Protein Engineering vol.12 no.2 pp.107–118, 1999

Protein subcellular location prediction

Kuo-Chen Chou[1] **and David W.Elrod**

Computer-Aided Drug Discovery, Pharmacia & Upjohn, Kalamazoo,
MI 49007-4940, USA

[1]To whom correspondence should be addressed.
E-mail: kuo-chen.chou@am.pnu.com

The function of a protein is closely correlated with its subcellular location. With the rapid increase in new protein sequences entering into data banks, we are confronted with a challenge: is it possible to utilize a bioinformatic approach to help expedite the determination of protein subcellular locations? To explore this problem, proteins were classified, according to their subcellular locations, into the following 12 groups: (1) chloroplast, (2) cytoplasm, (3) cytoskeleton, (4) endoplasmic reticulum, (5) extracell, (6) Golgi apparatus, (7) lysosome, (8) mitochondria, (9) nucleus, (10) peroxisome, (11) plasma membrane and (12) vacuole. Based on the classification scheme that has covered almost all the organelles and subcellular compartments in an animal or plant cell, a covariant discriminant algorithm was proposed to predict the subcellular location of a query protein according to its amino acid composition. Results obtained through self-consistency, jackknife and independent dataset tests indicated that the rates of correct prediction by the current algorithm are significantly higher than those by the existing methods. It is anticipated that the classification scheme and concept and also the prediction algorithm can expedite the functionality determination of new proteins, which can also be of use in the prioritization of genes and proteins identified by genomic efforts as potential molecular targets for drug design.

Keywords: amino acid composition/bioinformatics/covariant discriminant/organelles/subcellular compartments

Introduction

Given the sequence of a protein, how can its cellular location and biological function be determined? This is a problem vitally important to both cell biologists and bioinformatists today. Since the number of sequences entering into data banks has been rapidly increasing, it is time consuming and costly to approach this problem entirely by performing various locational and functional experimental tests. For example, in the recent release 35.0 (November 1997) of SWISS-PROT (Bairoch and Apweiler, 1997), the number of sequence entries has reached 69 113, which represents an increase of 17.10% over release 34.0 (October 1996). In view of this, it is highly desirable to develop an algorithm for rapidly predicting the subcellular compartments in which a new protein sequence could be located.

In a pioneering study, Nakashima and Nishikawa (1994) proposed an algorithm to discriminate between intracellular and extracellular proteins by amino acid composition and residue-pair frequencies. In their method, the training set consisted of 894 proteins, of which 649 were intracellular and 245 extracellular; the testing set consisted of 379 proteins, of which 225 were intracellular and 154 extracellular. Recently, Cedano *et al*. (1997) extended the discriminative classes from two to five, i.e. extracellular, integral membrane, anchored membrane, intracellular and nuclear. This represents remarkable progress in this area. Furthermore, in an attempt to improve the prediction quality of protein cellular location, they proposed an algorithm called ProtLock. The idea of predicting the cellular location of a protein according to its amino acid composition alone, as done in ProtLock, is actually stimulated by the encouraging results of structural class prediction, where the only input is also the amino acid composition (see, e.g., P.Y.Chou, 1980, 1989; Nakashima *et al*., 1986; K.C.Chou, 1995; Chou and Zhang, 1995). An analysis in an attempt to understand the correlation of the structural class and subcellular location of a protein with its amino acid composition was recently given by Bahar *et al*. (1997) and Andrade *et al*. (1998), respectively.

Approaching the problem in a different way, Nakai and Kanehisa (1992) and Claros *et al*. (1997) proposed to predict the cellular location of proteins based on their N-terminal sorting signals. Obviously, these algorithms rely strongly on the existence of leader sequences. However, as pointed out recently by Reinhardt and Hubbard (1998), ‘In large genome analysis projects genes are usually automatically assigned and these assignments are often unreliable for the 5′-regions’. ‘This can lead to leader sequences being missing or only partially included, thereby causing problems for prediction algorithms depending on them’. Therefore, a method based on the amino acid composition would be more useful in practical applications.

As stated in the paper by Cedano *et al*. (1997), the ProtLock algorithm is mainly based on the procedure reported by Chou and Zhang (1995) for the prediction of protein structural classes according to Mahalanobis distances. Since the least Mahalanobis distance algorithm (K.C.Chou, 1995; Chou and Zhang, 1995) is valid only when the training subset sizes are the same or approximately the same or poor predictions will otherwise result (Chou *et al*., 1998; Chou and Maggiora, 1988), in the ProtLock algorithm the training set for each class was chosen to contain the same number of proteins. However, as shown later, when the cellular protein classification is conducted at a deeper level, it is found that proteins located in some organelles are much more abundant in the SWISS-PROT databank than in others. Besides, for a real cell the number of cellular locations is much greater than five considered by Cedano *et al*. (1997). For example, the number of proteins described as being located in a nucleus is much greater than that in a lysosome, and the number of proteins in cytoplasm is much greater than that in a Golgi apparatus. In view of this, can we develop an algorithm to predict effectively the locations of proteins in cells at a much more discriminative level? The current study was initiated in an attempt to solve this problem.

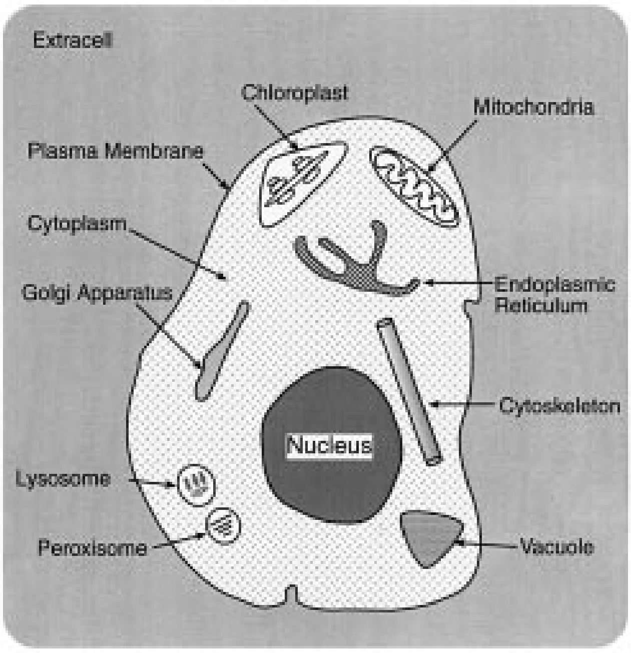

Fig. 1. Schematic diagram showing the subcellular locations of proteins. For simplification, indices 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12 are used to represent chloroplast, cytoplasm, cytoskeleton, endoplasmic reticulum, extracell, Golgi, lysosome, mitochondria, nucleus, peroxisome, plasma membrane and vacuole, respectively. Note that the vacuole and chloroplast proteins exist only in a plant cell.

Location classification

According to their subcellular locations, proteins are classified into the following 12 discriminative groups: (1) chloroplast, (2) cytoplasm, (3) cytoskeleton, (4) endoplasmic reticulum, (5) extracell, (6) Golgi apparatus, (7) lysosome, (8) mitochondria, (9) nucleus, (10) peroxisome, (11) plasma membrane and (12) vacuole (Figure 1). Such a classification covers almost all the organelles in an animal or plant cell (see, e.g., Alberts *et al.*, 1994; Lodish *et al.*, 1995). Note that the vacuole and chloroplast exist only in a plant cell. Membrane proteins such as transmembrane and anchored-membrane proteins actually reflect the protein types rather than subcellular locations. For example, a membrane protein can be associated with the membrane of endoplasmic reticulum, Golgi apparatus, lysosome or any other organelle enveloped by a lipid bilayer structure. Therefore, if associated with endoplasmic reticulum, the membrane protein is located at the endoplasmic reticulum; if associated with the Golgi apparatus, it is located at the Golgi apparatus; and so forth. Plasma membrane proteins are located at the cell envelope (Figure 1).

The classification was based on release 35.0 of SWISS-PROT (Bairoch and Apweiler, 1997). In order to obtain a high-quality, well defined training set, the data were screened strictly according to the following procedures:

1. Included are only those sequences with clear locational descriptions; those with ambiguous or uncertain words such as ‘location unspecified’, ‘probable’, ‘potential’ and ‘by similarity’ were omitted.

2. Sequences annotated by two or more locations are not included because of a lack of uniqueness. For example, a protein sequence labeled with ‘Golgi and nuclear’ or ‘chloroplast or mitochondria’ was omitted. Also note that secreted proteins should be assigned to the extracellular group and proteins annotated with ‘microtubule’ or ‘filament’ should be assigned to the cytoskeletal group (Alberts *et al.*, 1994).

3. For protein sequences with the same name but from different species, only one of them was included. After the above screening procedures we obtained a dataset, S^{12}, of 12 categories that contains 2319 protein sequences, of which 154 are chloroplast proteins, 592 cytoplasmic, 37 cytoskeletal, 53 endoplasmic reticulum, 230 extracellular, 26 Golgi apparatus, 38 lysosomal, 86 mitochondrial, 288 nuclear, 32 peroxisomal, 758 plasma membrane and 25 vacuoles (column 2 of Table I).

4. In order to observe the impact of the number of subcellular locations considered on the prediction rate, two more datasets were constructed. These two datasets are S^7 and S^5 (columns 4 and 6 of Table I, respectively), which were obtained by simply removing the small subsets from S^{12}. The datasets S^7 was derived from S^{12} by removing the cytoskeleton, Golgi apparatus, lysosome, peroxisome and vacuole subsets, none of which contains more than 50 proteins in S^{12}. The dataset S^5 was derived from S^7 by further removing endoplasmic reticulum and mitochondrial subsets, none of which contains more than 100 proteins in S^{12}.

5. In order to test the consistency, three corresponding independent datasets were constructed. They are $\overline{S}^{12}$, $\overline{S}^7$ and $\overline{S}^5$ (columns 3, 5 and 7 of Table I, respectively), none of which contains a protein that occurs in the datasets S^{12}, S^7 and S^5.

Table I. Breakdown of the datasets used in this study

Cellular location	Dataset[a]					
	S^{12}	$\overline{S}^{12}$	S^7	$\overline{S}^7$	S^5	$\overline{S}^5$
(1) Chloroplast	154	119	154	119	154	119
(2) Cytoplasm	592	786	592	786	592	786
(3) Cytoskeleton	37	19	–	–	–	–
(4) Endoplasmic reticulum	53	108	53	108	–	–
(5) Extracell	230	101	230	101	230	101
(6) Golgi apparatus	26	4	–	–	–	–
(7) Lysosome	38	31	–	–	–	–
(8) Mitochondria	86	165	86	165	–	–
(9) Nucleus	288	431	288	431	288	431
(10) Peroxisome	32	24	–	–	–	–
(11) Plasma membrane	758	803	758	803	758	803
(12) Vacuole	25	0	–	–	–	–
Total proteins	2319	2591	2161	2513	2022	2240

[a] The datasets were extracted from release 35.0 of SWISS-PROT (Bairoch and Apweiler, 1997). Dataset S^{12} was obtained by following procedures 1–3 as described in Location classification. Datasets S^7 and S^5 were derived from S^{12}. Datasets $\overline{S}^{12}$, $\overline{S}^7$ and $\overline{S}^5$ are the three independent datasets, none of which contains a protein that occurs in the datasets S^{12}, S^7 and S^5, respectively, as described in Location classification, point 5.

For the convenience of further study or practical application, the names of the 2319 proteins in S^{12} are listed in Appendix A, from which the datasets S^7 and S^5 can also be easily obtained. In this study, the datasets S^{12}, S^7 and S^5 were used as the training datasets to predict the subcellular location of a protein among the 12, seven and five categories of classification, respectively. Owing to limitations on space, the protein names in the datasets $\overline{S}^{12}$, $\overline{S}^7$ and $\overline{S}^5$ are not given here, but they are available upon request.

Prediction algorithm

For brevity, let us use indices 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12 to represent chloroplast, cytoplasm, cytoskeleton, endoplasmic reticulum, extracell, Golgi apparatus, lysosome, mitochondria, nucleus, peroxisome, plasma membrane and vacuole, respectively. We use G_1 to represent the chloroplast subset consisting of only chloroplast proteins, G_2 to represent the cytoplasm subset consisting of only cytoplasmic proteins, and so forth.

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

$$S = G_1 \cup G_2 \cup G_3 \cup G_4 \cup \ldots \cup G_m \tag{1}$$

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 G_ξ. Obviously, $N = \sum_{\xi=1}^{m} n_\xi$. For example, for the dataset in Appendix A, we have $m = 12$, $n_1 = 154$, $n_2 = 592$, . . ., $n_{11} = 758$, $n_{12} = 25$ and $N = 2319$.

The prediction algorithm is established based on the correlation between the subcellular location of a protein and its amino acid composition. Suppose the 20 amino acids are ordered alphabetically according to their single-letter codes: A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W and Y. Thus, any protein in S will correspond to a vector or a point in the 20-D (dimensional) space, i.e. it can be described by (K.C.Chou, 1995)

$$\mathbf{X}^\xi_k = \begin{bmatrix} x^\xi_{k,1} \\ x^\xi_{k,2} \\ \vdots \\ x^\xi_{k,20} \end{bmatrix}, (k = 1, 2, \ldots, n_\xi;\ \xi = 1, 2, 3, \ldots, m) \tag{2}$$

where $x^\xi_{k,1}$, $x^\xi_{k,2}$, . . ., $x^\xi_{k,20}$ are the normalized occurrence frequencies of the 20 amino acids in the kth protein $\mathbf{X}^\xi_k$ of the subset G_ξ. The *standard vector* for the subset G_ξ is defined by

$$\overline{\mathbf{X}}^\xi = \begin{bmatrix} \bar{x}^\xi_1 \\ \bar{x}^\xi_2 \\ \vdots \\ \bar{x}^\xi_{20} \end{bmatrix}, (\xi = 1, 2, 3, \ldots, m) \tag{3}$$

where

$$\bar{x}^\xi_i = \frac{1}{n_\xi} \sum_{k=1}^{n_\xi} x^\xi_{k,i}\,, (i = 1, 2, \ldots, 20). \tag{4}$$

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

The similarity between the standard vector $\mathbf{X}^\xi$ and the protein $\mathbf{X}$ is characterized by the covariant discriminant, as defined by Liu and Chou (1998):

$$F(\mathbf{X}, \overline{\mathbf{X}}^\xi) = D^2(\mathbf{X}, \overline{\mathbf{X}}^\xi) + \ln(\lambda^\xi_2\, \lambda^\xi_3\, \lambda^\xi_4 \ldots \lambda^\xi_{20}) \tag{5}$$

where the first term is the squared Mahalanobis distance between $\overline{\mathbf{X}}^\xi$ and $\mathbf{X}$ (Mahalanobis, 1936; Pillai, 1985; K.C.Chou, 1995):

$$D^2(\mathbf{X}, \overline{\mathbf{X}}^\xi) = (\mathbf{X} - \overline{\mathbf{X}}^\xi)^{\mathbf{T}} \mathbf{C}^{-1}_\xi (\mathbf{X} - \overline{\mathbf{X}}^\xi),\ (\xi = 1, 2, 3, \ldots, m) \tag{6}$$

where $\mathbf{C}_\xi$ is the covariance matrix for subset G_ξ, given by

$$\mathbf{C}_\xi = \begin{bmatrix} c^\xi_{1,1} & c^\xi_{1,2} & \cdots & c^\xi_{1,20} \\ c^\xi_{2,1} & c^\xi_{2,2} & \cdots & c^\xi_{2,20} \\ \vdots & \vdots & \ddots & \vdots \\ c^\xi_{20,1} & c^\xi_{20,2} & \cdots & c^\xi_{20,20} \end{bmatrix} \tag{7}$$

the superscript $\mathbf{T}$ is the transposition operator and $\mathbf{C}^{-1}_\xi$ is the inverse matrix of $\mathbf{C}_\xi$. The matrix elements of $c^\xi_{i,j}$ in Equation 7 are given by

$$c^\xi_{i,j} = \frac{1}{n_\xi - 1} \sum_{k=1}^{n_\xi} [x^\xi_{k,i} - x^\xi_i]\,[x^\xi_{k,j} - x^\xi_j], (i, j = 1, 2, \ldots, 19). \tag{8}$$

Because the amino acid composition must be normalized, i.e. constrained by

$$\sum_{i=1}^{20} x^\xi_{k,i} = 1,\ (k = 1, 2, \ldots, N_\xi;\ \xi = 1, 2, 3, \ldots, m), \tag{9}$$

we have (cf. Equation 8)

$$\begin{cases} \sum_{j=1}^{20} c^\xi_{i,j} = 0,\ (i = 1, 2, \ldots, 20) \\ \sum_{i=1}^{20} c^\xi_{i,j} = 0,\ (j = 1, 2, \ldots, 20) \end{cases} \tag{10}$$

Therefore, $\mathbf{C}_\xi$ defined by Equation 8 is a singular matrix, and its inverse matrix $\mathbf{C}^{-1}_\xi$ must be of divergence and meaninglessness. To overcome such a difficulty, one way is to reduce the amino acid composition space from 20-D to 19-D by removing any one of its 20 components, as described by K.C.Chou (1995). Another way is to use an eigenvalue–eigenvector approach to calculate the Mahalanobis distance so as to avoid dealing with any inverse matrix. According to the eigenvalue–eigenvector approach (Chou and Zhang, 1995), Equation 6 can be written as

$$D^2(\mathbf{X}, \overline{\mathbf{X}}^\xi) = \sum_{i=2}^{20} \frac{1}{\lambda^\xi_i} \left[\sum_{j=1}^{20} (x_j - \bar{x}^\xi_j)\psi^\xi_{i,j} \right]^2 \tag{11}$$

where λ^ξ_i, the eigenvalue, and $\psi^\xi_{i,j}$, the jth component of the eigenvector $\mathbf{\Psi}^\xi_i$, are given by the following equation:

$$\mathbf{C}_\xi\, \mathbf{\Psi}^\xi_i = \lambda^\xi_i\, \mathbf{\Psi}^\xi_i = \lambda^\xi_i \begin{bmatrix} \psi^\xi_{i,1} \\ \psi^\xi_{i,2} \\ \vdots \\ \psi^\xi_{i,20} \end{bmatrix} \quad (i = 1, 2, \ldots, 20) \tag{12}$$

The second term of Equation 5 reflects the difference of covariance matrices for different subsets, in which λ^ξ_i is the ith eigenvalue of the covariance matrix $\mathbf{C}_\xi$ ($i = 2, 3, 4, \ldots, 20$), as defined by Equation 12. It can be proved (Appendix B) that for the covariance matrix $\mathbf{C}_\xi$ as defined by Equation 8, there is no negative eigenvalue. Actually, owing to Equation 10, $\mathbf{C}_\xi$ must have one eigenvalue, denoted by λ^ξ_1, equalto zero (Chou and Zhang, 1995); all the other 19 eigenvalues λ^ξ_2, λ^ξ_3, . . ., λ^ξ_{20} are generally greater than zero. Incorporation of the term ln ($\lambda^\xi_2\, \lambda^\xi_3\, \lambda^\xi_4 \ldots \lambda^\xi_{20}$) into

Table II. Self-consistency test results for the 2319 proteins in Appendix A

Methods	Rate of correct prediction for each subcellular location			
	(1) Chloroplast	(2) Cytoplasm	(3) Cytoskeleton	(4) Endoplasmic ret.
This paper (eq.13)	$\frac{114}{154} = 74.0\%$	$\frac{447}{592} = 75.5\%$	$\frac{33}{37} = 89.2\%$	$\frac{42}{53} = 79.3\%$
ProtLock (Cedano et al., 1997)	$\frac{66}{154} = 42.9\%$	$\frac{182}{592} = 30.7\%$	$\frac{15}{37} = 40.5\%$	$\frac{27}{53} = 50.9\%$

Rate of correct prediction for each subcellular location				
(5) Extracellular	(6) Golgi	(7) Lysosomal	(8) Mitochondrial	(9) Nuclear
$\frac{159}{230} = 69.1\%$	$\frac{26}{26} = 100\%$	$\frac{38}{38} = 100\%$	$\frac{68}{86} = 79.1\%$	$\frac{222}{288} = 77.1\%$
$\frac{65}{230} = 28.3\%$	$\frac{13}{26} = 50.0\%$	$\frac{24}{38} = 63.2\%$	$\frac{45}{86} = 52.3\%$	$\frac{156}{288} = 54.2\%$

Rate of correct prediction for each subcellular location			**Overall rate** of correct prediction
(10) Peroxisomal	(11) Plasma membrane	(12) Vacuole	
$\frac{32}{32} = 100\%$	$\frac{647}{758} = 85.4\%$	$\frac{24}{25} = 96.0\%$	$\frac{1852}{2319} = \mathbf{79.9\%}$
$\frac{11}{32} = 34.4\%$	$\frac{453}{758} = 59.8\%$	$\frac{8}{25} = 32.0\%$	$\frac{1065}{2319} = 45.9\%$

the discriminant function is important, especially when the subset sizes in the training dataset are much different (Chou *et al.*, 1998). It is due to the second term that the covariant discriminant F as defined by Equation 5 is no longer a distance because it does not satisfy the condition of $F(\mathbf{X},\overline{\mathbf{X}}^{\xi}) = 0$ when $\mathbf{X} \equiv \overline{\mathbf{X}}^{\xi}$, and 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. Accordingly, the prediction rule is formulated by

$$F(\mathbf{X}, \overline{\mathbf{X}}^{\lambda}) = \mathbf{Min}\{F(\mathbf{X}, \overline{\mathbf{X}}^{1}), F(\mathbf{X}, \overline{\mathbf{X}}^{2}), F(\mathbf{X}, \overline{\mathbf{X}}^{3}), \ldots, F(\mathbf{X}, \overline{\mathbf{X}}^{m})] \qquad (13)$$

where λ can be 1, 2, 3, . . ., m, and the operator **Min** means taking the least one among those in the parentheses and the superscript λ is the subcellular location predicted for the protein $\mathbf{X}$. If there is a tie case, λ is not uniquely determined, but that did not occur in our datasets.

The eigenvalue–eigenvector approach and the 19-D space approach should give the same results. It is instructive to point out that, if using the 19-D space approach, the covariant discriminant value as defined by Equation 5 will be the same regardless of which one of the 20 amino acid components is left out for constructing a 19-D space. This can be elucidated as follows. The covariant discriminant of Equation 5 consists of two terms. The first term is the squared Mahalanobis distance and its invariability has already been proved by a theorem given by K.C.Chou (1995). The second term is a logarithm, and its argument is actually equal to the determinant value of the matrix obtained by deleting the 20th row and 20th column from the matrix $\mathbf{C}_{\xi}$. As shown by Equation A17 of K.C.Chou (1995), such a determinant value would remain the same regardless of which row and column were removed from $\mathbf{C}_{\xi}$ as long as the removed row and column were the same in order. This indicates the invariability of the second term, and hence also the invariability of the covariant discriminant of Equation 5.

Table III. Overall rates of correct prediction by self-consistency, jackknife and independent dataset tests

Algorithm	Self-consistency test Dataset[a] S^{12}	 S^{7}	 S^{5}
This paper (eq.13)	$\frac{1852}{2319} = 79.9\%$	$\frac{1728}{2161} = 80.0\%$	$\frac{1680}{2022} = 83.1\%$
ProtLock (Cedano et al., 1997)	$\frac{1065}{2319} = 45.9\%$	$\frac{1233}{2161} = 57.1\%$	$\frac{1423}{2022} = 70.4\%$

Algorithm	Jackknife test Dataset[a] S^{12}	 S^{7}	 S^{5}
This paper (eq.13)	$\frac{1586}{2319} = 68.4\%$	$\frac{1579}{2161} = 73.1\%$	$\frac{1584}{2022} = 78.3\%$
ProtLock (Cedano et al., 1997)	$\frac{1017}{2319} = 43.9\%$	$\frac{1201}{2161} = 55.6\%$	$\frac{1405}{2022} = 69.5\%$

Algorithm	Independent-dataset test[b] Dataset[a] $\overline{S}^{12}$	 $\overline{S}^{7}$	 $\overline{S}^{5}$
This paper (eq.13)	$\frac{1966}{2591} = 75.9\%$	$\frac{1948}{2513} = 77.5\%$	$\frac{1833}{2240} = 81.8\%$
ProtLock (Cedano et al., 1997)	$\frac{1036}{2591} = 40.0\%$	$\frac{1275}{2513} = 50.7\%$	$\frac{1528}{2240} = 68.2\%$

[a] See Table I.
[b] The subcellular locations of proteins in the independent testing datasets $\overline{S}^{12}$, $\overline{S}^{7}$ and $\overline{S}^{5}$ were predicted using the rule parameters derived from the training datasets S^{12}, S^{7} and S^{5}, respectively. The same protein did not occur in both training and testing datasets.

Results and discussion

The prediction quality was examined by two test methods, the self-consistency test and the jackknife test. In the self-consistency test, the subcellular location for each of the proteins in a given dataset was predicted using the rules derived from the same dataset, the so-called development dataset or training dataset. In the jackknife test, each protein in the training dataset was singled out in turn as a 'test protein' and all the rule parameters were determined from the remaining $N - 1$ proteins. Jackknife tests are thought one of the most effective and objective methods for cross-validation in statistics (Mardia *et al.*, 1979).

Listed in Table II are the self-consistency test results for discriminating the 12 subcellular locations of proteins in the dataset S^{12} (Appendix A) by using the covariant discriminant algorithm (Equation 13) and ProtLock algorithm (Cedano *et al.*, 1997), respectively. For a detailed prediction process by the current algorithm, see Appendix C, where the covariant discriminant values calculated according to Equation 5 for the 37 proteins in the cytoskeleton subset and their predicted results are given as a demonstration. As can be seen from Table II, the overall rate of correct prediction by the current algorithm is 30% higher than that by the ProtLock algorithm (Cedano *et al.*, 1997). Similar calculations were also carried out for the dataset S^{7} and S^{5}. Furthermore, a jackknife test by the current algorithm and the ProtLock algorithm was performed for each of these three datasets. The results obtained are summarized in Table III, from which the following can be observed.

1. The overall rates of correct prediction obtained by the

current algorithm using the jackknife and self-consistency tests for dataset S^{12} were 68.4 and 79.9%, respectively. Imagine: if the samples of proteins are completely randomly assigned among m possible subsets, the rate of correct assignment would generally be $1/m$; if the random assignment is weighted according to the sizes of subsets, then the rate of correct prediction would be $p_1^2 + p_2^2 + p_3^2 + \ldots + p_m^2$, where $p_i = n_i \Big/ \sum_{\xi}^{m} n_\xi = n_i/N$ (see Equation 1 and the relevant text). Hence the correct rate by a completely random assignment for a classification of 12 categories would be $1/12 \approx 8.3\%$, and the corresponding rate by the weighted random assignment would be $(154/2319)^2 + (592/2319)^2 + (37/2319)^2 + (53/2319)^2 + (230/2319)^2 + (26/2319)^2 + (38/2319)^2 + (86/2319)^2 + (288/2319)^2 + (32/2319)^2 + (758/2319)^2 + (25/2319)^2 \approx 20.5\%$, provided one uses the number of proteins in each subcellular location as given in Appendix A to represent the size of each subset. Therefore, the rates of correct prediction obtained by using the covariant discriminant algorithm in both the self-consistency and jackknife tests are much higher than the corresponding completely randomized rate and weighted randomized rate, implying that the cellular location of a protein is considerably correlated with its amino acid composition.

2. When the number of subcellular locations considered was reduced from 12 (S^{12}) to seven (S^7) and five (S^5) by excluding small subsets (see Table I), the corresponding rates were increased to 73.1 and 80.0% and 78.3 and 83.1%, respectively. This indicates that the prediction quality can be substantially improved if one can (i) 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 is from an animal organism, one can exclude the chloroplast and vacuole subsets from consideration and the prediction will be made among 10 possible subcellular locations instead of 12); and (ii) 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.

3. As a demonstration of a practical application, predictions were also performed for the three independent datasets $\overline{S}^{12}$, $\overline{S}^7$ and $\overline{S}^5$ using the rule parameters derived from the datasets S^{12}, S^7 and S^5, respectively. The overall rates of correct prediction thus obtained are also given in Table III, from which it can be seen that the rates of correct prediction by the current algorithm are in the range 75.9–81.8%, fully consistent with the results obtained by the self-consistency and jackknife tests.

4. No matter whether the self-consistency test, the jackknife test or the independent dataset test is used, the overall rates of correct prediction obtained by the current algorithm are significantly higher than those obtained by the ProtLock algorithm (Cedano *et al*., 1997). For the case of five subcellular locations, the rates of correct predictions by the current algorithm are 8.8–13.6% higher, for seven subcellular locations 17.5–26.8% higher and for 12 subcellular locations 24.5–35.9% higher. The above data also clearly indicate that the greater the number of subcellular locations considered, the more significant the improvement of prediction quality would be by using the current algorithm. In other words, the covariant discriminant algorithm is particularly powerful when used to deal with a classification with many possible categories.

5. The comparison of prediction quality was also extended to cover other algorithms, such as the least city-block distance algorithm (P.Y.Chou, 1980, 1989), and the least Euclidean algorithm (Nakashima *et al*., 1986). Both of these algorithms were developed for predicting the structural class of a protein according to its amino acid composition, and hence can be directly applied to predicting the protein subcellular locations based on the same datasets as used here. It was found that for the case of 12 subcellular locations, the overall rates of correct prediction by using the least city-block distance algorithm (P.Y.Chou, 1980, 1989) for the self-consistency, jackknife and independent dataset tests were 47.9, 46.4 and 45.4%, respectively, and the corresponding rates by the least Euclidean algorithm (Nakashima *et al*., 1986) were 48.1, 46.7 and 46.6%. Compared with these results, the overall rates of correct prediction by using the current algorithm are about 22–32% higher.

The current algorithm was also used to test the dataset studied by Nakai and Kanehisa (1991). From Gram-negative bacteria these authors extracted 106 proteins, of which 34 are inner membrane proteins, 21 periplasmic proteins, 22 outer membrane proteins and 29 cytoplasmic proteins (see Table 1 in Nakai and Kanehisa, 1991). According to their report, the self-consistency by using the expert system to predict the localization sites of the 106 proteins was 83%. No cross-validation was performed in their study. For the same database, when using the ProtLock algorithm (Cedano *et al*., 1997), the corresponding rate was 85%. However, when using the current algorithm, the corresponding rate was 99%, further indicating its power.

To demonstrate its power further, the current algorithm was also used to test the dataset recently studied by Reinhardt and Hubbard (1998). After discarding those groups in which the amount of data available is too small for statistical analysis, these authors classified 997 prokaryotic proteins into three different subcellular locations: 688 cytoplasmic, 107 extracellular and 202 periplasmic proteins. Within each group none had >90% sequence identity with any other. According to their report, for such a dataset the rate of correct prediction by them using the neural network method for a subsampling test was 81%. This is the highest accuracy rate so far reported for a cross-validation test in protein cellular location prediction. Now for the same dataset, when using the discriminant function algorithm to perform prediction, we found that the rate of correct prediction was 91% by self-consistency test and 86% by jackknife test; both are considerably higher than 81%. Further, in their subsampling procedure, only a very small fraction of the possible divisions were investigated (Chou and Elrod, 1998), and the results thus obtained would certainly bear considerable arbitrariness. Actually, compared with the limited subsampling test, the jackknife test is much more objective and rigorous (Mardia, 1979). Accordingly, from both the percentage of correct prediction and the rationality of cross-validation, a higher prediction quality can be obtained by using the current algorithm.

That the current algorithm can lead to the best prediction quality is because it takes into account the coupling effect among different amino acid components, which is a kind of collective interaction, as formulated by a set of covariance matrices in Equation 7, $\mathbf{C}_\xi(\xi = 1, 2, \ldots, m)$, that is the core of the current algorithm. It is through each of these matrices that a more reasonable statistical distance (K.C.Chou, 1995; Chou and Zhang, 1995), the Mahananobis distance, in the amino acid composition space is defined (see the first term of Equation 5), and it is through the eigenvalues of these matrices that the coupling effects in different subsets as well as their sizes are reflected (see the second term of Equation 5). It

Table IV. The standard vector derived from the training dataset of Appendix A for each of the 12 protein subcellular locations

	Subcellular location of proteins											
	Chloro-plast	Cytop-lasmic	Cytoske-letal	Endoplasmic reticulum	Extra-cellular	Golgi apparatus	Lyso-some	Mitochon-drial	Nuc-lear	Peroxi-some	Plasma membrane	Vacuo-lar
Amino acid code	$\overline{\mathbf{X}}^1$	$\overline{\mathbf{X}}^2$	$\overline{\mathbf{X}}^3$	$\overline{\mathbf{X}}^4$	$\overline{\mathbf{X}}^5$	$\overline{\mathbf{X}}^6$	$\overline{\mathbf{X}}^7$	$\overline{\mathbf{X}}^8$	$\overline{\mathbf{X}}^9$	$\overline{\mathbf{X}}^{10}$	$\overline{\mathbf{X}}^{11}$	$\overline{\mathbf{X}}^{12}$
	Components of the standard vector (normalized to 1)											
A	0.086	0.079	0.078	0.068	0.080	0.063	0.070	0.084	0.083	0.086	0.080	0.074
C	0.016	0.015	0.014	0.017	0.020	0.016	0.023	0.013	0.015	0.013	0.020	0.023
D	0.052	0.058	0.055	0.063	0.053	0.056	0.052	0.039	0.046	0.056	0.036	0.058
E	0.064	0.072	0.096	0.075	0.053	0.070	0.049	0.048	0.064	0.059	0.043	0.065
F	0.038	0.041	0.030	0.046	0.040	0.043	0.044	0.050	0.029	0.041	0.059	0.041
G	0.071	0.075	0.049	0.064	0.077	0.058	0.080	0.075	0.066	0.077	0.068	0.076
H	0.016	0.024	0.021	0.028	0.022	0.020	0.025	0.020	0.026	0.024	0.019	0.025
I	0.056	0.059	0.047	0.053	0.049	0.061	0.045	0.060	0.036	0.060	0.073	0.047
K	0.064	0.064	0.086	0.070	0.058	0.062	0.046	0.062	0.075	0.066	0.042	0.056
L	0.086	0.093	0.089	0.092	0.083	0.098	0.097	0.099	0.080	0.088	0.113	0.078
M	0.025	0.025	0.022	0.020	0.021	0.027	0.022	0.028	0.022	0.020	0.030	0.018
N	0.041	0.040	0.046	0.041	0.053	0.048	0.049	0.040	0.044	0.044	0.037	0.059
P	0.050	0.047	0.043	0.048	0.049	0.043	0.061	0.047	0.072	0.051	0.044	0.042
Q	0.033	0.036	0.055	0.038	0.042	0.045	0.040	0.038	0.051	0.036	0.030	0.047
R	0.050	0.049	0.056	0.044	0.039	0.046	0.040	0.048	0.058	0.048	0.044	0.037
S	0.085	0.058	0.077	0.063	0.077	0.078	0.077	0.075	0.096	0.065	0.073	0.080
T	0.055	0.052	0.053	0.051	0.061	0.059	0.054	0.062	0.053	0.052	0.057	0.053
V	0.075	0.070	0.054	0.067	0.071	0.067	0.062	0.065	0.048	0.071	0.078	0.070
W	0.010	0.013	0.007	0.015	0.015	0.010	0.023	0.015	0.008	0.012	0.018	0.012
Y	0.027	0.032	0.022	0.036	0.038	0.030	0.042	0.035	0.028	0.031	0.035	0.039

should be pointed out that although the ProtLock algorithm (Cedano *et al.*, 1997) also contained a covariance matrix, it did not reflect the special character for each of the individual subsets. Particularly, in the ProtLock algorithm, a critical term, i.e. the second term of Equation 5, was completely missed. For a detailed discussion of this aspect, see Appendix D, where two important differences between the current algorithm and ProtLock are illustrated.

To show the difference in amino acid compositions that distinguish the subcellular locations of proteins, the 20-D standard vector derived from the proteins in the training dataset of Appendix A for each of the 12 subcellular locations is given in Table IV. Further, to provide an intuitive picture, each such 20-D standard vector is projected on to a 2-D radar diagram as given in Figure 2. In addition, the 19 positive eigenvalues for each of the 12 corresponding covariance matrices (see Equations 7 and 12) are given in Table V that might be of use for investigating the component-coupled effects at a deeper level, especially for understanding the important contribution from the second term of Equation 5 as illustrated in Figure 3. This is a vitally important term for dealing with the case where the sizes of subsets are different. However, such an important term and also the denominator $n_\xi - 1$ in Equation 8 were not included in the original least Mahalanobis distance algorithm (K.C.Chou, 1995), although good results were still obtained because the case studied there consisted of subsets with the same size. It is very important to realize this, otherwise the prediction algorithm might be misused, leading to poor results and an incorrect conclusion, as elaborated in a recent paper (Chou *et al.*, 1998).

Conclusion

The idea of predicting the subcellular location of a protein according to its amino acid composition is based on the following rationale. (i) Different compartments of a cell usually have different physio-chemical environments which might be very sensitive in selectively accommodating a protein according to its structural feature, particularly its surface physical chemistry

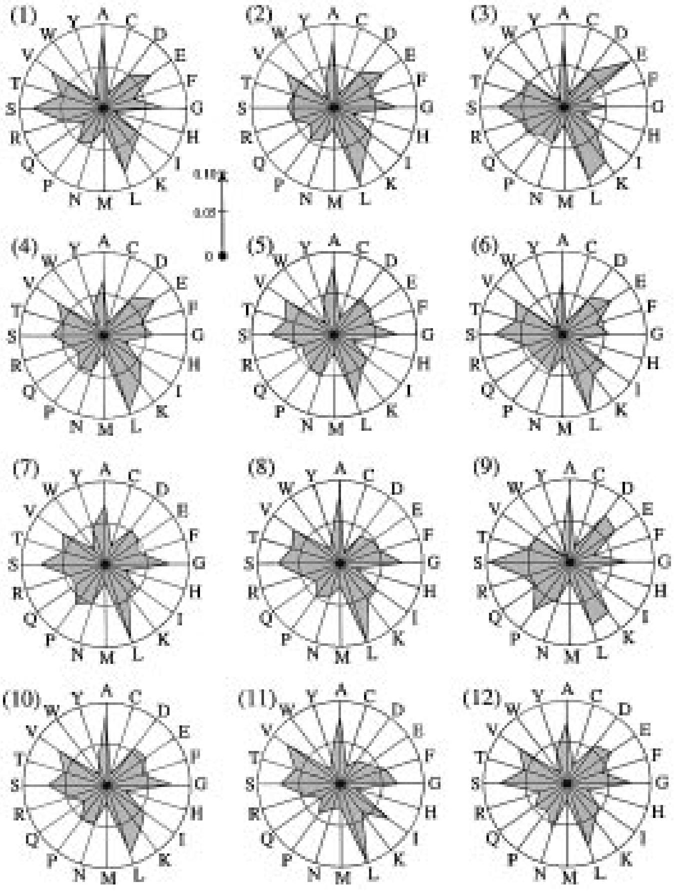

Fig. 2. Radar diagrams to show the difference of the 20-D standard vectors, i.e. the average amino acid compositions for the proteins in the following subcellular locations: (1) chloroplast, (2) cytoplasm, (3) cytoskeleton, (4) endoplasmic reticulum, (5) extracell, (6) Golgi apparatus, (7) lysosome, (8) mitochondria, (9) nucleus, (10) peroxisome, (11) plasma membrane and (12) vacuole. Amino acids are denoted by their single-letter codes (see Table IV).

Table V. The 19 positive eigenvalues of the covariance matrix derived from the training dataset of Appendix A for each of the 12 protein subcellular locations

	Subcellular location											
	Chloroplast	Cytoplasmic	Cytoskeletal	Endoplasmic reticulum	Extracellular	Golgi apparatus	Lysosome	Mitochondrial	Nuclear	Peroxisome	Plasma membrane	Vacuolar
Order	λ_i^1	λ_i^2	λ_i^3	λ_i^4	λ_i^5	λ_i^6	λ_i^7	λ_i^8	λ_i^9	λ_i^{10}	λ_i^{11}	λ_i^{12}
i	Eigenvalues $\times 10^5$											
2	0.4	0.6	0.1	0.4	1.0	0.01	0.3	0.9	0.6	0.1	0.5	0.1
3	3.0	5.9	0.4	2.1	6.1	0.3	0.8	3.9	3.2	0.6	6.4	0.3
4	5.3	6.6	0.9	2.7	9.9	0.8	1.3	5.5	8.6	1.0	7.2	1.1
5	6.7	8.7	2.2	5.0	11.2	1.2	2.2	7.5	10.2	1.1	7.5	2.1
6	7.4	9.7	2.7	5.3	14.9	2.5	2.9	8.4	13.7	1.2	8.2	3.2
7	9.1	11.6	3.4	7.3	18.6	4.3	3.8	10.8	13.9	1.8	9.4	3.8
8	11.7	13.1	4.6	8.3	20.0	5.6	4.9	13.8	16.2	3.4	11.5	6.5
9	12.3	13.6	5.7	11.4	23.2	7.3	5.9	16.4	19.8	3.9	11.8	7.8
10	14.1	14.9	9.1	11.9	24.5	9.1	7.1	20.9	28.5	4.8	12.6	11.2
11	18.2	18.1	14.2	13.0	29.5	12.6	9.9	22.7	29.9	6.0	16.9	12.8
12	18.4	19.5	17.6	21.1	34.7	13.8	10.1	29.9	32.8	7.1	17.1	14.9
13	22.8	22.2	19.5	24.9	40.7	18.2	15.0	34.2	48.6	9.9	18.4	22.7
14	33.0	27.6	33.1	28.6	45.9	26.6	18.1	36.2	66.3	12.3	21.6	35.7
15	36.5	29.7	41.5	50.6	53.8	34.9	23.2	41.8	77.4	14.9	28.5	54.3
16	38.2	33.1	54.8	52.9	76.2	49.2	27.9	49.0	87.6	26.4	31.0	99.5
17	45.6	36.6	69.6	66.4	80.4	65.5	36.0	64.5	106.0	33.0	35.6	142.9
18	54.7	57.4	108.9	86.6	118.8	101.9	48.0	85.8	139.4	45.0	80.4	204.9
19	82.4	86.9	172.4	115.0	121.7	110.7	73.7	101.9	239.2	92.4	90.6	241.9
20	117.5	122.3	329.1	220.0	200.3	209.7	206.2	166.0	462.2	108.4	127.7	472.4

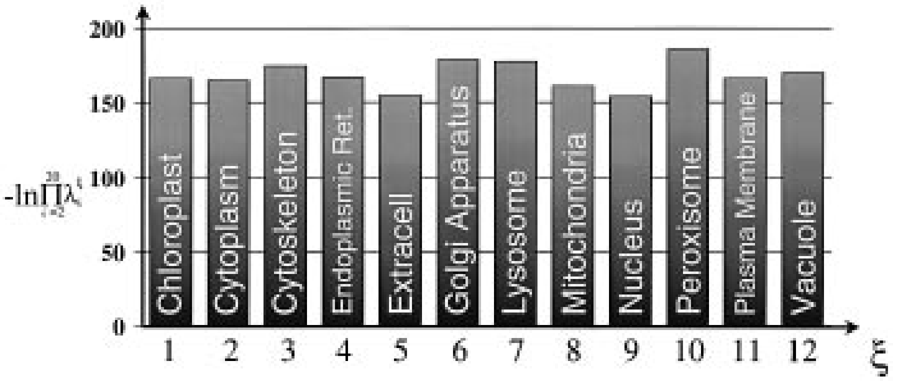

Fig. 3. Histograms to show the contributions of $\ln(\lambda_2^\xi\,\lambda_3^\xi\,\lambda_4^\xi\,\ldots\,\lambda_{20}^\xi)$ from different subsets to the covariant discriminant function of Equation 5. As can be seen, the heights of the 12 histograms are considerably different. Only when the heights are the same can the second term of Equation 5 be omitted from the prediction algorithm.

character. (ii) The structural class of a protein, one of the most basic structural features, is correlated with its amino acid composition, as reflected by many encouraging reports of predicting the former based on the latter alone (see, e.g., P.Y.Chou, 1980; Klein and Delisi, 1986; Nakashima *et al.*, 1986; K.C.Chou, 1995; Chou and Zhang, 1995; Bahar *et al.*, 1997). (iii) The character of a protein surface, which is directly exposed to the environment of a cellular compartment, is also very likely correlated with the amino acid composition because it is determined by a sequence-folding process during which the interaction among different amino acid components might also play an important role. (iv) The above correlations suggest that the total amino acid composition might carry a 'signal' that identifies the subcellular location. (v) Compared with the existing algorithms, the covariant discriminant algorithm proposed in this paper can give the best prediction quality for the protein subcellular location.

Acknowledgements

Valuable discussions with Professor Ferenc J.Kézdy and Dr Reqiang Yan are gratefully acknowledged. The authors are also indebted to Dr Viv Junker for interpreting the annotations in the Swiss Protein Data Bank, to Dr Nakai and Dr A.Reinhardt for providing their datasets for testing the covariant discriminant algorithm and to Raymond B.Moeller, Cynthia A.Ludlow and Diane M.Ulrich for drawing the figures.

References

Alberts,B., Bray,D., Lewis,J., Raff,M., Roberts,K. and Watson,J.D. (1994) *Molecular Biology of the Cell*, 3rd edn. Garland Publishing, New York, London, Ch. 1.

Andrade,M.A., O'Donoghue,S.I. and Rost,B. (1998) *J. Mol. Biol.*, **276**, 517–525.

Bahar,I., Atilgan,A.R., Jernigan,R.L. and Erman,B. (1997) *Proteins*, **29**, 172–185.

Bairoch,A. and Apweiler,R. (1997) *Nucleic Acids Res.*, **25**, 31–36.

Cedano,J., Aloy,P., Pérez-Pons,J.A. and Querol,E. (1997) *J. Mol. Biol.*, **266**, 594–600.

Chou,K.C. (1995) *Proteins: Struct. Funct. Genet.*, **21**, 319–344.

Chou,K.C. and Elrod,D.W. (1998) *Biochem. Biophys. Res. Commun.*, **252**, 63–68.

Chou,K.C. and Maggiora,G.M. (1998) *Protein Engng*, **11**, 523–538.

Chou,K.C. and Zhang,C.T. (1995) *Crit. Rev. Biochem. Mol. Biol.*, **30**, 275–349.

Chou,K.C., Liu,W., Maggiora,G.M. and Zhang,C.T. (1998) *Proteins: Struct. Funct. Genet.*, **31**, 97–103.

Chou,P.Y. (1980) In *Abstracts of Papers, Part I, Second Chemical Congress of the North American Continent*, Las Vegas.

Chou,P.Y. (1989) In Fasman,G.D. (ed.), *Prediction of Protein Structure and the Principles of Protein Conformation*. Plenum Press, New York, pp. 549–586.

Claros,M.G., Brunak,S. and von Heijne,G. (1997) *Curr. Opin. Struct. Biol.*, **7**, 394–398.

Klein,P. and Delisi,C. (1986) *Biopolymers*, **25**, 1569–1672.

Liu,W. and Chou,K.C. (1998) *J. Protein Chem.*, **17**, 209–217.

Lodish,H., Baltimore,D., Berk,A., Zipursky,S.L., Matsudaira,P. and Darnell,J. (1995) *Molecular Cell Biology*, 3rd edn. Scientific American Books, New York, Ch. 3.

Mahalanobis,P.C. (1936) *Proc. Natl Inst. Sci. India*, **2**, 49–55.

Mardia,K.V., Kent,J.T. and Bibby,J.M. (1979) *Multivariate Analysis*. Academic Press, London, pp. 322 and 381.

Nakashima,H. and Nishikawa,K. (1994) *J. Mol. Biol.*, **238**, 54–61.

Nakashima,H., Nishikawa,K. and Ooi,T. (1986) *J. Biochem.*, **99**, 152–162.

Nakai,K. and Kanehisa,M. (1991) *Proteins: Struct. Funct. Genet.*, **11**, 95–110.

Nakai,K. and Kanehisa,M. (1992) *Genomics*, **14**, 897–911.

Pillai,K.C.S. (1985) In Kotz,S. and Johnson,N.L. (eds), *Encyclopedia of Statistical Sciences*, Vol. 5. Wiley, New York, pp. 176–181.

Reinhardt,A. and Hubbard,T. (1998) *Nucleic Acids Res.*, **26**, 2230–2236.

Received July 28, 1998; revised October 16, 1998; accepted October 21, 1998

Appendix A

List of the 2319 proteins located in 12 different subcellular locations, with codes according to the SWISS-PROT data bank

(1) 154 chloroplast proteins

ACCA_ANTSP ACP1_CUPLA ACP2_CUPLA ACP3_CUPLA ACP4_CUPLA AKH1_MAIZE AKH2_MAIZE ALFC_SPIOL ALFD_PEA ARO1_TOBAC
AROA_ARATH AROC_CORSE AROF_ARATH AROG_ARATH AROL_LYCES BCCP_PORPU BGLC_MAIZE CAHC_ARATH CH10_SPIOL CHMU_ARATH
CLAA_LYCES CLAB_LYCES CLPA_PEA CLPC_ODOSI CRTI_ARATH CYP4_ARATH CYSL_ARATH DAP1_WHEAT DAP2_WHEAT DAPA_MAIZE
DHAB_ATRHO DPEP_SOLTU EFGC_SOYBN EFTS_GALSU EFTU_ARATH ELI5_HORVU ELI6_HORVU ELI9_HORVU ELI_PEA F16P_ARATH
FABB_ARATH FABG_ARATH FABH_ARATH FABI_BRANA FER1_DUNSA FER2_DUNSA FER3_MAIZE FER5_MAIZE FER_ARATH FRI2_MAIZE
FRI_PEA FTRC_MAIZE FTRD_SPIOL FTRV_SPIOL G3PA_ARATH G3PB_ARATH G6PI_CLAUN GGPP_ARATH GLB1_CHLEU GLB2_CHLEU
GLG1_BETVU GLG2_SOLTU GLG3_SOLTU GLGS_HORVU GLN2_HORVU GLN4_PHAVU GLNC_MAIZE GLSF_ANTSP GSA_ARATH HEM1_ARATH
HEM2_SELMA HEMZ_ARATH HISX_BRAOC HO_PORPU HS2C_ARATH HS7S_PEA IF2_PORPU IF3C_EUGGR ILV5_ARATH KADC_MAIZE
LEU3_BRANA MAOC_FLAPR MDHC_FLABI MDHD_SORVU METC_ARATH ODPA_PORPU ODPB_PORPU PGKH_CHLRE PHS1_SOLTU PHS2_SOLTU
PHSL_IPOBA PLSB_CUCMO PMGI_ANTSP PODK_FLATR PPOA_LYCES PPOB_LYCES PPOC_LYCES PPOD_LYCES PPOE_LYCES PPOF_LYCES
PPO_MALDO PSY1_LYCES PSY_ARATH PUR1_SOYBN PUR3_ARATH PUR5_ARATH RBL_ABIMA RBS0_SOLTU RBS1_ACEME RBS2_ARATH
RBS3_ACECL RBS4_ACECL RBS5_ACECL RBS6_LEMGI RBS8_NICPL RBSA_SOLTU RBSB_SOLTU RBSC_SOLTU RBSX_TOBAC RBS_ANTSP
RCA_ARATH RK15_ARATH RK18_PEA RK22_MEDSA RK24_PEA RK40_SPIOL RK9_ARATH RO28_NICSY RO30_NICPL RO31_ARATH
RO33_NICSY RR13_ARATH RR17_ARATH RR30_SPIOL RUB2_BRANA RUBA_PEA RUBB_ARATH S17P_ARATH SECA_ANTSP SODF_SOYBN
SODP_LYCES SR5C_ARATH STAD_BRANA SYFB_PORPU SYH_PORPU THD1_LYCES THIF_PEA THIM_PEA THIO_CYACA TPIC_SECCE
UCRA_TOBAC UCRB_TOBAC UCRI_CHLRE UGST_HORVU

(2) 592 cytoplasmic proteins

143F_BOVIN 143G_BOVIN 3HAO_HUMAN 3O5B_HUMAN 5NTC_HUMAN AACA_STAAU AAT1_MEDSA AATC_BOVIN AAT_ECOLI ABFA_STRLI
ABL1_HUMAN ABL2_HUMAN ABL_DROME ACEA_CORGL ACEK_ECOLI ACKA_CLOTS ACLY_HUMAN ACOC_ARATH ACT1_FUGRU ACT2_FUGRU
ACT3_BOMMO ACT5_CHICK ACT8_XENLA ACTA_CHICK ACTB_CRIGR ACTG_HUMAN ACTH_HUMAN ACY1_HUMAN ADH1_ALLMI ADH2_HORVU
ADH3_COTJA ADH6_HUMAN ADH7_HUMAN ADHA_HUMAN ADHB_HUMAN ADHE_HORSE ADHG_HUMAN ADHI_RHOSH ADHP_HUMAN ADHS_HORSE
ADHX_HORSE ADH_FRAAN ADI_ECOLI ADO_BOVIN ALAT_HUMAN ALDR_BOVIN ALF1_PEA ALF2_PEA ALF_ARATH ALKH_BACSU
ALKK_PSEOL AMOH_ARTGO AMPL_ARATH AMPN_LACHE AMY1_DICTH AMY2_DICTH AMY3_DICTH APT1_ARATH APT2_YEAST APT_CRILO
APX1_ARATH ARGI_HUMAN ARGJ_CORGL ARI1_PENRO ARY1_HUMAN ARY2_HUMAN ARY3_MOUSE ASG1_ECOLI ASPG_BACLI ASRB_SALTY
ASRC_SALTY ATDA_HUMAN ATE1_YEAST BAXB_HUMAN BAXC_HUMAN BCAT_CAEEL BGLB_MICBI BIEA_HUMAN BLMH_RAT BNC2_RAT
BODG_PSESK BTUR_ECOLI BUP_RAT BV1C_BETVE C1TC_HUMAN CAFA_ECOLI CAH1_HORSE CAH2_BOVIN CAH3_HORSE CAIB_ECOLI
CAN1_HUMAN CAN2_CHICK CAN3_HUMAN CANX_CHICK CAN_DROME CAP1_FLAPR CAP2_FLATR CAP3_SORVU CAPP_AMAHP CARA_YEAST
CARB_TRICU CATA_MICLU CATR_PSEPU CATT_YEAST CBS_RAT CC2H_PLAFK CFA1_MYCTU CFA2_MYCTU CFA_ECOLI CGL_HUMAN
CHEA_ECOLI CHEB_ECOLI CHLR_HUMAN CHMU_BACSU CILA_ECOLI CILB_ECOLI CKI1_SCHPO CKI2_SCHPO CNTF_CHICK COA1_HUMAN
COA2_HUMAN COAC_CHICK COBO_PSEDE CSCA_ECOLI CSW_DROME CTK_HUMAN CYAA_AERHY CYG1_BOVIN CYG2_RAT CYG3_BOVIN
CYG4_HUMAN CYG5_HUMAN CYP4_BOVIN CYPB_ECOLI CYPC_ECOLI CYPH_BLAGE CYS3_YEAST CYSE_ECOLI CYSK_SPIOL DAPD_ACTPL
DBDD_HUMAN DCK_HUMAN DCP_ECOLI DCUP_HUMAN DDLA_ECOLI DDLB_ECOLI DEOC_BACSU DEXB_STRMU DHA6_YEAST DHAC_BOVIN
DHAP_HUMAN DHAR_RAT DHAS_MOUSE DHA_BACSH DHB1_HUMAN DHCA_HUMAN DHGY_METEX DHQU_HUMAN DHQV_HUMAN DIDH_RAT
DLD1_BACST DLD2_PSEPU DLD3_PSEPU DLDH_ALCEU DLTA_LACCA DPS1_PINST DPS2_PINST DPSS_PINSY DPYD_HUMAN DUS6_HUMAN
DYHC_CAEEL DYL1_HUMAN E4PD_ECOLI EF10_XENLA EF1C_PORPU ENO1_ENTHI ENO2_MAIZE ENOA_ANAPL ENOB_CHICK ENOG_HUMAN
ENO_ARATH ENP2_BACSH EPSC_BURSO ERG8_YEAST EXOA_BACSU F16Q_BETVU F3ST_FLABI F4ST_FLACH FABA_ECOLI FABB_ECOLI
FACC_HUMAN FAD1_YEAST FKB1_BOVIN FKBP_CANAL FPPS_ARATH FTDH_RAT FTHC_HUMAN FUCI_ECOLI FUMC_BRAJA G3P1_AGABI
G3P2_AGABI G3P3_ANAVA G3PC_ANTMA G3PX_HORVU G3P_ASPNG G6P1_CLALE G6P2_CLALE G6PA_BACST G6PB_BACST G6PI_ARATH
GAL_PSEFL GAPN_MAIZE GCY_YEAST GGPP_NEUCR GLB1_SCAIN GLMS_BACSU GLMT_RAT GLN1_ALNGL GLN2_BRAJA GLN3_HORVU
GLN4_MAIZE GLN5_MAIZE GLNA_AGABI GLPD_BACSU GLYA_ACTAC GLYC_HUMAN GNO_GLUOX GPDA_DROME GPP1_YEAST GPP2_YEAST
GSHC_BOVIN GSHR_ANASP GTA1_HUMAN GTA2_HUMAN GTA3_CHICK GTA_PLEPL GTC1_RAT GTC2_RAT GTC_MOUSE GTH_SILCU
GTM1_HUMAN GTM2_CHICK GTM3_HUMAN GTM4_HUMAN GTM5_HUMAN GTMU_CAVPO GTS_OMMSL GTT1_CHICK GTT2_HUMAN GT_ECOLI
GUAA_HUMAN HEM6_ECOLI HEMG_BACSU HGXR_TOXGO HMC6_DESVH HMCS_CHICK HMT_HUMAN HOSC_YEAST HOXF_ALCEU HOXH_ALCEU
HOXU_ALCEU HOXY_ALCEU HPRT_BACSU HXKG_ECOLI I1BC_HUMAN IADA_ECOLI ICE6_HUMAN ICE7_HUMAN IDE_HUMAN IDHC_RAT
IDH_SYNY3 IFEA_HELAS IFEB_HELAS IFE_BRALA IFRH_MAIZE INO1_ARATH INVA_ZYMMO IPPI_SCHPO IPYR_BACP3 IREB_MOUSE
ISP1_BACSU ISPA_BACST ITK_HUMAN JNK3_HUMAN KAD1_BOVIN KAD_BACST KC1A_BOVIN KC1B_BOVIN KC1D_HUMAN KCOT_HUMAN
KCRB_CANFA KCRM_CANFA KDSA_CHLPS KICH_HUMAN KIME_HUMAN KKA4_BACCI KPYC_SOLTU KRB1_VACCC LB4D_PIG LEU3_AGRTU
LIK1_HUMAN LIK2_HUMAN LIPA_ECOLI LKHA_CAVPO LON1_BACSU LON2_MYXXA LON_BACBR LOX1_ARATH LOX2_BOVIN LOX3_PEA
LOX4_SOYBN LOX5_HUMAN LOXA_LYCES LOXB_LYCES LOXL_MOUSE LOXP_MOUSE LOXX_SOYBN LPCA_ECOLI LPLA_ECOLI MALQ_ECOLI
MALZ_ECOLI MANA_YEAST MAOX_ANAPL MASY_CORGL MCH_METTH MDHC_ECHGR MEPD_HUMAN METB_ECOLI METC_BORAV METH_HUMAN
METK_ECOLI MLER_LACLA MT17_YEAST MTD1_YEAST MURF_ECOLI NAT1_YEAST NCK_HUMAN NDKC_DICDI NDK_BACSU NEUA_ECOLI
NIRD_ECOLI NMT_AJECA NNMT_HUMAN NODA_AZOCA NODB_AZOCA NRDG_ECOLI O16G_BACCE OAT_EMENI OMP_HUMAN OTC1_ECOLI
OTC2_BACSU OTCA_MYCBO OTCC_CLOPE OTC_HAEIN P2A1_ARATH P2A2_ARATH P2A3_ARATH P2A4_ARATH P2AA_CHICK P2AB_HUMAN
PA1F_HUMAN PA1S_HUMAN PCP_BACAM PDXK_HUMAN PE2R_RABIT PEPC_LACHE PEPE_ECOLI PEPT_BACSU PEPX_LACLA PFLA_CLOPA
PFLB_ECOLI PFPN_ENTHI PGDH_HUMAN PGF2_BOVIN PGFS_BOVIN PGK1_TRYCO PGKB_CRIFA PGKC_ALCEU PGKE_TRYBB PGKP_ALCEU
PGKY_TOBAC PGK_BACME PGM1_YEAST PGM2_YEAST PH2M_TRICU PHAB_ACISP PHBB_ALCEU PHBC_ALCEU PHEA_ECOLI PHHC_PSEAE
PHSH_SOLTU PIMT_ARATH PKN5_MYXXA PLSI_HUMAN PLSL_HUMAN PMGI_MAIZE PMM1_HUMAN PMM_CANAL PNPA_BACSU PNP_ECOLI
POLO_DROME PP11_YEAST PP12_DROME PP1A_HUMAN PP1G_HUMAN PPAC_BOVIN PPAL_SCHPO PPCC_CHICK PPCE_HUMAN PPCK_UROPA
PPV_DROME PRCA_METJA PRCB_METJA PROB_BACSU PROC_ARATH PT1A_ECOLI PT1_ALCEU PTCA_ECOLI PTCB_ECOLI PTFA_BACSU
PTFB_BACSU PTGA_ECOLI PTHA_ECOLI PTH_ECOLI PTI8_HUMAN PTI9_HUMAN PTKA_ECOLI PTKB_ECOLI PTLA_LACCA PTMA_ENTFA
PTN2_HUMAN PTN6_HUMAN PTN8_MOUSE PTNA_ECOLI PTNB_HUMAN PTNC_HUMAN PTP1_YEAST PTP2_YEAST PTP3_DICDI PTRA_KLEPN
PTRB_KLEPN PTWB_ECOLI PTWX_ECOLI PUA2_MOUSE PUR4_YEAST PYC1_YEAST PYC2_YEAST PYC_PICPA PYP1_SCHPO PYP2_SCHPO
PYP3_SCHPO PYR1_DICDI PYRD_YEAST QOR_CAVPO RET3_BOVIN RET4_HUMAN RFFE_ECOLI RIMI_ECOLI RIMJ_ECOLI RIML_ECOLI
RIP3_MAIZE RIP9_MAIZE RIR1_HUMAN RIR2_HUMAN RNB_ECOLI RNC_BACSU RND_ECOLI RNE_ECOLI RURE_ACICA SAHH_HUMAN
SAOX_ARTSP SBMC_ECOLI SCRB_KLEPN SDHL_HUMAN SERC_ECOLI SHC_HUMAN SLYD_ECOLI SOD1_ORYSA SOD2_ORYSA SOD4_MAIZE
SOD5_MAIZE SODC_ACTPL SODD_XENLA SODF_SULSO SOXA_CORSP SOXD_CORSP SOXG_CORSP SPRE_HUMAN SRPH_SYNP7 ST20_YEAST
SUAR_RAT SUDY_RAT SUH1_MOUSE SUH2_MOUSE SUH3_RAT SUHA_HUMAN SUHB_CAVPO SUHS_RAT SUO3_RAT SUO6_RAT
SUOE_BOVIN SUOT_MOUSE SUP1_HUMAN SUP2_HUMAN SUPM_HUMAN SUPP_BOVIN SYAC_YEAST SYA_BARBA SYC_BACSU SYDC_YEAST
SYD_ECOLI SYEC_YEAST SYE_AZOBR SYFA_BACSU SYFB_BACSU SYF_METJA SYGA_BACSU SYGB_BACSU SYG_CHLTR SYH1_SYNY3
SYH2_SYNY3 SYH_ECOLI SYIP_STAAU SYI_CAEEL SYK1_ECOLI SYK2_ECOLI SYKC_YEAST SYK_ACICA SYLC_NEUCR SYL_BACSU
SYMC_YEAST SYM_BACST SYNC_YEAST SYN_BACSU SYP_CHLTR SYQ_ECOLI SYRC_YEAST SYR_BRELA SYSC_YEAST SYS_BACSU
SYT1_BACSU SYT2_BACSU SYTC_HUMAN SYT_BUCAP SYV_BACST SYWC_YEAST SYW_BACST SYY1_BACSU SYY2_BACSU SYYC_YEAST
SYY_BACCA TAGD_BACSU TAGE_BACSU TAGF_BACSU TBUD_BURPI THGA_ECOLI THIK_ECOLI THIL_ALCEU THL_BACSU THS1_ARAHY
THS2_VITVI THS3_ARAHY TPIS_HORVU TPMT_HUMAN TPP2_HUMAN TRB1_ARATH TREA_YEAST TREC_BACSU TRXB_ECOLI TSA1_YEAST
TSA2_YEAST TYRA_ECOLI TYRB_ECOLI TYSY_ECOLI TYTR_CRIFA UBC1_HUMAN UBIC_ECOLI UBL1_HUMAN UBL3_HUMAN UBL_APLCA
UDPG_BOVIN UGPQ_ECOLI UVRB_ECOLI UVRC_BACSU VATE_BOVIN VATF_HUMAN VDH_STRCO VGB_STAAU XGPT_ECOLI XYLA_ACTMI
YJ9M_YEAST YPR1_YEAST

(3) 37 cytoskeletal proteins

ABP1_SACEX CISY_TETTH CP23_CHICK CYLI_BOVIN NINL_DROME NINS_DROME PAS5_PICPA REST_HUMAN BNK_DROME CALD_CHICK
DCPY_NEUCR MYSA_CAEEL MYSB_CAEEL MYSC_CAEEL MYSD_CAEEL MYSE_CHICK MYSG_CHICK MYSP_CAEEL MYSQ_DROME MYSS_CHICK
MYST_RABIT MYS_AEQIR N214_HUMAN N358_HUMAN NULL_DROME CIN8_YEAST DYN1_CAEEL DYN2_HUMAN DYN3_RAT DYN_DROME

Appendix A. *Continued*

KCRF_STRPU KIP1_YEAST KLP1_CHLRE MAPX_DROME SCP1_MOUSE SCP2_MOUSE VP22_ASFB7

(4) 53 endoplasmic reticulum proteins

ABP1_ARATH ABP2_TOBAC ABP4_MAIZE ANTA_HYDMA CBP2_HUMAN CNBP_MOUSE CRT1_BOVIN CRT2_BOVIN CRTC_CAEEL CRU4_BRANA
CRUA_BRANA CYPB_BOVIN CYPD_YEAST CYSP_PHAVU ENPL_CATRO ER31_RAT ER55_HUMAN ER60_RAT ER72_HUMAN ERG2_MAGGR
ES10_RAT EST1_CAEBR EUG1_YEAST FD31_BRANA FD32_BRANA FD3E_ARATH FD61_SOYBN FD62_SOYBN FD6E_ARATH G6PE_RABIT
GR74_TOBAC GR75_TOBAC GR78_HUMAN GSBP_CHICK HEMA_CVBF HS47_CHICK HS7C_CAEEL IOD1_RAT KRE5_YEAST LHS1_YEAST
MAN1_RAT MTP_HUMAN P4H2_MOUSE P4HA_CAEEL PDI_BOVIN PNOC_HUMAN PTN1_HUMAN RCN_HUMAN SLS1_YARLI SYN5_RAT
UGGG_DROME VS09_ROTB4 VS10_ROTBN

(5) 230 extracellular proteins

A1AF_RABIT A1AS_CAVPO A1AT_BOMMO A1BG_HUMAN AACT_HUMAN ABP_HUMAN ACH1_BOMMO ACH2_LONAC AFAM_HUMAN AGAR_ALTAT
ALB1_SALSA ALB2_SALSA ALBU_BOVIN ALS_HUMAN AMT4_PSESA AMT6_BACS7 AMY1_HORVU AMYB_BACPO AMYG_HORRE AMYP_HUMAN
AMYR_BACS8 ANT3_BOVIN APA1_BOVIN APA2_HUMAN APAR_PIG APC3_CANFA APC4_HUMAN APE_BOVIN API_ACHLY APL3_LOCMI
ARY1_CALVI ARY2_CALVI ARYA_MANSE ARYB_MANSE B2MG_BARIN BAR1_YEAST BTD_HUMAN CAC3_BOVIN CAS1_BOVIN CAS2_BOVIN
CAS3_MOUSE CASB_BOVIN CASK_BOVIN CBG_HUMAN CBPN_HUMAN CETP_HUMAN CFAI_HUMAN CFH1_HUMAN CFHD_HUMAN CFHE_HUMAN
CHI4_BRANA CHIA_CICAR CHIB_LYCES CHIP_BETVU CHOD_BREST CL43_BOVIN COTR_CAVPO CTR1_PENVA CTR2_CANFA CTRA_BOVIN
CTRB_BOVIN CTRL_HALRU CUDP_METAN CUTI_ALTBR DEXT_ARTSP E13A_LYCES E13G_TOBAC E13H_TOBAC E13K_TOBAC E13L_TOBAC
EBA1_FLAME EBA2_FLAME EBA3_FLAME ELAS_PSEAE EP45_XENLA ESP4_LACVV FA8_HUMAN FBP3_STRPU FETA_GORGO FGF6_HUMAN
GDN_HUMAN GLBH_TRICO GLB_ASCSU GP39_HUMAN GRP1_RAT GRP2_RAT GSHP_BOVIN GTF1_STRDO GTF2_STRDO GTFB_STRMU
GTFC_STRMU GUN_ASPAC HCY2_LIMPO HCY6_ANDAU HCYA_EURCA HCYB_PANIN HCYD_EURCA HCYE_EURCA HEMO_HUMAN HIG_DROME
HLT_VIBPA HP20_TAMAS HP25_TAMAS HP27_TAMAS HPT1_HUMAN HPT2_HUMAN HPT_ATEGE HYPB_HYPLI IBP3_BOVIN IBP5_MOUSE
IGUP_HUMAN IML2_DROME INIG_HUMAN INU1_KLUMA KNH1_BOVIN KNH2_BOVIN KNH_HUMAN KNL1_BOVIN KNL2_BOVIN KNL_HUMAN
KNT1_RAT KNT2_RAT LIP_PSESP LP1_BOMMO LP2_BOMMO LP3_BOMMO LP4_BOMMO LP5_BOMMO LSTP_STASI MASP_HUMAN
MIG_HUMAN MIP_TRYCR MS2A_DROMA MS2B_DROMA NKG5_HUMAN NUC_SERMA NUP1_PENCI NUP3_PENSQ OLFM_RANCA PAC6_MOUSE
PAPA_ECOLI PAPH_ECOLI PBPA_STRPN PEDF_HUMAN PEL1_ERWCA PEL3_ERWCA PELA_ERWCA PELB_ERWCA PELC_ERWCA PELD_ERWCH
PELE_ERWCH PELF_ERWCH PEL_BACSU PERL_BOVIN PHB_ALCFA PHL1_BACCE PHL2_BACCE PHL3_BACCE PHLD_BOVIN PHL_LEPIN
PHO2_YARLI PHOA_ASPNG PIL1_ECOLI PIL4_ECOLI PIL5_ECOLI PIL6_ECOLI PIL7_ECOLI PON2_CANFA PON_HUMAN PPT_BOVIN
PROA_LEGPN PROB_STRAG PRSH_ECOLI PRT1_ERWCA PRTS_BOVIN PRTZ_BOVIN PSPA_CANFA PSPB_BOVIN PSPC_BOVIN PSPD_BOVIN
RNBR_BACAM RNLE_LYCES RNS2_NICAL RN_BACCI SACB_STRMU SAP_RAT SAX_RANCA SELP_HUMAN SEPA_STAEP SERA_MANSE
SODE_BRUPA SODF_MYCTU SSP1_BOMMO STAT_HUMAN STRK_STRGR SUBE_BACSU SUBF_BACSU SUBV_BACSU SVS4_RAT SXA2_SCHPO
TCPA_VIBCH THBG_HUMAN THET_THEVU THRB_BOVIN TRY1_ANOGA TRY2_ANOGA TRY3_AEDAE TRY4_ANOGA TRY5_ANOGA TRY6_ANOGA
TRY7_ANOGA TRYA_DROER TRYB_DROER TRYD_DROER TRYE_DROER TRYG_DROME TRYI_DROME TRYP_ASTFL TRYT_DROER TRYU_DROER
TRYZ_DROER UFBP_PIG VTDB_HUMAN VTNC_HUMAN XYN1_COCCA XYNA_STRLI XYNB_STRLI XYNC_PSEFL YGP1_YEAST ZA2G_HUMAN

(6) 26 Golgi apparatus proteins

A471_RAT A472_HUMAN A47H_DISOM ADG_MOUSE AP19_MOUSE AP47_CAEEL ASPX_HUMAN CB45_MOUSE COPA_BOVIN COPB_DROME
COPD_BOVIN COPE_BOVIN COPG_BOVIN COPP_BOVIN COPZ_BOVIN FURI_BOVIN LDLC_CAEEL RAB1_LYMST RAB6_HUMAN RB1A_HUMAN
SFT1_YEAST SPC3_STRPU SYN5_HUMAN TGN3_RAT VP15_YEAST VP34_YEAST

(7) 38 lysosomal proteins

AGAL_HUMAN ARSA_HUMAN ARSB_FELCA ASM_HUMAN ASPG_HUMAN ASPP_AEDAE BGAL_HUMAN BGLR_HUMAN CATB_BOVIN CATC_HUMAN
CATD_CHICK CATH_HUMAN CATL_BOVIN CATS_BOVIN CYS1_DICDI CYS2_DICDI CYS4_DICDI CYS5_DICDI CYSP_TRYBB DIAC_HUMAN
FUCO_CANFA GA6S_HUMAN GALC_HUMAN GL6S_CAPHI HEXA_DICDI HEXB_HUMAN IDS_HUMAN IDUA_CANFA LIPA_HUMAN LYAG_HUMAN
NAGA_HUMAN PCP_HUMAN PPA5_HUMAN PPAL_HUMAN PRTP_HUMAN SAP3_HUMAN SAP_HUMAN SPHM_HUMAN

(8) 86 mitochondrial proteins

ACR1_YEAST ADT1_BOVIN ADT2_ARATH ADT3_BOVIN ADT_CHLKE ATM1_YEAST ATPY_YEAST BPL1_HUMAN C560_BOVIN COQ2_SCHPO
COX1_ALBCO COX2_ACHDO COXT_YEAST COXW_YEAST COXX_YEAST COXY_YEAST CY1_NEUCR CYPH_NEUCR CYT1_CAEBR DCMC_ANSAN
DHSD_CHOCR FABH_BOVIN FLX1_YEAST FOLC_HUMAN FUMH_HUMAN GDC_BOVIN IM17_YEAST IM23_YEAST IMP1_YEAST IMP2_YEAST
LCF2_YEAST LEU1_YEAST M2OM_BOVIN MD10_YEAST MMM1_YEAST MPCP_BOVIN MRS3_YEAST MRS4_YEAST MSP1_CAEEL NEUL_PIG
NI9M_BOVIN NLTP_BOVIN NUAM_BOVIN NUGM_BOVIN NUHM_BOVIN NUJM_NEUCR NUPM_NEUCR NURM_NEUCR NUXM_NEUCR NUYM_NEUCR
OM06_YEAST OM07_YEAST OM20_NEUCR OM22_NEUCR OM37_YEAST OM40_NEUCR OM70_NEUCR PET8_YEAST PMT_YEAST RIM2_YEAST
SDH3_YEAST SDH4_YEAST SHM1_YEAST SMF1_YEAST SMF2_YEAST SYH_YEAST SYV_NEUCR TXTP_HUMAN UCP_HUMAN YAD8_SCHPO
YB8E_YEAST YD1K_SCHPO YDBA_SCHPO YDE9_SCHPO YEA6_YEAST YEO3_YEAST YFL5_YEAST YG2O_YEAST YG5F_YEAST YHG2_YEAST
YIA6_YEAST YMC1_YEAST YMC2_YEAST YMX1_RAPSA YNI3_YEAST ZRC1_YEAST

(9) 288 nuclear proteins

A33_PLEWA AANT_HDVAM ABP1_SCHPO ACE1_YEAST AD4B_BOVIN ADF1_DROME ADR6_YEAST AFLR_ASPFL AG_BRANA ALCR_EMENI
AMT1_CANGA AP2_HUMAN APN1_YEAST AREA_EMENI ARG2_YEAST ARP1_HUMAN ATF2_RAT ATF4_HUMAN ATH5_ARATH ATH7_ARATH
ATO_DROME AX11_ARATH AXI6_PEA B1_USTMA B3_USTMA B5_USTMA B7_USTMA BAF1_KLULA BASO_HUMAN BCL3_HUMAN
BF1_HUMAN BIMB_EMENI BRAC_MOUSE BRC2_DROME BRLA_EMENI BTEB_RAT BUB1_YEAST C46H_HUMAN CB20_HUMAN CB33_YEAST
CB80_HUMAN CBFA_HUMAN CBFX_HUMAN CBF_HUMAN CBP_MOUSE CC16_YEAST CC23_YEAST CCG1_DROME CDK7_HUMAN CDNB_HUMAN
CDX2_MOUSE CDX4_MOUSE CEBB_CHICK CEBG_HUMAN CEB_DROME CENA_HUMAN CF1A_DROME CF1_BOMMO CF23_DROME CF2_DROME
CGM2_SCHPO CHD1_MOUSE CID_DROME CLK1_HUMAN CPC1_NEUCR CPH1_CANAL CPO_DROME CPR1_PETCR CREA_ASPNG CREM_MOUSE
CSE1_YEAST CSE4_YEAST CST2_HUMAN CTF4_CHICK CTK2_YEAST CUT1_SCHPO CYCH_XENLA CYS3_NEUCR DA80_YEAST DAX1_HUMAN
DA_DROME DBP2_SCHPO DBX_MOUSE DET1_ARATH DNL3_HUMAN DNLI_CANAL DP30_CAEEL DPOA_DROME DPOL_EBV DSRA_HUMAN
E74A_DROME EGR1_BRARE EGR4_RAT ELF1_DROME ELG_DROME ELK1_HUMAN ELT2_CAEEL EMC_DROME EMP1_WHEAT ENL_HUMAN
ENP1_YEAST ERC1_HUMAN ERF_HUMAN ERM_HUMAN ESCA_DROME ESP1_YEAST ESTR_CHICK ETS2_CHICK ETV1_MOUSE EVX1_HUMAN
FKB2_BOVIN FKH_DROME FLI1_HUMAN FOSB_HUMAN FOS_CHICK FRA1_HUMAN FTFB_DROME FUS_HUMAN GA15_CRILO GA1B_XENLA
GA5B_XENLA GAGA_DROME GAT1_CHICK GAT3_CHICK GAT5_CHICK GATB_BOMMO GBF2_ARATH GBF4_ARATH GCF_HUMAN GCN4_YEAST
GLI3_HUMAN GLI_HUMAN GLN3_YEAST GROU_DROME GRP2_SINAL GSBP_DROME GSCB_XENLA GSC_BRARE GSH1_MOUSE GSP1_YEAST
H101_CHICK H114_BRARE H11R_CHICK H11_ARATH H13_GLYBA H15_MOUSE H1B_CHITE H1D_HUMAN H1G_STRPU H1O_CHITH
H2A2_HUMAN H2A4_CHICK H2AL_STRPU H2AO_CHITH H2AV_CHICK H2AZ_HUMAN H2A_ACRFO H2B0_HUMAN H2B2_CHLRE H2B4_CHLRE
H2BE_STRPU H2BN_STRPU H5B_XENLA H5_ANSAN H5_CAIMO HAP2_KLULA HAP4_YEAST HAT5_ARATH HBPB_ARATH HDF1_YEAST
HES2_RAT HES5_RAT HEXP_LEIMA HG14_BOVIN HG17_BOVIN HIBN_XENLA HIR2_YEAST HM22_CAEEL HM8_XENLA HMAB_DROME
HME1_BRARE HME3_BRARE HMEV_DROME HMG2_CHICK HMGB_CHITE HMGD_DROME HMGI_HUMAN HMGY_HUMAN HMG_TETPY HMIX_XENLA
HMMD_BRARE HMPR_DROME HMX1_CHICK HMZ1_DROME HN3A_HUMAN HN3G_HUMAN HNFA_HUMAN HOX3_BRAFL HP1_DROME HPR1_CHICK
HSF1_ARATH HSF3_LYCPE HSP2_ALOSE HTF4_HUMAN HX1A_MAIZE HX3_XENLA HXA1_HUMAN HXA4_CHICK HXA7_COTJA HXAB_CHICK
HXB2_HUMAN HXB4_CHICK HXB6_BRARE HXB8_MOUSE HXC4_HUMAN HXC6_HUMAN HXC9_MOUSE HXD1_MOUSE HXD4_CHICK HXD9_HUMAN
HXDB_CHICK HXDD_CHICK ID2_HUMAN ID4_HUMAN IKAR_MOUSE ILF_HUMAN IPF1_HUMAN IRF1_HUMAN IRTF_HUMAN ISL1_BRARE
ISL3_BRARE JUNB_HUMAN KE2_MOUSE KEM1_YEAST KNRL_DROME KU70_HUMAN LAC9_KLULA LAM0_DROME LAMC_HUMAN LEUR_YEAST
LOLL_DROME LOS1_YEAST MA1R_YEAST MA6R_YEAST MAF2_MOUSE MAT2_YEAST MAX_BRARE MBP1_KLULA MCM1_YEAST MCM3_HUMAN
MCR_HUMAN ME18_MOUSE MEF2_HUMAN MET4_YEAST MIG1_KLULA MKS1_YEAST MOT1_YEAST MRF1_YEAST MSSP_HUMAN MTA1_YEAST
MYBA_CHICK NAB2_YEAST NAM8_YEAST NECD_MOUSE NFI2_CHICK NFIC_CHICK NFIR_MESAU NGFI_CANFA NHPA_YEAST NIL2_HUMAN
NIT4_NEUCR NOT2_YEAST NUC2_SCHPO NUMB_DROME NUR1_MOUSE OC3A_HUMAN OC3N_HUMAN OCT1_CHICK OCT6_HUMAN OP2_MAIZE
ORC1_KLULA ORC3_YEAST ORC5_YEAST P53_CERAE PAN2_RAT PAX1_MOUSE PAX3_HUMAN PAX6_BRARE

Appendix A. *Continued*

(10) 32 peroxisomal proteins

ACEA_CANTR ALOX_CANBO AMO_HANPO CAO1_CANTR CAO2_CANTR CAO4_CANMA CAO_YEAST CAT1_GOSHI CAT2_GOSHI CATA_BOVIN
CISZ_YEAST DAS_HANPO DHGY_CUCSA ECHP_CAVPO FOX2_YEAST GOX_RAT HDE_CANTR LUCI_PHOPY MDHP_YEAST OXDA_HUMAN
OXDD_BOVIN PX18_CANMA SPYA_RABIT THI1_RAT THI2_RAT THIK_CANTR THIL_CANTR THIM_CANTR UBCX_PICPA URIC_ASPFL
URID_CANLI XDH_BOVIN

(11) 758 plasma membrane proteins

5H1A_HUMAN 5H1B_CRIGR 5H1D_CANFA 5H1E_HUMAN 5H1F_HUMAN 5H2A_CRIGR 5H2B_HUMAN 5H2C_HUMAN 5H5A_HUMAN 5H5B_MOUSE
5H6_HUMAN 5H7_CAVPO 5HT1_DROME 5HT3_HUMAN 5HTA_DROME 5HTB_DROME AA1R_BOVIN AA2A_CANFA AA2B_HUMAN AA3R_HUMAN
AAAT_MOUSE AC22_STRCO ACH1_CAEEL ACH2_CAEEL ACH3_BOVIN ACH4_CAEEL ACH5_CHICK ACH6_CHICK ACH7_BOVIN ACH9_RAT
ACHA_BOVIN ACHB_BOVIN ACHD_BOVIN ACHE_BOVIN ACHG_BOVIN ACHN_CHICK ACHO_CARAU ACHP_CARAU ACTR_BOVIN ADT_RICPR
AFQ2_STRCO AG22_MOUSE AG2R_BOVIN AG2S_MOUSE AGG2_HUMAN ALCP_THEP3 ALKB_PSEOL AMT_CORGL APJ_HUMAN APRD_PSEAE
AQP1_BOVIN AQP2_HUMAN AQP3_RAT AQP4_HUMAN AQP5_HUMAN AQPA_RANES AQPL_YEAST AQUA_ATRCA AT7B_HUMAN ATA1_SYNY3
ATC1_DICDI ATC2_YEAST ATC3_SCHPO ATC4_YEAST ATC5_YEAST ATCF_RAT 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_ARTSA ATP6_ALBCO ATR1_YEAST ATSY_SYNP7 ATU1_YEAST ATXA_LEIDO ATXB_LEIDO B3A2_HUMAN
B3A3_HUMAN B3AT_CHICK BAC1_HALS1 BAC2_HALS2 BACH_HALSP BACR_HALHA BACS_HALHA BACT_HALVA BENE_ACICA BETP_CORGL
BFR1_SCHPO BIOX_BACSH BLR1_HUMAN BMR1_BACSU BMR2_BACSU BMRP_CANAL BOFA_BACSU BRB1_HUMAN BRB2_HUMAN BRNQ_LACDL
BROW_DROME BRS3_CAVPO BRS4_BOMOR C24B_HUMAN C550_BACSU C561_HUMAN CADA_STAAU CADD_STAAU CALR_HUMAN CAMG_HUMAN
CAN1_YEAST CAR1_SCHPO CASR_BOVIN CB11_RABIT CB12_RABIT CB1R_HUMAN CB21_RABIT CB22_RABIT CB2R_HUMAN CBIN_SALTY
CBIQ_SALTY CCKR_HUMAN CCP1_RAT CCT1_RAT CD20_HUMAN CD2R_HUMAN CD47_HUMAN CD97_HUMAN CFTR_BOVIN CGCC_BOVIN
CGOC_BOVIN 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 CINA_ELEEL CITN_KLEPN CKR1_HUMAN CKR2_HUMAN CKR3_MOUSE CKRV_MOUSE CLC1_HUMAN CLC2_HUMAN CLC3_HUMAN
CLC4_HUMAN CLC5_HUMAN CLC6_HUMAN CLC7_RAT CMLR_STRLI COMA_STRPN COMP_BACSU COX2_BACFI COX3_SYNVU COX4_THEP3
COXM_BRAJA CPSD_STRAG CRF2_RAT CRFR_HUMAN CRNA_EMENI CSG2_YEAST CTK1_RABIT CTPA_MYCLE CTPB_MYCLE CTR1_YEAST
CTR2_MOUSE CVAB_ECOLI CX32_ARATH CX33_MICUN CX41_XENLA CX56_CHICK CXA1_BOVIN CXA2_XENLA CXA3_BOVIN CXA4_HUMAN
CXA5_CANFA CXA6_CANFA CXA7_RAT ÇXA8_CHICK CXB1_HUMAN CXB2_HUMAN CXB3_MOUSE CXB4_MOUSE CXB5_MOUSE CY14_NEUCR
CYA1_BOVIN CYA2_RAT CYA3_RAT CYA4_RAT CYA5_CANFA CYA6_CANFA CYA7_HUMAN CYA8_HUMAN CYA9_MOUSE CYAB_BORPE
CYBH_ALCEU CYB_SULAC 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 DOPR_DROME
DTPT_LACLA DUR3_YEAST EAT1_BOVIN EAT2_HUMAN EAT3_HUMAN EAT4_HUMAN EBI1_HUMAN EDG1_HUMAN EDG2_SHEEP EMP1_HUMAN
EMP2_HUMAN EMP3_HUMAN ER21_CAEEL ER22_CAEEL ERD1_KLULA ERD2_ARATH ERS1_YEAST ET1R_BOVIN ET3R_XENLA ETBR_BOVIN
EXOQ_RHIME EXOY_RHIME EXUT_ECOLI FCEB_HUMAN FCY2_YEAST FDNH_ECOLI FDNI_ECOLI FDOH_ECOLI FDOI_ECOLI FDXH_HAEIN
FET4_YEAST FEUB_BACSU FEUC_BACSU FIXG_RHIME FIXI_RHIME FML1_HUMAN FML2_HUMAN FMLR_HUMAN FRIZ_DROME FSHR_BOVIN
FTSH_BACSU FUR4_YEAST G10D_MOUSE GAA1_BOVIN GAA2_BOVIN GAA3_BOVIN GAA4_BOVIN GAA5_HUMAN GAA6_MOUSE 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 GASR_HUMAN GC96_HUMAN GCRC_MOUSE GCRT_CHICK
GCY4_HUMAN GCY6_HUMAN GEF1_YEAST GLCP_SYNY3 GLHR_ANTEL GLPF_BACSU GLPR_HUMAN GLPT_BACSU GLR1_HUMAN GLR2_HUMAN
GLR3_HUMAN GLR4_HUMAN GLR5_HUMAN GLR6_RAT GLR7_RAT GLRK_CHICK GLR_HUMAN GLTP_BACSU GLTT_BACCA 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 GPRE_RAT GPRF_HUMAN GRA1_HUMAN GRA2_BACSU GRA3_RAT GRB2_BACSU GRB_HUMAN
GRFR_HUMAN GRHR_BOVIN GRPR_HUMAN GTR1_BOVIN GTR2_HUMAN GTR3_CANFA GTR4_HUMAN GTR5_HUMAN GTRL_DROME GU27_RAT
GUDT_BACSU GUSB_BOVIN H218_RAT HAK1_SCHOC HEX6_RICCO HGT1_KLULA HH1R_BOVIN HIP1_YEAST HLY2_ECOLI HLYB_ACTAC
HM74_HUMAN 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 IDD_MOUSE IL8A_HUMAN IL8B_HUMAN
INA1_TRIHA IRK0_RAT IRK1_HUMAN IRK2_CAVPO IRK3_HUMAN IRK4_HUMAN IRK5_HUMAN IRK7_HUMAN IRK9_RAT IRKG_MOUSE
IRKX_MOUSE ITR1_YEAST ITR2_YEAST KBAA_BACSU KDGT_BACSU KHT2_KLULA KINB_BACSU KINC_BACSU LACP_KLULA LCN3_LACLA
LCNC_LACLA LCR1_BOVIN LMRA_STRLN LPLB_BACSU LSHR_HUMAN LSPA_STAAU LYSI_CORGL M6A_MOUSE M6B_MOUSE MA3T_YEAST
MA6T_YEAST 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_CAEEL MEP1_YEAST MEP2_YEAST
MEP3_YEAST MESD_LEUME ML1A_CHICK ML1B_HUMAN MMR_BACSU MOTA_BACSU MRED_BACSU MRG_HUMAN MRP1_HUMAN MSCL_CLOPE
MSHR_BOVIN MTR_NEUCR MYP1_XENLA MYP2_XENLA MYPR_BOVIN NAAA_PIG NABA_RAT NAC1_BOVIN NAC2_RAT NAGC_HUMAN
NAGL_HUMAN NAH1_CRIGR NAH2_RABIT NAH3_HUMAN NAH4_RAT NAH_SCHPO NAMI_BOVIN NANU_RABIT NAPA_ENTHR NAPT_HUMAN
NARI_BACSU NARK_BACSU NARV_ECOLI NASA_BACSU NDHF_BACSU NHAC_BACFI NIST_LACLA NK1R_CAVPO NK2R_BOVIN NK3R_HUMAN
NKC1_HUMAN NKC2_MOUSE NMBR_HUMAN NME1_MOUSE NME2_MOUSE NME3_MOUSE NME4_MOUSE NMZ1_HUMAN NQO7_PARDE NQO8_PARDE
NQOA_PARDE NQOB_PARDE NQOC_PARDE NQOD_PARDE NQOE_PARDE NSR_LACLA NTBE_CANFA NTCH_RAT NTCR_HUMAN NTDO_BOVIN
NTG1_HUMAN NTG2_MOUSE NTG3_HUMAN NTGL_HUMAN NTNO_BOVIN NTPI_ENTHR NTPJ_ENTHR NTPR_RAT NTRY_AZOCA NTR_HUMAN
NTS1_RAT NTS2_RAT NTSE_DROME NTT4_RAT NTT7_RAT NTTA_CANFA NUOA_ECOLI NUOH_ECOLI NUOJ_ECOLI NUOK_ECOLI
NUOL_ECOLI NUOM_ECOLI NUON_ECOLI NUPC_BACSU NY1R_HUMAN NY2R_HUMAN NY4R_HUMAN NYR_DROME OAR_DROME OL1E_HUMAN
OLF0_RAT OLF1_CHICK OLF2_CHICK OLF3_CHICK OLF4_CHICK 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 OPSR_ANOCA OPSU_BRARE OPSV_CHICK
OPUB_BACSU OPUD_BACSU OXYR_HUMAN P2X1_RAT P2Y4_HUMAN PACR_HUMAN PAFR_CAVPO PAR2_HUMAN PATC_DROME PBP4_NOCLA
PBUX_BACSU PDR5_YEAST PDUF_SALTY PECM_ERWCH PEDD_PEDAC PER1_HUMAN PER2_HUMAN PER3_BOVIN PER4_HUMAN PET1_HUMAN
PET2_RABIT PF2R_BOVIN PGSA_BACSU PI2R_HUMAN PIGF_HUMAN PIP_LACLA PKBS_BOVIN PLLP_RAT PM1_HUMAN PM22_HUMAN
PMA1_AJECA PMA2_ARATH PMA3_ARATH PMA4_NICPL PPA1_YEAST PRA1_USTMA PRA2_USTMA PRO1_LEIEN PSAA_SYNEN PSAB_SYNEN
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 PTTR_PIG PUR8_STRLP P_HUMAN QAY_NEUCR QOX1_BACSU QOX2_ACEAC QOXM_SULAC QUTD_EMENI
RAFP_PEDPE RAG1_KLULA RBS1_RAT RBSC_BACSU RCEL_CHLAU RCEM_CHLAU RDC1_CANFA RDS_BOVIN RDXA_RHOSH RFAL_ECOLI
RFE_ECOLI RGR_BOVIN RH50_HUMAN RHOM_DROME RH_HUMAN 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_BACLI
SENR_RAT SLY4_YEAST SNF3_YEAST SNQ2_YEAST SP5E_BACSU SPAB_BACSU 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_HUMAN TAP1_HUMAN TAP2_HUMAN TAT2_YEAST TCR2_BACSU TCRB_BACSU TCR_BACST TERC_ALCSP TH11_TRYBB TH12_TRYBB
TH2A_TRYBB THAS_HUMAN THRR_CRILO TIPW_LYCES TJ6_MOUSE TLR2_DROME TOK1_YEAST TRA2_CAEEL 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_CANFA TXKR_HUMAN UAPC_EMENI UL33_HCMVA UN17_CAEEL UN36_CAEEL UNC7_CAEEL US27_HCMVA US28_HCMVA
V1AR_HUMAN V1BR_HUMAN V28_HUMAN V2R_BOVIN VAL1_YEAST VC03_SPVKA VG74_HSVSA VGLB_HSVA1 VIPR_HUMAN VIPS_HUMAN
VK02_SPVKA VM11_YEAST VU51_HSV6U WC1B_ARATH WC1C_ARATH WHIT_DROME Y736_HAEIN YAG7_YEAST YG90_HAEIN YKH3_CAEEL
YMN2_CAEEL YNZ3_CAEEL YOPB_YEREN YOPD_YEREN YOR1_YEAST YRO2_YEAST YTP1_YEAST YZN4_CAEEL

(12) 25 vacuole proteins

ABRA_PLAFC ALEU_HORVU APE3_YEAST AVE3_AVESA CARP_YEAST CARV_CANAL CBPS_YEAST CBPY_CANAL CHLY_HEVBR CYS2_MAIZE
DP87_DICDI FAB1_YEAST GRA5_TOXGO INV1_LYCES INVA_PHAAU P34_SOYBN PPB_YEAST PR1A_TOBAC PR1B_TOBAC PR1C_TOBAC
PRTB_YEAST RAB4_DICDI SANT_PLAF7 SERA_PLAFG THGF_TOBAC

Appendix B

For the reader's convenience, let us prove that the covariance matrix $\mathbf{C}_\xi$ as defined by Equations 7 and 8 has no negative eigenvalues.

Suppose

$$\mathbf{B}_\xi = \mathbf{S}_\xi - \mathbf{x}^\xi\,\mathbf{e}^{\mathrm{T}} \qquad \text{(B1)}$$

where $\mathbf{S}_\xi$ is a $20 \times n_\xi$ matrix consisting of the n_ξ vectors of Equation 2 and $\mathbf{e}$ is the n_ξ-dimensional column vector with all components equal to 1. Then we have

$$\mathbf{C}_\xi = \mathbf{B}_\xi\,\mathbf{B}_\xi^{\mathrm{T}} \qquad \text{(B2)}$$

Suppose

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{20} \end{bmatrix} \qquad \text{(B3)}$$

is any real vector in the 20-D composition space. Left and right multiplying both sides of Equation B2 by $\mathbf{y}^{\mathrm{T}}$ and $\mathbf{y}$, respectively, we can obtain

$$\mathbf{y}^{\mathrm{T}}\mathbf{C}_\xi\,\mathbf{y} = \mathbf{y}^{\mathrm{T}}\mathbf{B}_\xi\,\mathbf{B}_\xi^{\mathrm{T}}\,\mathbf{y} = (\mathbf{B}_\xi^{\mathrm{T}}\,\mathbf{y})^{\mathrm{T}}(\mathbf{B}_\xi^{\mathrm{T}}\,\mathbf{y}) \geqslant 0 \qquad \text{(B4)}$$

Suppose $\mathbf{\Psi}$ is an eigenvector of $\mathbf{C}_\xi$, i.e.

$$\mathbf{C}_\xi\,\mathbf{\Psi} = \lambda\mathbf{\Psi} \qquad \text{(B5)}$$

where λ is the corresponding eigenvalue. Left multiplying both sides of the above equation by $\mathbf{\Psi}^{\mathrm{T}}$, we can obtain

$$\mathbf{\Psi}^{\mathrm{T}}\mathbf{C}_\xi\mathbf{\Psi} = \mathbf{\Psi}^{\mathrm{T}}\lambda\mathbf{\Psi} = \lambda\mathbf{\Psi}^{\mathrm{T}}\mathbf{\Psi} \qquad \text{(B6)}$$

Because Equation B4 and the fact that an eigenvector is a non-zero vector, it follows that

$$\lambda = \frac{\mathbf{\Psi}^{\mathrm{T}}\mathbf{C}_\xi\,\mathbf{\Psi}}{\mathbf{\Psi}^{\mathrm{T}}\mathbf{\Psi}} \geqslant 0 \qquad \text{(B7)}$$

This completes the proof.

Appendix C

Covariant discriminant values computed according to Equation 5 for the 37 proteins in the cytoskeleton subset of the dataset S^{12} (see Appendix A) and the subcellular location predicted for each of these proteins according to Equation 13

Protein Code	$F(\mathbf{X},\mathbf{X}^1)$	$F(\mathbf{X},\mathbf{X}^2)$	$F(\mathbf{X},\mathbf{X}^3)$	$F(\mathbf{X},\mathbf{X}^4)$	$F(\mathbf{X},\mathbf{X}^5)$	$F(\mathbf{X},\mathbf{X}^6)$	$F(\mathbf{X},\mathbf{X}^7)$	$F(\mathbf{X},\mathbf{X}^8)$	$F(\mathbf{X},\mathbf{X}^9)$	$F(\mathbf{X},\mathbf{X}^{10})$	$F(\mathbf{X},\mathbf{X}^{11})$	$F(\mathbf{X},\mathbf{X}^{12})$	Predicted location[a]
ABP1_SACEX	-47.94	-64.15	-146.20	-85.92	-108.27	165.30	247.34	-107.20	-121.34	275.88	-79.87	105.90	Cytoskeleton
CISY_TETTH	-132.25	-145.58	-144.84	-133.41	-141.52	47.94	-65.20	-126.04	-138.39	-45.97	-128.18	-16.30	Cytoplasm[b]
CP23_CHICK	127.78	40.91	-146.89	96.69	-77.75	2364.53	731.82	-71.49	-114.97	671.92	39.37	477.30	Cytoskeleton
CYLI_BOVIN	126.21	52.65	-142.02	96.16	-51.35	1071.65	607.89	84.43	-114.24	790.11	117.66	418.40	Cytoskeleton
NINL_DROME	-152.93	-149.19	-155.29	-124.59	-141.66	1827.02	-68.35	-133.57	-143.46	-74.25	-141.41	-60.09	Cytoskeleton
NINS_DROME	-152.23	-154.26	-157.30	-136.62	-142.12	43.53	-3.37	-138.97	-140.70	-34.93	-146.30	-90.72	Cytoskeleton
PAS5_PICPA	-159.05	-152.37	-157.89	-150.39	-146.36	0.97	-147.06	-150.08	-144.95	-116.34	-151.92	-86.53	Chloroplast[b]
REST_HUMAN	-88.59	-112.63	-158.12	-125.92	-127.06	120.00	196.66	-129.55	-135.67	233.20	-97.89	-102.93	Cytoskeleton
BNK_DROME	-108.04	-98.99	-147.25	-56.71	-124.88	1691.28	-36.96	-94.87	-133.94	398.65	-92.74	63.38	Cytoskeleton
CALD_CHICK	59.68	13.77	-144.92	52.52	-36.48	2129.80	1314.42	-62.52	-99.94	2690.52	52.09	197.40	Cytoskeleton
DCPY_NEUCR	-152.59	-155.03	-149.19	-116.95	-145.20	175.86	-127.13	-150.93	-137.74	-89.29	-150.50	29.99	Cytoplasm[b]
MYSA_CAEEL	-118.27	-130.66	-169.19	-129.50	-127.08	104.22	174.67	-131.87	-140.38	228.96	-115.72	-67.74	Cytoskeleton
MYSB_CAEEL	-112.50	-125.66	-167.74	-135.83	-128.29	230.45	165.19	-132.62	-142.70	230.48	-113.83	-7.88	Cytoskeleton
MYSC_CAEEL	-116.33	-127.13	-168.74	-126.12	-128.53	336.70	147.74	-132.38	-142.98	231.57	-111.15	-32.96	Cytoskeleton
MYSD_CAEEL	-124.24	-132.02	-165.77	-128.17	-127.81	315.20	141.46	-133.49	-142.42	193.17	-120.03	-70.99	Cytoskeleton
MYSE_CHICK	-114.32	-129.13	-167.24	-128.04	-126.21	239.42	224.74	-134.05	-138.76	191.64	-103.79	-111.10	Cytoskeleton
MYSG_CHICK	-100.50	-117.80	-163.04	-121.89	-122.65	671.17	197.11	-123.83	-137.23	354.04	-94.47	-29.91	Cytoskeleton
MYSP_CAEEL	-66.19	-97.04	-155.59	-98.51	-101.21	150.54	376.69	-95.79	-122.91	808.92	-68.22	80.27	Cytoskeleton
MYSQ_DROME	-109.26	-103.01	-144.83	-70.54	-109.28	209.46	171.84	-117.41	-128.23	793.90	-94.59	-30.23	Cytoskeleton
MYSS_CHICK	-114.51	-130.52	-166.92	-126.93	-125.36	254.38	233.14	-133.41	-137.83	230.13	-104.66	-107.36	Cytoskeleton
MYST_RABIT	-101.25	-119.31	-167.07	-125.66	-121.19	692.19	223.14	-123.73	-136.86	363.43	-93.16	-46.89	Cytoskeleton
MYS_AEQIR	-115.15	-128.80	-161.66	-135.49	-127.09	720.69	201.40	-131.71	-142.06	231.48	-115.34	-88.17	Cytoskeleton
N214_HUMAN	-112.55	-86.98	-148.84	24.12	-119.86	841.08	-95.19	-86.05	-128.80	772.87	-74.94	72.59	Cytoskeleton
N358_HUMAN	-149.99	-146.55	-160.33	-141.04	-148.18	-9.84	-107.93	-149.16	-146.03	-50.64	-145.40	-78.42	Cytoskeleton
NULL_DROME	-116.08	-97.40	-151.00	-75.97	-109.37	203.61	101.33	-57.90	-137.15	216.07	-92.97	170.22	Cytoskeleton
CIN8_YEAST	-112.35	-127.15	-154.71	-118.29	-129.71	3960.65	-30.20	-117.48	-141.52	41.35	-118.42	-110.67	Cytoskeleton
DYN1_CAEEL	-147.12	-146.38	-160.89	-115.34	-142.29	-141.55	-25.83	-135.74	-143.50	-105.15	-146.96	20.24	Cytoskeleton
DYN2_HUMAN	-145.92	-148.57	-163.28	-132.45	-139.85	290.81	-51.34	-137.82	-143.73	-107.88	-149.16	-118.79	Cytoskeleton
DYN3_RAT	-154.48	-153.85	-163.76	-146.50	-146.18	-160.84	-65.80	-146.86	-147.09	-124.82	-153.68	-130.92	Cytoskeleton
DYN_DROME	-153.60	-152.79	-158.21	-141.59	-143.05	-68.18	1.55	-141.93	-144.09	-91.26	-152.98	-46.43	Cytoskeleton
KCRF_STRPU	-154.39	-155.55	-154.36	-140.75	-149.35	-90.86	-130.59	-144.97	-143.99	-98.75	-144.56	-53.28	Cytoplasm[b]
KIP1_YEAST	-124.46	-125.30	-149.85	-114.16	-129.13	1271.24	-21.15	-120.65	-137.50	22.84	-120.95	-81.66	Cytoskeleton
KLP1_CHLRE	-142.32	-142.05	-149.19	-105.79	-137.18	891.50	-33.91	-138.46	-133.97	212.63	-138.80	80.58	Cytoskeleton
MAPX_DROME	-116.72	-120.74	-151.75	-80.47	-140.37	-144.22	-65.92	-125.92	-141.54	118.07	-120.46	24.60	Cytoskeleton
SCP1_MOUSE	-110.58	-107.42	-149.90	-125.45	-129.35	6389.11	-101.50	-99.78	-111.37	217.81	-126.40	-20.45	Cytoskeleton
SCP2_MOUSE	-107.85	-113.55	-148.74	-132.78	-129.78	8248.70	-80.76	-94.95	-104.77	210.15	-127.77	30.31	Cytoskeleton
VP22_ASFB7	-126.90	-121.66	-145.74	-130.68	-136.70	305.38	-57.75	-100.77	-120.23	141.96	-117.35	-67.32	Cytoskeleton

The rate of correct prediction for the proteins in the cytoskeleton subset in $S^{12} = 33/37 = 89.2\%$.

[a] The indices 1, 2, 3, . . ., 12 represent the 12 subcellular locations (Figure 1) as defined in the text. The index for cytoskeleton is 3; when $F(\mathbf{X},\mathbf{X}^3)$ is the minimum, the corresponding protein is predicted to be located in cytoskeleton. The index for cytoplasm is 2; when $F(\mathbf{X},\mathbf{X}^2)$ is the minimum, the corresponding protein is predicted to be located in cytoplasm. And so forth.

[b] Incorrect prediction.

Appendix D

Although the coupling effects among different amino acid components are taken into account by both the ProtLock algorithm (Cedano *et al*., 1977) and the current algorithm via a covariance matrix, there are two important differences between these two.

Difference in covariance matrix

Rather than $\mathbf{C}_\xi$ as defined by Equations 7 and 8, the covariance matrix in the ProtLock algorithm was given by

$$\mathbf{C} = \begin{bmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,20} \\ c_{2,1} & c_{2,2} & \cdots & c_{2,20} \\ \vdots & \vdots & \ddots & \vdots \\ c_{20,1} & c_{20,2} & \cdots & c_{20,20} \end{bmatrix} \tag{D1}$$

where

$$c_{i,j} = \sum_{\xi=1}^{m} \sum_{k=1}^{n_\xi} [x_{k,i}^{\xi} - \bar{x}_i]\,[x_{k,j}^{\xi} - \bar{x}_j] \quad (i,\, j = 1, 2, \ldots, 20) \tag{D2}$$

where

$$\bar{x}_i = \frac{1}{N} \sum_{\xi=1}^{m} \sum_{k=1}^{n_\xi} x_{k,i}^{\xi} = \frac{1}{N} \sum_{\xi=1}^{m} n_\xi\, x_i^{\xi} \quad (i = 1, 2, \ldots, 20) \tag{D3}$$

Comparing Equation D1 with Equation 7, Equation D2 with Equation 8 and Equation D3 with Equation 4, one can easily see that there was only one covariance matrix **C** in ProtLock that was defined for the entire set *S*, rather than each of the *m* subsets G_ξ ($\xi = 1, 2, 3, \ldots, m$) having its own covariance matrix $\mathbf{C}_\xi$. Accordingly, the Mahananobis distance defined in ProtLock is a simplified form of the genuine Mahalanobis distance. This will certainly make the ProtLock algorithm lose some power in discriminating entries from different subsets.

It is instinctive to point out that the covariance matrix (Equation D1) given by Cedano *et al*. (1997) was defined in a 20-D space rather than 19-D space as originally formulated by K.C.Chou (1995). As mentioned in the prediction algorithm section, this would lead to a divergent difficulty when calculating the Mahalanobis distance in terms of the inverse matrix of **C** unless the user understood the use of the eigenvalue–eigenvector approach as described in this paper to avoid such a difficulty.

Difference in discriminative criterion

The prediction in ProtLock was based on Mahananobis distance as defined by

$$D_{\xi}^{2}(\mathbf{X}, \mathbf{X}^{\xi}) = (\mathbf{X} - \mathbf{X}^{\xi})^{\mathrm{T}} \mathbf{C}^{-1} (\mathbf{X} - \mathbf{X}^{\xi}) \quad (\xi = 1, 2, 3, \ldots) \tag{D4}$$

In contrast, the prediction in the current algorithm is based on the covariant discriminant function given by Equation 5. A comparison of Equation 5 with Equation D4 indicates that the contribution from the term $\ln(\lambda_2^{\xi}\, \lambda_3^{\xi}\, \lambda_4^{\xi} \ldots \lambda_{20}^{\xi})$, which reflects the difference of the covariance matrices $\mathbf{C}_\xi$ for different classes, was completely ignored in the ProtLock algorithm. This will further weaken the power of discriminativity.