSE	FANA_HELAS	FCEB_HUMAN	FCY2_YEAST	FDNH_ECOLI	FDNI_ECOLI	FDOH_ECOLI
FDOI_ECOLI	FDXH_HAEIN	FET4_YEAST	FEUB_BACSU	FEUC_BACSU	FIXG_RHIME	FIXI_RHIME	FLIP_ECOLI	FLIQ_ECOLI	FLIR_ECOLI
FML1_HUMAN	FML2_HUMAN	FMLR_HUMAN	FRE1_YEAST	FRIZ_DROME	FRP1_SCHPO	FSHR_BOVIN	FTH1_HAEIN	FTH2_SYNY3	FTH3_SYNY3
FTH4_SYNY3	FTSH_BACSU	FUR4_SCHPO	G10D_MOUSE	GAA1_BOVIN	GAA2_BOVIN	GAA3_BOVIN	GAA4_BOVIN	GAA5_HUMAN	GAA6_HUMAN
GAB1_BOVIN	GAB2_HUMAN	GAB3_CHICK	GAB4_CHICK	GABP_BACSU	GAB_DROME	GAC1_RAT	GAC2_BOVIN	GAC3_MOUSE	GAC4_CHICK
GAD_MOUSE	GAL2_YEAST	GALR_HUMAN	GAP1_YEAST	GAR1_HUMAN	GAR2_HUMAN	GAR3_RAT	GAS1_HUMAN	GASR_CANFA	GC96_HUMAN
GCRC_MOUSE	GCRT_CHICK	GCY6_HUMAN	GEF1_YEAST	GHSR_HUMAN	GIPR_HUMAN	GLCP_SYNY3	GLHR_ANTEL	GLPF_BACSU	GLPR_HUMAN
GLPT_BACSU	GLR1_HUMAN	GLR2_HUMAN	GLR3_HUMAN	GLR4_HUMAN	GLR5_HUMAN	GLR6_HUMAN	GLR7_HUMAN	GLRK_CHICK	GLR_HUMAN
GLTP_BACSU	GLTT_BACCA	GMS1_SCHPO	GNP1_YEAST	GNS1_YEAST	GNTP_BACLI	GPCR_LYMST	GPR1_HUMAN	GPR2_HUMAN	GPR3_HUMAN
GPR4_HUMAN	GPR5_HUMAN	GPR6_HUMAN	GPR7_HUMAN	GPR8_HUMAN	GPRA_HUMAN	GPRC_HUMAN	GPRD_HUMAN	GPRE_RAT	GPRF_HUMAN
GPRH_HUMAN	GPRJ_HUMAN	GPRK_HUMAN	GPRL_HUMAN	GPRM_HUMAN	GPRN_HUMAN	GPRO_HUMAN	GRA1_HUMAN	GRA2_BACSU	GRA3_RAT
GRB2_BACSU	GRB_HUMAN	GRFR_HUMAN	GRHR_BOVIN	GRPR_HUMAN	GTR1_BOVIN	GTR2_HUMAN	GTR3_CANFA	GTR4_BOVIN	GTR5_HUMAN
GTR7_RAT	GTRL_DROME	GU27_RAT	GUDT_BACSU	GUSB_BOVIN	H218_RAT	HAK1_SCHOC	HEX6_RICCO	HGT1_KLULA	HH1R_BOVIN
HH2R_CANFA	HIP1_YEAST	HLY2_ECOLI	HLYB_ACTAC	HM74_HUMAN	HMC2_DESVH	HMC3_DESVH	HMC4_DESVH	HMC5_DESVH	HNM1_YEAST
HS30_YEAST	HST6_CANAL	HUP1_CHLKE	HXT1_YEAST	HXT2_YEAST	HXT3_YEAST	HXT4_YEAST	HXT5_YEAST	HXT6_YEAST	HXT7_YEAST
HXTC_YEAST	HXTD_YEAST	HXTE_YEAST	HXTG_YEAST	HYBB_ECOLI	IL8A_GORGO	IL8B_BOVIN	INA1_TRIHA	IRK1_HUMAN	IRK2_CAVPO
IRK3_CHICK	IRK4_HUMAN	IRK5_HUMAN	IRK6_HUMAN	IRK8_HUMAN	IRK9_HUMAN	IRKA_HUMAN	IRKB_HUMAN	IRKC_HUMAN	IRKE_HUMAN
IRKI_HUMAN	IRKX_RAT	ITR1_SCHPO	ITR2_YEAST	KBAA_BACSU	KDGT_BACSU	KHT2_KLULA	KINB_BACSU	KINC_BACSU	LACP_KLULA
LCN3_LACLA	LCNC_LACLA	LMIP_BOVIN	LMRA_STRLN	LPLB_BACSU	LPLC_BACSU	LSHR_BOVIN	LSPA_BACSU	LYP1_YEAST	LYST_CORGL
M6A_HUMAN	M6B_MOUSE	MA3T_YEAST	MA6T_YEAST	MAE1_SCHPO	MALC_STRPN	MALD_STRPN	MAM2_SCHPO	MAP3_SCHPO	MAS_HUMAN
MC3R_HUMAN	MC4R_HUMAN	MC5R_HUMAN	MCBE_ECOLI	MDR1_CAEEL	MDR2_CRIGR	MDR3_CAEEL	MDR4_DROME	MDR5_DROME	MDR_LEITA
ME10_CAEEL	MEC4_CAEBR	MEP1_YEAST	MEP2_YEAST	MEP3_YEAST	MERT_STAAU	MESD_LEUME	MGR1_HUMAN	MGR2_HUMAN	MGR3_HUMAN
MGR4_HUMAN	MGR5_HUMAN	MGR6_RAT	MGR7_HUMAN	MGR8_HUMAN	MIP_BOVIN	ML1A_CHICK	ML1B_HUMAN	ML1C_CHICK	ML1X_HUMAN
MMR_BACSU	MOG_BOVIN	MOT1_CRILO	MOTA_BACSU	MRED_BACSU	MRG_HUMAN	MRP1_HUMAN	MSCL_CLOPE	MSHR_BOVIN	MTRC_METTH
MTRD_METTH	MTRE_METTH	MTR_NEUCR	MYP1_XENLA	MYP2_XENLA	MYPR_BOVIN	NAC1_BOVIN	NAC2_RAT	NAC3_RAT	NAG1_HUMAN
NAG2_HUMAN	NAG3_PIG	NAH1_BOVIN	NAH2_RABIT	NAH3_DIDMA	NAH4_RAT	NAH_SCHPO	NAM1_BOVIN	NANU_RABIT	NAPA_ENTHR
NARI_BACSU	NARK_BACSU	NARV_ECOLI	NASA_BACSU	NASU_RAT	NDHF_BACSU	NHAC_BACFI	NIST_LACLA	NK1R_CAVPO	NK2R_BOVIN
NK3R_HUMAN	NKC1_HUMAN	NKC2_MOUSE	NKCL_MANSE	NMBR_HUMAN	NME1_MOUSE	NME2_MOUSE	NME3_HUMAN	NME4_MOUSE	NMZ1_HUMAN
NPT1_HUMAN	NPT2_HUMAN	NQO7_PARDE	NQO8_PARDE	NQOA_PARDE	NQOB_PARDE	NQOC_PARDE	NQOD_PARDE	NQOE_PARDE	NSR_LACLA
NTBE_CANFA	NTCH_RAT	NTC1_CRIGR	NTCP_HUMAN	NTCR_HUMAN	NTDO_BOVIN	NTG1_HUMAN	NTG2_MOUSE	NTG3_HUMAN	NTGL_BOVIN
NTNO_BOVIN	NTPI_ENTHR	NTPJ_ENTHR	NTPR_RAT	NTR1_HUMAN	NTR2_MOUSE	NTRY_AZOCA	NTS1_RAT	NTS2_RAT	NTSE_DROME
NTT4_RAT	NTT7_RAT	NTTA_CANFA	NU2C_SYNP7	NU5C_SYNP2	NUOA_ECOLI	NUOH_ECOLI	NUOJ_ECOLI	NUOK_ECOLI	NUOL_ECOLI
NUOM_ECOLI	NUON_ECOLI	NUPC_BACSU	NY1R_HUMAN	NY2R_BOVIN	NY4R_HUMAN	NY5R_HUMAN	NY6R_MOUSE	NYR_DROME	OAR_BOMMO
OL1E_HUMAN	OLF0_RAT	OLF1_CANFA	OLF2_CANFA	OLF3_CANFA	OLF4_CANFA	OLF5_CHICK	OLF6_CHICK	OLF7_RAT	OLF8_RAT
OLF9_RAT	OLFD_CANFA	OLFE_HUMAN	OLFI_HUMAN	OLFJ_HUMAN	OPRD_HUMAN	OPRK_CAVPO	OPRM_HUMAN	OPRX_CAVPO	OPS1_CALVI
OPS2_DROME	OPS3_DROME	OPS4_DROME	OPSB_ANOCA	OPSD_ALLMI	OPSG_ASTFA	OPSH_ASTFA	OPSI_ASTFA	OPSP_CHICK	OPSR_ANOCA
OPSU_BRARE	OPSV_CHICK	OPUB_BACSU	OPUD_BACSU	OXYR_HUMAN	P2UR_HUMAN	P2X1_HUMAN	P2X2_RAT	P2X3_RAT	P2X4_HUMAN
P2X5_HUMAN	P2X6_RAT	P2X7_HUMAN	P2Y3_CHICK	P2Y4_HUMAN	P2Y6_HUMAN	P2Y7_HUMAN	P2Y8_XENLA	P2YR_BOVIN	PACR_BOVIN
PAFR_CAVPO	PAR2_HUMAN	PATC_DROME	PBP4_NOCLA	PBUX_BACSU	PD2R_HUMAN	PDR5_YEAST	PDUF_SALTY	PECM_ERWCH	PEDD_PEDAC
PER1_HUMAN	PER2_HUMAN	PER3_BOVIN	PER4_HUMAN	PET1_HUMAN	PET2_HUMAN	PF2R_BOVIN	PGSA_BACSU	PH84_YEAST	PI2R_HUMAN
PIGF_HUMAN	PIP_LACLA	PKBS_BOVIN	PKN6_MYXXA	PLLP_RAT	PM1_HUMAN	PM22_HUMAN	PMA1_AJECA	PMA2_ARATH	PMA3_ARATH
PMA4_NICPL	PMIP_NICAL	PNUC_ECOLI	PPA1_YEAST	PRA1_USTHO	PRA2_USTMA	PRO1_LEIEN	PROW_BACSU	PSAA_ANAVA	PSAB_ANAVA
PSAL_SYNEN	PSN1_HUMAN	PSN2_HUMAN	PSS1_CRILO	PSS_BACSU	PSY_NEUCR	PT2A_ARATH	PT2B_ARATH	PTBA_BACSU	PTFB_RHOCA
PTFC_BACSU	PTFD_BACSU	PTGA_BACSU	PTLB_LACCA	PTMA_BACSU	PTMB_BACST	PTNC_ECOLI	PTND_ECOLI	PTR2_CANAL	PTRR_DIDMA
PTSA_PEDPE	PTSB_BACSU	PUR8_STRLP	PUT4_YEAST	PUTX_EMENI	P_HUMAN	QACA_STAAU	QAY_NEUCR	QOX1_BACSU	QOX2_ACEAC
QOXM_SULAC	QUTD_EMENI	RAFP_PEDPE	RAG1_KLULA	RBSC_BACSU	RCEL_CHLAU	RCEM_CHLAU	RCO3_NEUCR	RDC1_CANFA	RDS_BOVIN
RDXA_RHOSH	RFAL_ECOLI	RFE_ECOLI	RGR_BOVIN	RGT2_YEAST	RHAG_HUMAN	RHCE_HUMAN	RHD_HUMAN	RHLA_PANTR	RHLC_GORGO
RHLD_GORGO	RHLF_PANTR	RHLR_PANTR	RHL_HYLPI	RHOM_DROME	ROCE_BACSU	ROM1_BOVIN	RT1B_ACTPL	RT3B_ACTPL	RTA_RAT
SAT1_RAT	SATT_HUMAN	SCAA_BOVIN	SCAB_HUMAN	SCAD_HUMAN	SCAG_HUMAN	SCRC_HUMAN	SCRT_DROME	SE12_CAEEL	SECY_ANTSP
SENR_RAT	SEO1_YEAST	SLY4_YEAST	SNF3_YEAST	SNQ2_YEAST	SP5E_BACSU	SPAT_BACSU	SPC1_YEAST	SPE4_CAEEL	SSR1_HUMAN
SSR2_BOVIN	SSR3_HUMAN	SSR4_HUMAN	SSR5_HUMAN	STE2_SACKL	STE3_YEAST	STE6_YEAST	STL1_YEAST	STP1_ARATH	STT3_CAEEL
SUL1_YEAST	SUR_CRICR	TA2R_BOVIN	TAP1_HUMAN	TAP2_HUMAN	TAT2_YEAST	TCMA_STRGA	TCR2_BACSU	TCRB_BACSU	TCR_BACST
TERC_ALCSP	TH11_TRYBB	TH12_TRYBB	TH2A_TRYBB	THAS_HUMAN	THI7_YEAST	THIX_YEAST	THIY_YEAST	THRR_CRILO	TIP1_TOBAC
TIP2_TOBAC	TIPA_ARATH	TIPE_DROME	TIPG_ARATH	TIPR_ARATH	TIPW_LYCES	TJ6_MOUSE	TLR1_DROME	TLR2_DROME	TOK1_YEAST
TRA2_CAEBR	TRBA_ECOLI	TRFR_HUMAN	TRK1_SACUV	TRK2_YEAST	TRK_SCHPO	TSAB_RICTS	TSAG_RICTS	TSAK_RICTS	TSAR_RICTS
TSAS_RICTS	TSAT_RICTS	TSAW_RICTS	TSCC_HUMAN	TSHR_BOVIN	TXKR_HUMAN	UAPA_EMENI	UAPC_EMENI	UGA4_YEAST	UGAT_HUMAN
UL33_HCMVA	UN17_CAEEL	UN36_CAEEL	UNC7_CAEEL	UNC8_CAEEL	US27_HCMVA	US28_HCMVA	UT1_HUMAN	UT2_HUMAN	V1AR_HUMAN
V1BR_HUMAN	V2R_BOVIN	VAL1_YEAST	VC03_SPVKA	VG74_HSVSA	VGLB_HSVA1	VIPR_CARAU	VIPS_HUMAN	VK02_SPVKA	VM11_YEAST
VQ3L_CAPVK	VU51_HSV6U	WC1A_ARATH	WC1B_ARATH	WC1C_ARATH	WC2A_ARATH	WC2B_ARATH	WC2C_ARATH	WHIT_ANOAL	Y4JF_RHISN
Y736_HAEIN	YADS_ECOLI	YAF3_YEAST	YBE2_YEAST	YCKJ_BACSU	YD19_METJA	YEP0_YEAST	YG90_HAEIN	YGT6_YEAST	YKH3_CAEEL
YMN2_CAEEL	YNZ3_CAEEL	YOPB_YEREN	YOPD_YEREN	YOR1_YEAST	YQJV_BACSU	YRO2_YEAST	YTP1_YEAST	YWOE_BACSU	YXEN_BACSU
YXX5_CAEEL	YZN4_CAEEL	41BB_HUMAN	A33_HUMAN	A4_DROME	ACET_HUMAN	ACE_HUMAN	AMA1_PLACH	AMFR_HUMAN	ANPA_HUMAN
ANPB_ANGJA	ANPC_BOVIN	APP2_RAT	APX1_CAEEL	AVR2_BOVIN	AVRB_HUMAN	BAS1_CHICK	BFR2_HUMAN	BGP1_HUMAN	BLVR_BOVIN
BUTY_BOVIN	C114_MOUSE	C166_BRARE	C22A_HUMAN	C22B_HUMAN	C79A_BOVIN	C79B_HUMAN	C8B1_HUMAN	C8B2_HUMAN	CAD1_CHICK
CAD2_CHICK	CAD3_HUMAN	CAD4_CHICK	CAD5_MOUSE	CAD6_HUMAN	CAD8_HUMAN	CADB_HUMAN	CADC_HUMAN	CADF_HUMAN	CADL_RAT
CADN_XENLA	CADO_XENLA	CAML_HUMAN	CD11_MOUSE	CD12_MOUSE	CD19_HUMAN	CD1A_HUMAN	CD1B_HUMAN	CD1C_HUMAN	CD1D_HUMAN
CD1E_HUMAN	CD27_HUMAN	CD28_BOVIN	CD2_HORSE	CD30_HUMAN	CD33_HUMAN	CD34_CANFA	CD36_BOVIN	CD3D_HUMAN	CD3E_CANFA
CD3G_HUMAN	CD3H_MOUSE	CD3Z_HUMAN	CD40_HUMAN	CD44_BOVIN	CD45_HUMAN	CD4_CANFA	CD5_BOVIN	CD6_HUMAN	CD7_HUMAN
CD80_HUMAN	CD83_HUMAN	CD86_HUMAN	CD8A_BOVIN	CD8B_MOUSE	CEK2_CHICK	CEK3_CHICK	CGM1_HUMAN	CINB_HUMAN	CINC_RAT
CR1_HUMAN	CR2_HUMAN	CRB_DROME	CRF4_HUMAN	CTL4_HUMAN	CYGD_BOVIN	CYGE_MOUSE	CYGF_HUMAN	CYGR_ARBPU	CYGS_STRPU
CYGX_RAT	CYRB_HUMAN	CYRG_BOVIN	DAF1_CAEEL	DAF4_CAEEL	DCC_HUMAN	DLK_HUMAN	DLL1_MOUSE	DL_DROME	DSC1_BOVIN
DSC2_HUMAN	DSC3_BOVIN	DSG1_BOVIN	DSG3_HUMAN	EDD1_HUMAN	EFB1_HUMAN	EFB2_HUMAN	EFB3_HUMAN	EG15_CAEEL	EGFR_HUMAN
EGF_HUMAN	EGLN_MOUSE	EPA1_HUMAN	EPA2_HUMAN	EPA3_CHICK	EPA4_CHICK	EPA5_CHICK	EPA6_MOUSE	EPA7_HUMAN	EPA8_MOUSE
EPB1_HUMAN	EPB2_CHICK	EPB3_HUMAN	EPB4_HUMAN	EPB5_CHICK	EPOR_HUMAN	ERB2_HUMAN	EV2A_HUMAN	EV2B_HUMAN	FAS2_SCHAM
FAS3_DROME	FASA_BOVIN	FAT_DROME	FCE1_RAT	FCEA_HUMAN	FCEG_CAVPO	FCG0_HUMAN	FCG1_HUMAN	FCG2_BOVIN	FCGA_HUMAN
FCGB_HUMAN	FCGC_HUMAN	FGR1_CHICK	FGR2_DROME	FGR3_HUMAN	FGR4_HUMAN	FLT3_HUMAN	FS21_DROME	G49A_MOUSE	G49B_MOUSE
G731_HUMAN	G732_HUMAN	GARP_HUMAN	GCSR_MOUSE	GHRH_MOUSE	GHR_BOVIN	GLP1_CAEEL	GLPA_HUMAN	GLPB_HUMAN	GLPE_HUMAN
GLP_HORSE	GP10_DICDI	GP38_CANFA	GP70_MOUSE	GPBA_HUMAN	GPBB_HUMAN	GPIX_HUMAN	GPV_HUMAN	GRK_DROME	HEMA_RACVI
HSER_CAVPO	I12R_HUMAN	I131_HUMAN	I132_HUMAN	ICA1_BOVIN	ICA2_HUMAN	ICA3_BOVIN	ICCR_DROME	IDD_HUMAN	IG1R_HUMAN

TABLE II. (Continued)

IL1R_HUMAN	IL1S_HUMAN	IL2A_BOVIN	IL2B_HUMAN	IL3A_MOUSE	IL3B_MOUSE	IL3R_HUMAN	IL4R_HUMAN	IL5R_HUMAN	IL6A_HUMAN
IL6B_HUMAN	IL7R_MOUSE	INGR_HUMAN	INGS_HUMAN	INLA_LISMO	INR1_BOVIN	INR2_HUMAN	INSR_DROME	IRR_CAVPO	ITA1_DROME
ITA2_DROME	ITA3_CRISP	ITA4_HUMAN	ITA5_HUMAN	ITA6_CHICK	ITA8_CHICK	ITA9_HUMAN	ITAB_HUMAN	ITAE_HUMAN	ITAL_HUMAN
ITAM_HUMAN	ITAV_CHICK	ITAX_HUMAN	ITB0_XENLA	ITB1_CHICK	ITB2_BOVIN	ITB3_HUMAN	ITB4_HUMAN	ITB5_HUMAN	ITB6_HUMAN
ITB7_HUMAN	ITB8_HUMAN	ITBX_DROME	KAPP_ARATH	KFMS_FELCA	KIR1_BOVIN	KIR2_HUMAN	KIR3_HUMAN	KIR4_HUMAN	KIR5_HUMAN
KIR6_CHICK	KKIT_BOVIN	KLTK_HUMAN	KPRO_MAIZE	KROS_HUMAN	LAG2_CAEEL	LAG3_HUMAN	LAGC_DICDI	LAR_DROME	LDL1_XENLA
LDL2_XENLA	LDLR_CRIGR	LDVR_CHICK	LEM1_BOVIN	LEM2_BOVIN	LEM3_BOVIN	LEPR_HUMAN	LEUK_HUMAN	LI12_CAEEL	LIN3_CAEEL
LRP1_CHICK	LRP_CAEEL	LT23_CAEEL	LU_HUMAN	LY9_MOUSE	MAGL_MOUSE	MAGS_MOUSE	MAG_HUMAN	MANR_HUMAN	MCP_HUMAN
MEPA_HUMAN	MEPB_HUMAN	MET_HUMAN	MINK_HUMAN	MS2_HUMAN	MU18_HUMAN	MYP0_BOVIN	NCA1_BOVIN	NCA2_RAT	NEU_RAT
NGCA_CHICK	NGFR_CHICK	NK10_HUMAN	NKR0_HUMAN	NKR1_HUMAN	NKR2_HUMAN	NKR3_HUMAN	NKR4_HUMAN	NKR5_HUMAN	NKR6_HUMAN
NKR7_HUMAN	NKR9_HUMAN	NOTC_BRARE	NRCA_CHICK	NRG_DROME	NRP_CHICK	NTC1_MOUSE	NTC4_MOUSE	OX2G_RAT	OX40_HUMAN
PA2R_BOVIN	PCP2_HUMAN	PD1_HUMAN	PEC1_BOVIN	PGDR_HUMAN	PGDS_HUMAN	PGG2_RAT	PRLR_BOVIN	PTP1_DROME	PTP6_DROME
PTP9_DROME	PTPA_HUMAN	PTPB_HUMAN	PTPD_HUMAN	PTPE_HUMAN	PTPF_HUMAN	PTPJ_HUMAN	PTPK_MOUSE	PTPM_HUMAN	PTPN_HUMAN
PTPO_RAT	PTPZ_HUMAN	PVDA_PLAKN	PVDB_PLAKN	PVDG_PLAKN	PVDR_PLAVI	PVR_MOUSE	RAGE_BOVIN	RET_HUMAN	RON_HUMAN
SDC1_CRIGR	SDC2_HUMAN	SDC3_CHICK	SDC4_CHICK	SDC_DROME	SEPL_HUMAN	SERR_DROME	SHAK_DROME	SL17_ENTHI	SPER_STRPU
SPIT_DROME	SRK6_BRAOL	TACT_HUMAN	TF_BOVIN	TGR2_HUMAN	TIE1_BOVIN	TIE2_BOVIN	TMK1_ARATH	TML1_ARATH	TNR1_HUMAN
TNR2_HUMAN	TNRC_HUMAN	TOLL_DROME	TOP_DROME	TOR_DROME	TPOR_HUMAN	TRBM_HUMAN	TRK3_HUMAN	TRKA_HUMAN	TRKB_HUMAN
TRKC_HUMAN	TSA4_GIALA	TYO3_HUMAN	UFO_HUMAN	UPK3_BOVIN	VCA1_HUMAN	VEGR_RAT	VGL2_CVH22	VGLI_HSVEB	VGP_EBOV
VGR1_HUMAN	VGR2_COTJA	VGR3_HUMAN	XMRK_XIPMA	ZAN_PIG	ZIPP_DROME	41BL_HUMAN	4F2_HUMAN	A15_HUMAN	AMPE_HUMAN
AMPN_HUMAN	ATHB_CANFA	ATNB_ANGAN	ATNC_BOVIN	ATND_BUFMA	ATNG_BOVIN	BST2_HUMAN	CD2L_HUMAN	CD37_HUMAN	CD38_HUMAN
CD3L_HUMAN	CD53_HUMAN	CD69_HUMAN	CD72_HUMAN	CD81_HUMAN	CD82_HUMAN	CD94_HUMAN	CD9_BOVIN	CO02_HUMAN	CYAG_DICDI
ECE1_BOVIN	ECE2_BOVIN	FCE2_MOUSE	HEPS_HUMAN	ILT4_HUMAN	IM23_SCHHA	KTR1_YEAST	KTR2_YEAST	KUCR_MOUSE	LECH_HUMAN
LECI_HUMAN	LEPS_BACSU	LY4A_MOUSE	LY4B_MOUSE	LY4D_MOUSE	LY4E_MOUSE	LY4F_MOUSE	LY4G_MOUSE	LYAH_MOUSE	LYTR_BACSU
MANX_MOUSE	MMGL_MOUSE	MNS1_YEAST	MOTB_BACSU	MSRE_BOVIN	NEP_HUMAN	NK11_MOUSE	NK12_MOUSE	NK13_RAT	NK14_MOUSE
NKGA_HUMAN	NKGC_HUMAN	NKGD_HUMAN	NKGE_HUMAN	NRT_DROME	OX4L_HUMAN	PBPB_BACSU	PC1_HUMAN	PSM_HUMAN	SC20_YEAST
SED4_YEAST	SKN1_CANAL	STUB_DROME	TAL6_HUMAN	TLPA_BRAJA	TRSR_HUMAN	UPKA_BOVIN	UPKB_BOVIN	AAS_ECOLI	ACRB_ECOLI
ACRF_ECOLI	ACSA_ACEXY	ANSP_ECOLI	AQPZ_ECOLI	ARAE_ECOLI	ARAH_ECOLI	BACA_RHIME	BCR_ECOLI	BCSA_ACEXY	BETT_ECOLI
BRAB_PSEAE	BRAD_PSEAE	BRAE_PSEAE	BRAZ_PSEAE	CAIT_ECOLI	CARR_MYXXA	CDSA_ECOLI	CIT1_ECOLI	CLD1_ECOLI	CLD2_ECOLI
CLD_SALTY	CMLA_PSEAE	CODB_ECOLI	CPXA_ECOLI	CRED_ECOLI	CSCB_ECOLI	CX1A_PARDE	CX1B_PARDE	CYDA_AZOVI	CYDB_ECOLI
CYOA_ECOLI	CYOB_ECOLI	CYOC_ECOLI	CYOD_ECOLI	CYOE_ECOLI	CYST_SYNP7	DCTA_ECOLI	DCTB_RHILE	DCTS_RHOCA	DHAQ_ACEPO
DHG_ECOLI	DHSC_COXBU	DHSD_COXBU	DMSC_ECOLI	DPPB_ECOLI	DPPC_ECOLI	DSBB_ECOLI	DSBD_ECOLI	EMRB_ECOLI	ENVZ_ECOLI
EXBB_ECOLI	FEOB_ECOLI	FEPD_ECOLI	FEPG_ECOLI	FIXL_AZOCA	FLHA_ECOLI	FRDC_ECOLI	FRDD_ECOLI	FTSW_ECOLI	GALP_ECOLI
GLF_ZYMMO	GLNP_ECOLI	GLTS_ECOLI	GUFA_MYXXA	HISM_ECOLI	HISQ_ECOLI	HMUU_YERPE	HOXN_ALCEU	HYCC_ECOLI	HYCD_ECOLI
HYFC_ECOLI	IMMA_CITFR	IMMB_ECOLI	IPAB_SHIDY	KDGL_ECOLI	KEFB_ECOLI	KEFC_ECOLI	KGTP_ECOLI	KUP_ECOLI	LACF_AGRRD
LACG_AGRRD	LACY_CITFR	LAFT_VIBPA	LCRD_YEREN	LEP3_PSEAE	LGT_ECOLI	LIMA_PSEGL	LIVH_ECOLI	LIVM_ECOLI	LLDP_ECOLI
LNT_ECOLI	LYSP_ECOLI	MALF_ECOLI	MALG_ECOLI	MCP1_ECOLI	MCP2_ECOLI	MCP3_ECOLI	MCP4_ECOLI	MCPC_SALTY	MCPD_ENTAE
MCPS_ENTAE	MDOH_ECOLI	MELB_ECOLI	MEXB_PSEAE	MGLC_ECOLI	MSBB_ECOLI	NANT_ECOLI	NDVB_RHIME	NFRB_ECOLI	NHAA_ECOLI
NHAB_ECOLI	NUPG_ECOLI	OPPB_ECOLI	OPPC_ECOLI	OUSA_ERWCH	PANF_ECOLI	PGPA_ECOLI	PGTB_SALTY	PGTP_SALTY	PHEP_ECOLI
PHOR_ECOLI	PNTA_ECOLI	PNTB_ECOLI	POTE_ECOLI	PROP_ECOLI	PROY_ECOLI	PRTD_ERWCH	PSTA_ECOLI	PSTC_ECOLI	PTAA_ECOLI
PTCC_ECOLI	PTDA_ECOLI	PTGB_ECOLI	PTHB_ECOLI	PTKC_ECOLI	PTOA_ECOLI	PTTB_BACSU	PUTP_ECOLI	RAFB_ECOLI	RFBP_SALTY
RHAT_ECOLI	RODA_ECOLI	SANA_ECOLI	SBMA_ECOLI	SECD_ECOLI	SECE_ECOLI	SECF_ECOLI	SECG_ECOLI	SFUB_SERMA	SHIA_ECOLI
SYRD_PSESY	TCR1_ECOLI	TCR3_ECOLI	TCR4_SALOR	TCR5_ECOLI	TCR7_VIBAN	TCR8_PASMU	TEHA_ECOLI	TNAB_ECOLI	TRBE_ECOLI
TRBI_ECOLI	TRD1_ECOLI	TRD2_ECOLI	TRG1_ECOLI	TRKG_ECOLI	TRKH_ECOLI	TUTB_ERWHE	TYRP_ECOLI	UBIA_ECOLI	UHPB_ECOLI
UHPC_ECOLI	UHPT_ECOLI	URAA_ECOLI	VLYS_LAMBD	XAPB_ECOLI	XYLE_ECOLI	Y4MJ_RHISN	Y889_HELPY	YCEE_ECOLI	YQGH_BACSU
YQGI_BACSU	CAFA_YERPE	FADL_ECOLI	FASD_ECOLI	FHUA_ECOLI	LAMB_ECOLI	NFRA_ECOLI	NMPC_ECOLI	OM11_HAEIN	OM12_HAEIN
OM21_HAEIN	OM22_HAEIN	OM23_HAEIN	OM24_HAEIN	OM25_HAEIN	OM32_COMAC	OM51_HAEIN	OM52_HAEIN	OM53_HAEIN	OM6B_CHLTR
OM6C_CHLTR	OM6E_CHLTR	OM6L_CHLTR	OM6_CHLPN	OMA1_NEIGO	OMA2_NEIME	OMB1_NEIGO	OMB2_NEIGO	OMB3_NEIME	OMB4_NEIME
OMB_NEILA	OMP1_CHLPN	OMP2_CHLPS	OMP3_CHLPS	OMP4_NEIME	OMPA_BORAV	OMPB_CHLTR	OMPC_CHLTR	OMPE_CHLTR	OMPF_CHLTR
OMPH_CHLTR	OMPL_CHLTR	OMPM_CHLTR	OMPN_CHLTR	OMPX_ECOLI	OMP_BORPE	PAGC_SALTY	PAPC_ECOLI	PHOE_CITFR	PORD_PSEAE
PORF_PSEAE	PORI_RHOBL	PORO_PSEAE	PORP_PSEAE	SCRY_KLEPN	TRAN_ECOLI	TSX_ECOLI	VLOM_LAMBD	A180_MOUSE	ADAA_MOUSE
ADAC_MOUSE	ADB1_HUMAN	ADB2_YEAST	ADB_HUMAN	ANK1_HUMAN	ANX2_BOVIN	ANXB_XENLA	ANXD_HUMAN	AP17_HUMAN	AP50_CAEEL
CC48_SOYBN	CC6_YEAST	CD63_RAT	CHAO_DROME	CHS1_CANAL	CHS4_NEUCR	CHSG_ASPFU	CHS_SAPMO	CK13_YEAST	ECTO_RAT
FATP_MOUSE	FIXH_RHIME	FIXS_RHIME	GEM_HUMAN	LSP1_HUMAN	LT60_CAEEL	MUC1_MOUSE	MYSB_DICDI	MYSD_DICDI	NOSY_PSEST
NR13_COTJA	NRL2_ARATH	NRL3_ARATH	NRL4_ARATH	PHY1_CERPU	PMP1_YEAST	PMP2_YEAST	PONA_DICDI	POR1_BOVIN	RAS1_PHYPO
RAS2_PHYPO	RAS3_RHIRA	RASH_MSVHA	RGC1_HUMAN	SECA_CAUCR	SP97_RAT	SPIR_SPICI	TREA_BOMMO	TS11_GIALA	VAT1_BOVIN
(9) 24 vacuolar membrane proteins									
AC45_BOVIN	ATC1_DICDI	ATC2_YEAST	AVP3_ARATH	CVCA_PEA	DAP1_YEAST	DAP2_YEAST	END1_YEAST	GLC5_SOYBN	GLCX_SOYBN
HMT1_SCHPO	IPYR_PHAAU	PEP3_YEAST	PHSA_PHAVU	PHSB_PHAVU	PLM1_PLAFA	VATL_ARATH	VCL1_PEA	VCLA_GOSHI	VCLB_GOSHI
VCLC_PEA	VCL_VICFA	VM13_YEAST	VPH1_YEAST						

[†]Codes are according to the SWISS-PROT data bank.

It is instructive to note that if the samples of proteins are randomly assigned among m possible categories, the rate of correct assignment would generally be $1/m$. For example, the correct rate by random assignment for a classification of five categories would generally be $1/5 = 20\%$, and that for nine categories $1/9 = 11.1\%$. Therefore, the greater the number of categories to be discriminated, the lower the correct rate obtained by random assignment.

Prediction of Membrane Protein Types

The predicted results for the 2,059 membrane proteins in Table I are summarized in Table III. For facilitating comparison, predictions for the same data were also made using the least city-block distance algorithm and least Euclidean distance algorithm, which were proposed for predicting protein structural classes by P.Y. Chou[14,15] and Nakashima et al.,[16] respectively. The corresponding results thus obtained are also given in Table III, from which we can see that the overall rates of correct prediction by the current covariant discriminant algorithm using self-consistency and jackknife tests are (1) 81.1% and 76.4%, respectively, much higher than the completely randomized rate = $1/5 = 20.0\%$, implying that the type of a membrane protein is considerably correlated with its amino acid composition; and (2) more than 17% and 13% higher than those by the simple geometry algorithms.

Moreover, as a demonstration of a practical application, predictions were also performed for 2,625 membrane proteins, which are not included in Table I and hence form an independent testing set. Of the 2,625 proteins that were derived from SWISS PROT (release 35) using the similar criteria discussed under Classification Schemes, Types of Membrane Proteins, 478 are type I transmembrane proteins, 180 type II transmembrane proteins, 1,867 multi-pass transmembrane proteins, 14 lipid-chain anchored membrane proteins, and 86 GPI-anchored membrane proteins. Because of space limitations, the names of these proteins are not given in this report but are available upon

TABLE III. Predicted Results for the Five Types of Membrane Proteins[†]

Test method	Algorithm	Rate of correct prediction for each type of membrane proteins (1) Type I	(2) Type II	(3) Multipass	(4) Lipid chain	(5) GPI- anchored	Overall rate of correct prediction
Self-consistency test[a]	This report[b]	$^{347}/_{435}$ = 79.8%	$^{96}/_{152}$ = 63.2%	$^{1116}/_{1311}$ = 85.1%	$^{43}/_{51}$ = 84.3%	$^{68}/_{110}$ = 61.8%	$^{1670}/_{2059}$ = 81.1%
	City-block distance[c]	$^{244}/_{435}$ = 56.1%	$^{77}/_{152}$ = 50.7%	$^{896}/_{1311}$ = 68.3%	$^{34}/_{51}$ = 66.7%	$^{42}/_{110}$ = 38.2%	$^{1293}/_{2059}$ = 62.8%
	Euclidean distance[d]	$^{280}/_{435}$ = 64.4%	$^{75}/_{152}$ = 49.3%	$^{874}/_{1311}$ = 66.7%	$^{33}/_{51}$ = 64.7%	$^{45}/_{110}$ = 40.9%	$^{1307}/_{2059}$ = 63.5%
Jackknife test[a]	This report[b]	$^{322}/_{435}$ = 74.0%	$^{79}/_{152}$ = 52.0%	$^{1097}/_{1311}$ = 83.7%	$^{25}/_{51}$ = 49.0%	$^{50}/_{110}$ = 45.5%	$^{1573}/_{2059}$ = 76.4%
	City-block distance[c]	$^{241}/_{435}$ = 55.4%	$^{75}/_{152}$ = 49.3%	$^{893}/_{1311}$ = 68.1%	$^{34}/_{51}$ = 66.7%	$^{36}/_{110}$ = 32.7%	$^{1279}/_{2059}$ = 62.1%
	Euclidean distance[d]	$^{277}/_{435}$ = 63.7%	$^{70}/_{152}$ = 46.1%	$^{872}/_{1311}$ = 66.5%	$^{32}/_{51}$ = 62.8%	$^{41}/_{110}$ = 37.3%	$^{1292}/_{2059}$ = 62.8%
Independent dataset test[e]	This report[b]	$^{374}/_{478}$ = 78.2%	$^{113}/_{180}$ = 62.8%	$^{1546}/_{1867}$ = 82.8%	$^{9}/_{14}$ = 64.3%	$^{43}/_{86}$ = 50.0%	$^{2085}/_{2625}$ = 79.4%
	City-block distance[c]	$^{317}/_{478}$ = 66.3%	$^{76}/_{180}$ = 42.2%	$^{1323}/_{1867}$ = 70.9%	$^{13}/_{14}$ = 92.9%	$^{22}/_{86}$ = 25.6%	$^{1751}/_{2625}$ = 66.7%
	Euclidean distance[d]	$^{325}/_{478}$ = 68.0%	$^{85}/_{180}$ = 47.2%	$^{1380}/_{1867}$ = 73.9%	$^{13}/_{14}$ = 78.6%	$^{22}/_{86}$ = 25.6%	$^{1817}/_{2625}$ = 69.2%

[†]As discussed under Classification Schemes, Types of Membrane Proteins.
[a]Conducted by using the 2,059 membrane proteins presented in Table I.
[b]See equation (9), which was also called the covariant discriminant algorithm.
[c]P.Y. Chou.[14,15]
[d]Nakashima et al.[16]
[e]The independent testing dataset contains 2,625 membrane proteins, of which 478 are type I membrane proteins, 180 type II membrane proteins, 1,867 multipass membrane proteins, 14 lipid chain-anchored membrane proteins, and 86 GPI-anchored membrane proteins. None of these proteins occurs in the training dataset of Table I.

TABLE IV. Standard Vector Derived from the Training Dataset of Table I for Each of the Five Membrane Protein Types

Amino acid code	Types of membrane proteins[a] (1) Type 1 transmembrane	(2) Type 2 transmembrane	(3) Multipass transmembrane	(4) Lipid chain-anchored membrane	(5) GPI-anchored membrane
A	0.067	0.073	0.086	0.107	0.088
C	0.027	0.021	0.018	0.016	0.029
D	0.052	0.049	0.036	0.060	0.045
E	0.064	0.056	0.042	0.059	0.059
F	0.035	0.048	0.057	0.026	0.034
G	0.068	0.064	0.073	0.078	0.071
H	0.023	0.026	0.018	0.010	0.021
I	0.048	0.055	0.071	0.045	0.041
K	0.052	0.056	0.042	0.091	0.051
L	0.093	0.102	0.112	0.082	0.097
M	0.019	0.022	0.029	0.022	0.018
N	0.044	0.044	0.037	0.051	0.049
P	0.061	0.048	0.043	0.036	0.056
Q	0.042	0.041	0.031	0.038	0.039
R	0.048	0.048	0.044	0.028	0.039
S	0.080	0.072	0.072	0.075	0.080
T	0.066	0.052	0.057	0.070	0.075
V	0.068	0.067	0.078	0.072	0.062
W	0.015	0.022	0.018	0.008	0.014
Y	0.031	0.033	0.034	0.025	0.030

[a]Components of the standard vector (normalized to 1).

request. The overall rates of correct prediction for the 2,625 membrane proteins in the testing set by various algorithms are also given in Table III, from which one can see that the prediction rate by the covariant discriminant algorithm is 79.4%, while those by the simple geometry distance algorithms are 66.7–69.2%. The former is more than 10% higher than the latter, fully consistent with the conclusion by the above self-consistency and jackknife tests.

To provide an intuitive picture about the difference in amino acid compositions that distinguish types of membrane proteins, the standard vectors derived from the training dataset of Table I for the five membrane protein types are given in Table IV. Furthermore, the 19 positive eigenvalues for each of the five membrane protein types are given in Table V that might be of use for investigating the component-coupled effects at a deeper level, especially

TABLE V. Nineteen Positive Eigenvalues of the Covariance Matrix Derived from the Training Dataset of Table I for Each of the Five Membrane Protein Types

	Type of membrane proteins[a]				
Order	(1) Type 1 transmembrane	(2) Type 2 transmembrane	(3) Multipass transmembrane	(4) Lipid chain-anchored membrane	(5) GPI-anchored membrane
1	0.4	0.6	0.6	0.2	0.4
2	4.5	7.2	6.7	2.1	4.0
3	5.8	8.7	8.7	2.9	4.5
4	7.2	12.3	8.9	5.2	5.8
5	10.8	12.8	10.7	8.3	9.1
6	12.6	16.6	11.4	8.4	12.6
7	13.9	17.3	11.8	10.4	13.4
8	14.8	18.8	14.9	10.9	16.6
9	16.7	26.2	15.5	14.8	17.7
10	18.8	29.0	19.4	20.3	19.1
11	23.9	30.0	20.2	23.4	26.9
12	28.9	41.2	22.1	36.2	36.3
13	31.1	46.7	26.4	45.5	39.4
14	37.0	59.2	33.3	58.9	59.4
15	44.2	65.2	35.4	82.1	78.1
16	64.1	73.9	39.6	98.4	107.2
17	73.8	108.1	77.9	153.9	113.8
18	105.5	110.8	111.0	307.8	176.9
19	143.4	217.7	147.7	578.6	328.4

[a]Eigenvalues $\times 10^5$.

TABLE VI. Standard Vector Derived from the Training Dataset of Table II for Each of the Nine Subcellular Locations of Membrane Proteins

	Subcellular location of membrane proteins[a]								
Amino acid code	(1) Chloroplast	(2) Endoplasmic reticulum	(3) Golgi apparatus	(4) Lysosome	(5) Mitochondrial	(6) Nuclear	(7) Peroxisome	(8) Plasma	(9) Vacuolar
A	0.105	0.069	0.066	0.073	0.082	0.075	0.069	0.080	0.059
C	0.008	0.014	0.016	0.032	0.012	0.011	0.012	0.022	0.010
D	0.042	0.051	0.054	0.043	0.045	0.048	0.045	0.040	0.050
E	0.056	0.061	0.059	0.050	0.056	0.084	0.066	0.048	0.074
F	0.052	0.056	0.049	0.051	0.047	0.035	0.047	0.051	0.054
G	0.089	0.059	0.062	0.069	0.071	0.050	0.061	0.072	0.063
H	0.012	0.023	0.029	0.018	0.023	0.020	0.015	0.019	0.020
I	0.054	0.063	0.046	0.052	0.051	0.049	0.055	0.065	0.061
K	0.055	0.058	0.053	0.050	0.064	0.060	0.066	0.044	0.059
L	0.095	0.104	0.097	0.094	0.099	0.104	0.120	0.108	0.101
M	0.025	0.023	0.023	0.021	0.024	0.022	0.023	0.027	0.016
N	0.042	0.042	0.043	0.061	0.038	0.038	0.054	0.039	0.061
P	0.053	0.049	0.060	0.041	0.053	0.048	0.043	0.047	0.040
Q	0.030	0.036	0.041	0.034	0.036	0.043	0.046	0.034	0.050
R	0.041	0.044	0.057	0.034	0.059	0.064	0.047	0.045	0.049
S	0.074	0.070	0.073	0.079	0.069	0.091	0.076	0.074	0.081
T	0.053	0.055	0.050	0.072	0.054	0.059	0.051	0.059	0.045
V	0.072	0.073	0.063	0.085	0.063	0.059	0.056	0.075	0.068
W	0.015	0.016	0.021	0.009	0.017	0.011	0.015	0.018	0.009
Y	0.028	0.035	0.038	0.033	0.036	0.028	0.034	0.034	0.031

[a]Components of the standard vector (normalized to 1).

for understanding the important contribution from the second term of equation (6).

Prediction of Membrane Protein Locations

Similar predictions were performed for the 2,105 membrane proteins classified under Classification Schemes, Locations of Membrane Proteins as listed in Table II. During the computation process, the nine standard vectors derived from the training dataset of Table II for the nine subcellular locations of membrane proteins and their corresponding nine sets of eigenvalues were automatically generated, and they are given in Tables VI and VII,

TABLE VII. Nineteen Positive Eigenvalues of the Covariance Matrix Derived from the Training Dataset of Table II for Each the Nine Subcellular Locations of Membrane Proteins

	Subcellular location of membrane proteins[a]								
Order	(1) Chloroplast	(2) Endoplasmic reticulum	(3) Golgi apparatus	(4) Lysosome	(5) Mitochondrial	(6) Nuclear	(7) Peroxisome	(8) Plasma	(9) Vacuolar
1	0.3	0.2	0.2	0.0001	0.9	0.03	0.1	0.6	0.007
2	1.6	1.7	1.5	0.005	7.7	0.2	0.4	6.7	0.1
3	2.4	4.1	2.1	0.05	9.2	0.3	1.8	8.0	0.4
4	3.8	6.1	3.2	0.1	11.5	0.4	2.4	9.5	0.6
5	7.1	6.6	3.5	0.2	14.5	1.1	3.1	11.1	1.1
6	9.1	8.9	4.6	0.3	18.9	1.6	4.2	11.5	1.7
7	9.8	10.1	5.8	0.8	25.1	1.9	5.3	15.4	2.6
8	12.3	10.7	7.8	1.2	28.3	4.2	8.2	16.5	3.0
9	17.3	12.2	8.4	1.4	30.1	5.2	12.0	19.7	5.5
10	20.3	16.3	9.4	3.3	32.1	7.9	12.9	20.6	6.2
11	26.3	18.3	11.2	4.0	37.5	11.4	16.2	24.0	7.8
12	30.1	20.3	13.2	4.4	43.7	13.5	20.0	24.9	13.1
13	36.3	25.0	20.7	5.8	50.2	16.1	28.1	26.3	13.5
14	42.5	28.0	22.3	10.1	57.5	17.5	32.6	33.9	14.8
15	54.9	31.6	24.8	11.1	63.1	31.4	37.5	36.7	31.3
16	67.0	46.8	39.1	29.5	76.9	64.0	65.5	45.1	41.8
17	79.7	56.6	50.6	51.1	87.1	109.0	97.2	81.1	73.2
18	182.5	79.7	87.3	117.8	160.2	144.0	156.0	110.8	173.4
19	222.1	134.9	157.6	165.9	167.1	391.4	213.6	174.4	387.7

[a]Eigenvalue $\times 10^5$.

TABLE VIII. Self-consistency Test Results for the 2,105 Membrane Proteins Classified into Nine Cellular Locations According to Fig. 4 as Listed in Table II

	Rate of correct prediction for each location				
Algorithms	(1) Chloroplast	(2) Endoplasmic	(3) Golgi	(4) Lysosome	(5) Mitochondria
This report, eq. (9)	$^{49}/_{55} = 74.5\%$	$^{56}/_{64} = 87.5\%$	$^{44}/_{44} = 100\%$	$^{21}/_{21} = 100\%$	$^{108}/_{154} = 70.1\%$
City-block distance[a]	$^{35}/_{55} = 63.6\%$	$^{18}/_{64} = 28.1\%$	$^{22}/_{44} = 50.0\%$	$^{13}/_{21} = 61.9\%$	$^{35}/_{154} = 22.7\%$
Euclidean distance[b]	$^{36}/_{55} = 65.4\%$	$^{21}/_{64} = 32.8\%$	$^{23}/_{44} = 52.3\%$	$^{16}/_{21} = 76.2\%$	$^{44}/_{154} = 28.6\%$

Rate of correct prediction for each location				
(6) Nucleus	(7) Peroxisome	(8) Plasma membrane	(9) Vacuole	Overall rate of correct prediction
$^{26}/_{26} = 100\%$	$^{34}/_{37} = 91.9\%$	$^{1207}/_{1680} = 71.9\%$	$^{24}/_{24} = 100\%$	$^{1569}/_{2105} = 74.5\%$
$^{12}/_{26} = 46.2\%$	$^{16}/_{37} = 43.2\%$	$^{643}/_{1680} = 38.3\%$	$^{14}/_{24} = 58.3\%$	$^{808}/_{2105} = 38.4\%$
$^{11}/_{26} = 42.3\%$	$^{19}/_{37} = 51.4\%$	$^{683}/_{1680} = 40.7\%$	$^{13}/_{24} = 54.2\%$	$^{866}/_{2105} = 41.1\%$

[a]P.Y. Chou.[14,15]
[b]Nakashima et al.[16]

respectively. Also, predictions were conducted for 2,698 membrane proteins which are not included in Table II and hence form an independent testing set. Of the 2,698 proteins derived from SWISS PROT (release 35) using the similar criteria given under Classification Schemes, Locations of Membrane Proteins, 293 are chloroplast membrane proteins, 79 endoplasmic membrane proteins, 35 Golgi membrane proteins, 433 mitochondria membrane proteins, 1,841 plasma membrane proteins, and 17 vacuolar membrane proteins. The names of these proteins are not given in this article but are available upon request. The predicted results thus obtained are summarized in Tables VIII and IX, from which we can see that the overall rate of correct prediction by the current algorithm for self-consistency test is 74.5%, and those for jackknife and independent dataset tests are 65.9% and 67.1%, respectively. All these rates are significantly higher than the

TABLE IX. Overall Rates of Correct Prediction by Jackknife Test and an Independent Dataset Test for the Nine Cellular Locations

	Overall rate of correct prediction	
Algorithms	Jackknife test[a]	Independent dataset test[b]
This report, eq. (9)	$^{1388}/_{2105} = 65.9\%$	$^{1810}/_{2698} = 67.1\%$
City-block distance[c]	$^{791}/_{2105} = 37.6\%$	$^{992}/_{2698} = 36.8\%$
Euclidean distance[d]	$^{791}/_{2105} = 37.6\%$	$^{991}/_{2698} = 36.7\%$

[a]The jackknife analysis was conducted for the 2,105 membrane proteins in Table II.
[b]The independent testing dataset contains 2,698 membrane proteins, of which 293 are chloroplast membrane proteins, 79 endoplasmic membrane proteins, 35 Golgi membrane proteins, 433 mitochondria membrane proteins, 1,841 plasma membrane proteins, and 17 vacuolar membrane proteins. None of these proteins occurs in the training dataset of Table II.
[c]P.Y. Chou.[14,15]
[d]Nakashima et al.[16]

TABLE X. List of 641 Protein Sequences Used as Training Data for Predicting the Attributes of Membrane Proteins[†]

440 inner membrane proteins

3O1D_COMTE	AAS_ECOLI	ACDV_BOVIN	ACRB_ECOLI	ACRF_ECOLI	ACSA_ACEXY	ADT1_BOVIN	ADT2_ARATH	ADT3_BOVIN	ADT_ANOGA
AMPE_ECOLI	AMPN_ECOLI	ANSP_ECOLI	AQPZ_ECOLI	AR11_YEAST	ARAE_ECOLI	ARAG_ECOLI	ARAH_ECOLI	ARM_MUSDO	AROP_ECOLI
ATKA_ECOLI	ATKB_ECOLI	ATKC_ECOLI	ATM1_YEAST	ATP0_BOVIN	ATP6_ECOLI	ATPA_HUMAN	BACA_RHIME	BCR_ECOLI	BCSA_ACEXY
BETT_ECOLI	BRAB_PSEAE	BRAD_PSEAE	BRAE_PSEAE	BRAZ_PSEAE	BRNQ_ECOLI	BTUD_ECOLI	C560_BOVIN	CACM_YEAST	CADB_ECOLI
CAIT_ECOLI	CARR_MYXXA	CBP4_YEAST	CBPA_SYNP7	CBRD_ERWCH	CCHL_CANAL	CDH_ECOLI	CDSA_ECOLI	CIT1_ECOLI	CLD1_ECOLI
CLD2_ECOLI	CLD_SALTY	CME1_BACSU	CMLA_PSEAE	CODB_ECOLI	COQ1_YEAST	COQ2_YEAST	COX1_ALBCO	COX2_ACHDO	COX3_PARDE
COX4_DICDI	COX5_DICDI	COX6_DICDI	COX7_YEAST	COX9_YEAST	COXA_BOVIN	COXB_BOVIN	COXC_HORVU	COXD_BOVIN	COXE_BOVIN
COXH_BOVIN	COXI_RAT	COXJ_BOVIN	COXK_BOVIN	COXZ_YEAST	CPT2_HUMAN	CPXA_ECOLI	CRED_ECOLI	CSCB_ECOLI	CX1A_PARDE
CX1B_PARDE	CY1_NEUCR	CYDA_AZOVI	CYDB_ECOLI	CYOA_ECOLI	CYOB_ECOLI	CYOC_ECOLI	CYOD_ECOLI	CYOE_ECOLI	CYSA_SYNP7
CYST_SYNP7	CYT2_YEAST	DACA_ECOLI	DACC_ECOLI	DADA_ECOLI	DCTA_ECOLI	DCTB_RHILE	DCTS_RHOCA	DHAQ_ACEPO	DHG_ECOLI
DHM1_METEX	DHSA_BOVIN	DHSB_CHOCR	DHSC_COXBU	DHSD_COXBU	DHSX_YEAST	DLD_ECOLI	DMSC_ECOLI	DPPB_ECOLI	DPPC_ECOLI
DPS1_YEAST	DSBB_ECOLI	DSBD_ECOLI	EMRA_ECOLI	EMRB_ECOLI	ENVZ_ECOLI	EST2_CAEEL	ETFD_HUMAN	EXBB_ECOLI	EXBD_ECOLI
FABI_ECOLI	FECE_ECOLI	FEOB_ECOLI	FEPD_ECOLI	FEPG_ECOLI	FHUB_ECOLI	FHUC_ECOLI	FIXL_AZOCA	FLHA_ECOLI	FLIG_ECOLI
FLIM_CAUCR	FLIN_CAUCR	FLX1_YEAST	FRDC_ECOLI	FRDD_ECOLI	FTSH_ECOLI	FTSL_ECOLI	FTSN_ECOLI	FTSQ_ECOLI	FTSW_ECOLI
FTSY_ECOLI	FTSZ_ECOLI	GABP_ECOLI	GALP_ECOLI	GATM_HUMAN	GEM_HUMAN	GLF_ZYMMO	GLNP_ECOLI	GLNQ_ECOLI	GLPF_ECOLI
GLPT_ECOLI	GLTP_ECOLI	GLTS_ECOLI	GLU2_MAIZE	GSPC_KLEPN	GUFA_MYXXA	HBPA_HAEIN	HEMZ_HUMAN	HISM_ECOLI	HISP_ECOLI
HISQ_ECOLI	HMUU_YERPE	HOXN_ALCEU	HYCC_ECOLI	HYCD_ECOLI	HYFC_ECOLI	IM17_YEAST	IM22_YEAST	IM23_YEAST	IM30_PEA
IM44_YEAST	IMMA_CITFR	IMMB_ECOLI	IMP1_YEAST	IMP2_YEAST	IN37_SPIOL	IPAB_SHIDY	KCRS_HUMAN	KCRU_CHICK	KDGL_ECOLI
KDGT_ECOLI	KDPF_ECOLI	KEFB_ECOLI	KEFC_ECOLI	KGTP_ECOLI	KST1_ECOLI	KST5_ECOLI	KUP_ECOLI	LACF_AGRRD	LACG_AGRRD
LACK_AGRRD	LACY_CITFR	LAFT_VIBPA	LAFU_VIBPA	LAM1_CHICK	LAM2_CHICK	LAM3_MOUSE	LAM4_XENLA	LBR_CHICK	LCRD_YEREN
LEP3_PSEAE	LEP_ECOLI	LGT_ECOLI	LHA1_RHOAC	LHA2_ECTHL	LHA3_RHOAC	LHA4_RHOAC	LHA5_RHOPA	LHA6_RHOAC	LHA7_RHOAC
LHA_CHLAU	LHB1_RHOAC	LHB2_ECTHL	LHB3_RHOAC	LHB4_RHOAC	LHB5_RHOAC	LHB6_RHOAC	LHB7_RHOAC	LHB_CHLAU	LIGE_PSEPA
LIMA_PSEGL	LIVH_ECOLI	LIVM_ECOLI	LLDP_ECOLI	LNT_ECOLI	LSPA_ECOLI	LYSP_ECOLI	M2OM_BOVIN	MALF_ECOLI	MALG_ECOLI
MALK_ECOLI	MBA1_YEAST	MCAT_RAT	MCP1_ECOLI	MCP2_ECOLI	MCP3_ECOLI	MCP4_ECOLI	MCPC_SALTY	MCPD_ENTAE	MCPS_ENTAE
MDOH_ECOLI	MELB_ECOLI	MERT_PSEAE	MEXB_PSEAE	MGLC_ECOLI	MIND_ECOLI	MODC_ECOLI	MOTA_ECOLI	MOTB_ECOLI	MOTX_VIBPA
MPCP_BOVIN	MPP1_SOLTU	MRS3_YEAST	MRS4_YEAST	MSBB_ECOLI	MSCL_ECOLI	MTR_ECOLI	MURG_ECOLI	N4AM_BOVIN	N4BM_BOVIN
NANT_ECOLI	NB2M_BOVIN	NB4M_BOVIN	NB5M_BOVIN	NB7M_BOVIN	NB8M_BOVIN	NDI1_YEAST	NDVB_RHIME	NFRB_ECOLI	NHAA_ECOLI
NHAB_ECOLI	NI2M_BOVIN	NI8M_BOVIN	NI9M_BOVIN	NIAM_BOVIN	NIDM_BOVIN	NIGM_BOVIN	NIKM_BOVIN	NIMM_BOVIN	NINM_BOVIN
NIPM_BOVIN	NISM_BOVIN	NLPA_ECOLI	NPL1_YEAST	NRTA_SYNP7	NU2M_ALBCO	NUBM_ASPNG	NUC1_YEAST	NUCM_BOVIN	NUFM_BOVIN
NUGM_ARATH	NUIM_NEUCR	NUML_BOVIN	NUMM_BOVIN	NUOM_BOVIN	NUPC_ECOLI	NUPG_ECOLI	NURM_NEUCR	NUYM_BOVIN	OAR_MYXXA
OPPB_ECOLI	OPPC_ECOLI	OPPD_ECOLI	OPPF_ECOLI	OTE_DROME	OUSA_ERWCH	P18_LEITA	PANF_ECOLI	PBP2_ECOLI	PBP3_ECOLI
PBPA_ECOLI	PBPB_ECOLI	PGPA_ECOLI	PGTB_SALTY	PGTP_SALTY	PHEP_ECOLI	PHOR_ECOLI	PLAS_ARATH	PLAT_POPNI	PLDB_ECOLI
PLSC_ECOLI	PNTA_ECOLI	PNTB_ECOLI	POTA_ECOLI	POTE_ECOLI	PPOX_HUMAN	PROP_ECOLI	PROV_ECOLI	PROW_ECOLI	PROY_ECOLI
PRTD_ERWCH	PRTE_ERWCH	PSBU_PHOLA	PSPA_ECOLI	PSPB_ECOLI	PSTA_ECOLI	PSTB_ECOLI	PSTC_ECOLI	PT22_YEAST	PT54_YEAST
PT94_YEAST	PTAA_ECOLI	PTBA_ECOLI	PTCC_ECOLI	PTDA_ECOLI	PTFB_ECOLI	PTGB_ECOLI	PTHB_ECOLI	PTKC_ECOLI	PTMA_ECOLI
PTOA_ECOLI	PTSB_KLEPN	PTTB_BACSU	PURT_PASHA	PUTP_ECOLI	PYRD_ARATH	RAFB_ECOLI	RAS1_PHYPO	RAS2_PHYPO	RASH_MSVHA
RBSA_ECOLI	RBSC_ECOLI	RFBP_SALTY	RHAT_ECOLI	RODA_ECOLI	SANA_ECOLI	SBMA_ECOLI	SCO1_YEAST	SDH3_YEAST	SDH4_YEAST
SECD_ECOLI	SECE_ECOLI	SECF_ECOLI	SECG_ECOLI	SECY_CHLTR	SFUB_SERMA	SFUC_SERMA	SHIA_ECOLI	STP_SPIOL	SULA_ECOLI
SYDP_ECOLI	SYRD_PSESY	TCPI_VIBCH	TCR1_ECOLI	TCR2_ECOLI	TCR3_ECOLI	TCR4_SALOR	TCR5_ECOLI	TCR7_VIBAN	TCR8_PASMU
TEHA_ECOLI	TNAB_ECOLI	TOLA_ECOLI	TOLQ_ECOLI	TOXR_PSEAE	TRAC_ECOLI	TRBE_ECOLI	TRBI_ECOLI	TRD1_ECOLI	TRD2_ECOLI
TRG1_ECOLI	TRKG_ECOLI	TRKH_ECOLI	TRM8_ECOLI	TUTB_ERWHE	TXTP_BOVIN	TYRP_ECOLI	UBIA_ECOLI	UCP1_HUMAN	UCP2_HUMAN
UCR1_BOVIN	UCR2_BOVIN	UCR3_TOBAC	UCR4_TOBAC	UCR5_TOBAC	UCR6_BOVIN	UCR7_YEAST	UCR9_EUGGR	UCRH_BOVIN	UCRI_BOVIN
UCRQ_BOVIN	UCRX_BOVIN	UCRY_BOVIN	UGPC_ECOLI	UHPB_ECOLI	UHPC_ECOLI	UHPT_ECOLI	URAA_ECOLI	VDHA_CHICK	VLYS_LAMBD
VR2A_BPT4	VR2B_BPT4	XAPB_ECOLI	XYLE_ECOLI	Y4MJ_RHISN	Y889_HELPY	YCEE_ECOLI	YQGH_BACSU	YQGI_BACSU	ZIPA_ECOLI

201 outer membrane proteins

AG43_ECOLI	AIDA_ECOLI	AIL_YEREN	ALGE_PSEAE	ALKL_PSEOL	AOFA_BOVIN	AOFB_HUMAN	AOF_ONCMY	APRF_PSEAE	BLC_CITFR
BTUB_ECOLI	CACM_YEAST	CAFA_YERPE	CHB_VIBHA	CIRA_ECOLI	COLY_YERPE	CPT1_HUMAN	CPTM_HUMAN	CTR1_NEIME	CTR2_NEIME
CUTF_ECOLI	DACB_BACSU	FADL_ECOLI	FASD_ECOLI	FATA_VIBAN	FCT_ERWCH	FCUA_YEREN	FECA_ECOLI	FEPA_ECOLI	FHUA_ECOLI
FHUE_ECOLI	FOXA_YEREN	FPTA_PSEAE	FPVA_PSEAE	FYUA_YEREN	GCA2_BOVIN	GLPQ_HAEIN	GSPD_KLEPN	H81_NEIGO	H82_NEIGO
HEL_HAEIN	HEMR_YEREN	HLYA_PROMI	HLYB_PROMI	HMUR_YERPE	HRPH_PSESY	HXB2_HAEIN	HXC2_HAEIN	HXK1_BOVIN	HXK2_HUMAN
ICEN_ERWHE	INVA_YEREN	IRGA_VIBCH	IROA_NEIME	IUTA_ECOLI	KCRS_HUMAN	KCRU_CHICK	LAMB_ECOLI	LASA_PSEAE	LPPB_HAEIN
MCR1_YEAST	MD10_YEAST	MIP_CHLTR	MMM1_YEAST	MP17_FRATU	MSP1_YEAST	MULI_ECOLI	MXID_SHIFL	NFRA_ECOLI	NLPB_ECOLI
NMPC_ECOLI	OM06_YEAST	OM07_YEAST	OM11_HAEIN	OM12_HAEIN	OM1C_CHLTR	OM1E_CHLTR	OM1L_CHLTR	OM20_NEUCR	OM21_HAEIN
OM22_HAEIN	OM23_HAEIN	OM24_HAEIN	OM25_HAEIN	OM32_COMAC	OM37_YEAST	OM3A_RHILV	OM3B_CHLTR	OM3L_CHLTR	OM3_CHLPS
OM40_NEUCR	OM45_YEAST	OM51_HAEIN	OM52_HAEIN	OM53_HAEIN	OM6B_CHLTR	OM6C_CHLTR	OM6E_CHLTR	OM6L_CHLTR	OM6_CHLPN
OM70_NEUCR	OMA1_NEIGO	OMA2_NEIME	OMB1_NEIGO	OMB2_NEIGO	OMB3_NEIME	OMB4_NEIME	OMB_NEILA	OMC9_CHLTR	OMC_NEIGO
OMLA_ACTPL	OMP1_CHLPN	OMP2_CHLPS	OMP3_CHLPS	OMP4_NEIME	OMPA_BORAV	OMPB_CHLTR	OMPC_CHLTR	OMPE_CHLTR	OMPF_CHLTR
OMPH_CHLTR	OMPL_CHLTR	OMPM_CHLTR	OMPN_CHLTR	OMPP_ECOLI	OMPT_ECOLI	OMPX_ECOLI	OMP_BORPE	OPR1_NEIME	OPR3_NEIME
OSA1_BORBU	OSA2_BORBU	OSA3_BORBU	OSA4_BORBU	OSA5_BORBU	OSA6_BORBU	OSA7_BORBU	OSB1_BORBU	OSB2_BORBU	PA1_ECOLI
PAGC_SALTY	PAL_ECOLI	PAPC_ECOLI	PBUA_PSESP	PCP_HAEIN	PERT_BORBR	PEXE_PICAN	PFEA_PSEAE	PGTE_SALTY	PHOE_CITFR
PMEB_ERWCH	POR1_WHEAT	POR2_HUMAN	POR4_SOLTU	POR6_SOLTU	PORD_PSEAE	PORF_PSEAE	PORI_DICDI	PORO_PSEAE	PORP_PSEAE
PRTF_ERWCH	PULS_KLEPN	PUPA_PSEPU	PUPB_PSEPU	SCRY_KLEPN	SLAP_AERSA	SLP_ECOLI	SSA1_PASHA	TB11_NEIME	TB12_NEIME
TBP1_NEIGO	TMPA_TREPH	TMPB_TREPA	TOLC_ECOLI	TRAN_ECOLI	TRAV_ECOLI	TRJ1_ECOLI	TRJ2_ECOLI	TRJ3_ECOLI	TRJ4_ECOLI
TRJ5_ECOLI	TRJ9_ECOLI	TRL1_ECOLI	TRT1_ECOLI	TRT2_ECOLI	TRT3_ECOLI	TRT4_ECOLI	TSX_ECOLI	VACJ_SHIFL	VIUA_VIBCH
VLOM_LAMBD	VM03_BORHE	VM07_BORHE	VM17_BORHE	VM21_BORHE	VM24_BORHE	VM25_BORHE	YADA_YEREN	YOPE_YEREN	YOPN_YEREN
YSCJ_YERPS									

[†]Codes are according to the SWISS-PROT data bank.

completely randomized rate, 1/9 = 11.1%, indicating that the cellular location of a membrane protein is also correlated with its amino acid composition. It can also be seen from Tables VIII and IX that the overall rates of correct prediction obtained by the covariant discriminant algorithm are about 28–33% higher than those by the simple geometry algorithms, indicating a significant improvement in both self-consistency and extrapolating effectiveness by taking into account the component-coupled effects, fully consistent with the case of protein structural class prediction.[7,12] Moreover, the prediction quality can be further improved if one can (1) narrow down the scope of subcellular location for a query protein according to its source and other relevant information (e.g., if a query protein was from an animal organism, one could exclude the chloroplast and vacuole subsets from consideration and perform the prediction among seven possible subcellular locations instead of nine; the corresponding rates thus obtained would be 76.7%, 70.2%, and 70.6% for the self-consistency, jackknife, and independent dataset tests, respectively); and (2) improve the training data of small subsets by adding into them more new proteins that have been found belonging to the locations defined by these subsets.

Finally, the covariant discriminant prediction algorithm was applied to a different classification level, where the membrane proteins are discriminated between the following two attributes as defined in release 35.0 of SWISS-PROT[4]: (1) inner membrane proteins and (2) outer membrane proteins. The prediction between the inner and

TABLE XI. Standard Vectors and Positive Eigenvalues Derived from Table X for Inner and Outer Membrane Proteins

(1) Inner membrane proteins		(2) Outer membrane proteins	
Standard vector[a]	Eigenvalue[b] ($\times 10^5$)	Standard vector[a]	Eigenvalue[b] ($\times 10^5$)
0.098	0.7	0.094	0.8
0.009	6.9	0.011	4.6
0.039	10.5	0.058	6.5
0.048	11.3	0.051	8.6
0.049	14.7	0.037	10.4
0.078	17.5	0.084	12.9
0.020	20.3	0.014	16.0
0.065	20.9	0.046	18.8
0.045	27.8	0.065	20.4
0.114	28.2	0.083	25.8
0.031	34.9	0.019	32.6
0.034	36.7	0.054	35.6
0.046	39.6	0.035	41.4
0.036	48.8	0.041	45.1
0.051	52.3	0.042	60.1
0.062	64.5	0.080	82.4
0.053	89.9	0.068	108.0
0.075	108.6	0.068	143.5
0.019	235.8	0.012	224.0
0.029	—	0.039	—

[a]See Table IV or VI for the order of vector components.
[b]See Table V or VII for the order of eigenvalues.

TABLE XII. Predicted Results for the Classification Scheme in Which the Proteins Are Discriminated Between Inner and Outer Membrane Proteins

	Rate of correct prediction for each class		Overall rate of correct prediction
Test methods	(1) Inner membrane	(2) Outer membrane	
Self-consistency test[a]	$^{409}/_{440} = 92.6\%$	$^{178}/_{201} = 88.6\%$	$^{587}/_{641} = 91.6\%$
Jackknife test[a]	$^{401}/_{440} = 91.1\%$	$^{162}/_{201} = 80.6\%$	$^{563}/_{641} = 87.8\%$
Independent test[b]	$^{544}/_{587} = 92.7\%$	$^{75}/_{92} = 81.5\%$	$^{619}/_{679} = 91.2\%$

[a]Conducted by using the 641 membrane proteins in Table X.
[b]Conducted by using an independent dataset of 679 membrane proteins, none of which occurs in Table X.

outer membrane proteins is also important because the lipid compositions of the inner and outer monolayers are different, reflecting the different functions of the two faces of a cell membrane. By following the same screening procedures as described under Classification Schemes, we obtained a dataset of 641 protein sequences, of which 440 are inner membrane proteins and 201 outer membrane proteins. The names of the 641 membrane proteins are given in Table X, from which the standard vectors and the eigenvalue sets for the inner and outer membrane proteins were derived, as given in Table XI. The rates of correct prediction using the current algorithm for the 641 proteins by resubstitution and jackknifing are given in Table XII, from which we can see that the overall rate of correct prediction by self-consistency test is 91.6% and that by jackknifing is 87.8%. Predictions were also performed for 679 membrane proteins that were classified according the same scheme but are not included in Table X and hence formed an independent testing set. Of the 679 proteins, 587 are inner membrane proteins and 92 outer membrane proteins. For the space-limited reason, the 679 independent proteins are not given in this report but are available upon request. The predicted results thus obtained are also given in Table XII, from which we can see that the overall rate of correct prediction for such an independent testing dataset is 91.2%, fully consistent with the jackknife-analyzed result.

CONCLUSION

The types of membrane proteins as well as their cellular locations and other attributes are, to a considerable degree, predictable on the basis of their amino acid composition. Compared with the simple geometry distance algorithms in which the composition of each of the 20 amino acids is treated as an independent variable, the rates of correct prediction by using the current covariant discriminant algorithm are significantly higher. The component-coupled effect is a kind of collective interaction, as formulated by a set of covariance matrices in equation (7), $\mathbf{C}_\xi$ ($\xi = 1, 2, \ldots, m$), which are the core of the current algorithm. It is through each of these matrices that a more reasonable statistical distance,[7,8] the Mahananobis distance [the first term of equation (6)], in the amino acid composition space is defined, and it is through the eigenvalues of these matrices [the second term of equation (6)] that the coupling effects in different subsets as well as their sizes are reflected. The covariant discriminant algorithm can also be used in distinguishing membrane proteins from other proteins, as will be discussed in another report.[17]

ACKNOWLEDGMENTS

Valuable discussions with Professor Ferenc J. Kezdy, Dr. Reqiang Yan, and Dr. Jinhe Li are gratefully acknowledged. The authors are indebted to Dr. Viv Junker for interpreting the annotations in Swiss Protein Data Bank and to Raymond B. Moeller, Cynthia A. Ludlow, and Diane M. Ulrich for drawing Figures 1–5. The authors also thank the anonymous referees for their constructive comments in strengthening the presentation of this work.

REFERENCES

1. Rost B, Casadio R, Fariselli P, Sander C. Transmembrane helices predicted at 95% accuracy. Protein Sci 1995;4:521–533.
2. Resh MD. Myristylation and palmitylation of Src family members: the fats of the matter. Cell 1994;76:411–413.
3. Casey PJ. Protein lipidation in cell signalling. Science 1995;268: 221–225.
4. Bairoch A, Apweiler R. "The SWISS-PROT protein sequence data bank and its supplement TrEMBL. Nucleic Acids Res 1997;25:31–36.
5. Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD. In: Molecular biology of the cell. 3rd ed. New York: Garland Publishing. 1994.

6. Lodish H, Baltimore D, Berk A, Zipursky SL, Matsudaira P, Darnell J. (1995) in Molecular Cell Biology, 3rd ed., New York: Scientific American Books.
7. Chou KC. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins Structure Function Genet 1995;21:319–344.
8. Chou KC, Zhang CT. "Prediction of Protein Structural Classes." Crit Rev Biochem Mol Biol 1995;30:275–349.
9. Johnson RA, Wichem DW. Applied multivariate statistical analysis. 3rd ed., Prentice-Hall, Englewood Cliffs, New Jersey.
10. Mahalanobis PC. On the generalized distance in statistics. Proc Natl Inst Sci India 1936;2:49–55.
11. Pillai KCS. (1985) Mahalanobis D^2. Encyclopedia of Statistical Sciences, ed. Kotz, S. Johnson, NL., Vol. 5, pp. 176–181, John Wiley & Sons, New York.
12. Chou KC, Liu W, Maggiora GM, Zhang CT. Prediction and classification of domain structure classes. Protein Structure Function Genet 1998;31:97–103.
13. Mardia KV, Kent JT, Bibby JM. Multivariate analysis. p. 322 and p. 381, Academic Press, London, 1979.
14. Chou PY. Amino acid composition of four classes of proteins. In: Abstracts of Papers, Part I, Second Chemical Congress of the North American Continent, Las Vegas. 1980.
15. Chou PY. Prediction of protein structural classes from amino acid composition. in Prediction of protein structure and the principles of protein conformation. Fasman, G. D., p 549–586. New York: Plenum Press. 1989.
16. Nakashima H, Nishikawa K, Ooi T. The folding type of a protein is relevant to the amino acid composition. J Biochem 1986;99:152–162.
17. Chou KC, Elrod DW. Protein subcellular location prediction. Protein Engineering, 1999, in press.