PROTEINS: Structure, Function, and Genetics 50:44–48 (2003)

Subcellular Location Prediction of Apoptosis Proteins

Guo-Ping Zhou[1]* and Kutbuddin Doctor[2]
[1]*Department of Medicine, Harvard Medical School, Boston, Massachusetts*
[2]*Department of Medicine, The Burnham Institute, La Jolla, California*

ABSTRACT **Apoptosis proteins have a central role in the development and homeostasis of an organism. These proteins are very important for understanding the mechanism of programmed cell death. Many efforts in pharmaceutical research have been aimed at understanding their structure and function. Unfortunately, thus far, very few apoptosis protein structures have been determined. In contrast, many apoptosis protein sequences are known, and many more are expected to come in the near future. Because of the extremely unbalanced state, it would be worthwhile to develop a fast sequence-based method to identify their subcellular location so as to gain some insight about their biological function. In view of this, a study was initiated in an attempt to identify the subcellular location of apoptosis proteins according to their sequences by means of the covariant discriminant function, which was established based on the Mahalanobis distance and Chou's invariance theorem (Chou, Proteins 1995;21:319–344). The results were quite promising, indicating that the subcellular location of apoptosis proteins are predictable to a considerably accurate extent if a good training data set can be established. It is expected that, with a continuous improvement of the training data set by incorporating more and more new data, the current method might eventually become a useful tool in this area because the function of an apoptosis protein is closely related to its subcellular location. Proteins 2003;50:44–48.** © 2002 Wiley-Liss, Inc.

Key words: covariant discriminant algorithm; amino acid composition; dimension-reducing procedure; Mahalanobis distance; Chou's invariance theorem

INTRODUCTION

Apoptosis, or programmed cell death, is a fundamental process controlling normal tissue homeostasis by regulating a balance between cell proliferation and death.[1–8] This process entails the autolytic degradation of cellular components, and is characterized by blebbing of cell membranes, shrinkage of cell volumes, and condensation of nuclei,[9] and is currently an area of intense investigation. Cell death and renewal are responsible for maintaining the proper turnover of cells, which ensures a constant controlled flux of fresh cells. Programmed cell death and cell proliferation are tightly coupled. When apoptosis malfunctions, a variety of formidable diseases can ensue: blocking apoptosis is associated with cancer[10,11] and autoimmune disease, whereas unwanted apoptosis can possibly lead to ischemic damage[12] or neurodegenerative disease.[13] Apoptosis is considered to have a key role in these several devastating diseases and, in principle, provides many targets for therapeutic intervention.[3,4,14]

To understand the apoptosis mechanism and functions of various apoptosis proteins, it would be helpful to obtain information about their subcellular location. This is because the subcellular location of apoptosis proteins is closely related to their function (see, e.g., Refs.12, 15, 16). It has been known that there are 732 archetypical proteins with "apoptosis" domains. Recently, in one of our ongoing projects, we were dealing with a number of protein sequences already known belonging to apoptosis proteins. However, it is both time-consuming and costly to determine which specific subcellular location a given apoptosis protein belongs to. Confronted with such a situation, can we develop a fast and effective way to predict the subcellular location for a given apoptosis protein based on its amino acid sequence? The present study was initiated in an attempt to address this problem.

MATERIALS AND METHODS

As demonstrated in an incisive analysis by Chou,[17] because of the extremely large number of possible protein sequences, it is almost unrealistic to develop a statistical method purely based on the input of sequences without adopting any simplifying procedures. One of the effective simplifying procedures is using the amino acid composition. The idea of predicting the subcellular location of a protein according to its amino acid composition is based on the following rationales.

1. Different compartments of a cell usually have different physiochemical environments that might be very sensitive in selectively accommodating a protein according to its structural feature, particularly its surface physical chemistry character.[18]
2. 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.[19–27]
3. The character of a protein surface, which is directly

*Correspondence to: Guo-Ping Zhou, Harvard Medical School, Boston, MA 02115. E-mail: gzhou@caregroup.harvard.edu

Received 20 June 2002; Accepted 26 July 2002

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 have an important role.[23,28]

4. The above correlations suggest that the total amino acid composition might carry a "quasi-signal" that identifies the subcellular location.[29]

Actually, many efforts have been made to predict both the structural class and subcellular location of a protein according to its amino acid composition (see, e.g., Refs.19–21, 25, 26, 30–33). Nakashima and Nishikawa[31] first proposed an algorithm to discriminate between intracellular and extracellular proteins by amino acid composition and residue-pair frequency. Three years later, Cedano et al.[32] extended the discriminative classes from two to five, i.e., extracellular, integral membrane, anchored membrane, intracellular, and nuclear. Furthermore, in an attempt to improve the prediction quality of protein cellular location, they proposed an algorithm called Protlock. The Protlock algorithm[32] is mainly based on the procedure reported by Chou and Zhang[21] for the prediction of protein structural classes according to Mahalanobis distances. Because the least Mahalanobis distance algorithm[21,22] is valid only when the training subset sizes are the same or approximately the same, or poor predictions will otherwise result.[34,35] In this report, we shall adopt the covariant discriminant algorithm developed by Chou and his coworkers.[34,36,37] For reader's convenience, a brief introduction about the covariant discriminant algorithm is provided below.

Suppose the kth apoptosis protein in the class m is represented by the following vector:

$$\mathbf{A}_k^m = \begin{bmatrix} a_{k,1}^m \\ a_{k,2}^m \\ \vdots \\ a_{k,20}^m \end{bmatrix}, (k = 1, 2, \cdots, n_m; \quad m = 1, 2, \cdots) \quad (1)$$

where $a_{k,1}^m$, $a_{k,2}^m$, ..., $a_{k,20}^m$ are the amino acid composition[22,24] for the kth apoptosis protein of class m, and n_m the total number of apoptosis proteins in class m. The standard vector for class m is defined by[22]:

$$\overline{\mathbf{A}}^m = \begin{bmatrix} \bar{a}_1^m \\ \bar{a}_2^m \\ \vdots \\ \bar{a}_{20}^m \end{bmatrix}, \quad (m = 1, 2, \cdots) \quad (2)$$

where

$$\bar{a}_i^m = \frac{1}{n_m} \sum_{k=1}^{n_m} a_{k,i}^m, (i = 1, 2, \cdots, 20). \quad (3)$$

Suppose $\mathbf{A}$ is a query apoptosis protein whose cellular location is to be identified. It can also be represented by a point or vector in the 20-D space with the components of (a_1, a_2, ..., a_{20}), where a_i has the same meaning as $a_{k,1}^m$ of Eq. (1) but is associated with protein $\mathbf{A}$ instead of $\mathbf{A}_k^m$. The difference between the query protein $\mathbf{A}$ and the norm of class m is measured by the following covariant discriminant function, as defined by Chou et al.[34]:

$$\Delta(\mathbf{A},\overline{\mathbf{A}}^m) = D_M^2(\mathbf{A},\bar{\mathbf{A}})^m) + \ln|\mathbf{S}^m|, \quad (m = 1, 2, \cdots), \quad (4)$$

where

$$D_{\mathbf{M}}^2(\mathbf{A},\bar{\mathbf{A}})^m) = (\mathbf{A} - \overline{\mathbf{A}}^m)^T \mathbf{S}_m^{-1} (\mathbf{A} - \overline{\mathbf{A}}^m) \quad (5)$$

is the squared Mahalanobis distance,[22,38,39] $\mathbf{T}$ is the transposition operator, and $|\mathbf{S}^m|$ and $\mathbf{S}_m^{-1}$ are respectively the determinant and inverse matrix of $\mathbf{S}_m$. The latter is the covariance matrix for class m and defined by:

$$\mathbf{S}_m = \begin{bmatrix} s_{1,1}^m & s_{1,2}^m & \cdots & s_{1,20}^m \\ s_{2,1}^m & s_{2,2}^m & \cdots & s_{2,20}^m \\ \vdots & \vdots & \ddots & \vdots \\ s_{20,1}^m & s_{20,2}^m & \cdots & s_{20,20}^m \end{bmatrix} \quad (6)$$

where the matrix elements are given by:

$$S_{i,j}^m = \frac{1}{n_m - 1} \sum_{k=1}^{n_m} [a_{k,i}^m - \bar{a}_i^m][a_{k,j}^m - \bar{a}_j^m], \quad (i,j = 1, 2, \ldots, 20). \quad (7)$$

According to the principle of similarity, the smaller the difference between the query apoptosis protein $\mathbf{A}$ and the norm of class m, the higher the probability that the protein $\mathbf{A}$ belongs to class m. Accordingly, the identification rule can be formulated as follows:

$$\Delta(\mathbf{A},\overline{\mathbf{A}}^\mu) = \mathbf{Min}\{\Delta(\mathbf{A},\overline{\mathbf{A}}^1), \Delta(\mathbf{A},\overline{\mathbf{A}}^2), \Delta(\mathbf{A},\overline{\mathbf{A}}^3), \cdots\}, \quad (8)$$

where μ can be 1, 2, 3, ..., and the operator **Min** means taking the minimal one among those in the brackets. The value of the superscript μ derived from Eq. (8) indicates which class the query apoptosis protein $\mathbf{A}$ belongs to. If there is a tie case, μ is not uniquely determined, but that did not happen for the data sets studied here.

Before using the above equations for practical calculations, we draw the reader's attention to the following point. Because of the normalization condition imposed on amino acid composition, of the 20 components in Eq. (1), only 19 are independent,[22] and hence the covariance matrix $\mathbf{S}_m$ as defined by Eq. (7) must be a singular one.[24] This implies that the Mahalanobis distance defined by Eq. (5) and the covariant discriminant function by Eq. (4) would be divergent and meaningless. To overcome such a difficulty, the dimension-reducing procedure[22] was adopted in practical calculations; i.e., instead of 20-D space, a protein is defined in a (20-1)-D space by leaving out one of its 20 amino acid components. The remaining 19 components would be completely independent and hence the corresponding covariance matrix $\mathbf{S}_m$ no longer singular. In such a 19-D space, the Mahalanobis distance (Eq. 5) and covariant the discriminant function (Eq. 4) can be defined without the divergence difficulty. However, which one of the 20 components can be left out during the dimension-reducing procedure? The answer is: any one of them. Will it lead to a different predicted result by

TABLE I. List of the Accession Numbers for the 98 Apoptosis Proteins Classified Into Four Categories According to Their Subcellular Location†

(1) 43 Cytoplasmic proteins			
XP_013050	P55212	P42574	P39429
P55867	P55865	Q02357	NP_033941
NP_033940	NP_033939	NP_031637	NP_031570
NP_031563	NP_031490	NP_033447	P29452
NP_036246	NP_001218	NP_004041	O54786
Q60989	Q62210	NP_065209	NP_001151
NP_071610	NP_071567	NP_066961	NP_037054
NP_036894	NP_005649	NP_004392	NP_004315
NP_001187	NP_001159	NP_001157	NP_001156
P22366	P55866	Q60431	P55214
P55269	P29466	O70201	
(2) 30 Plasma membrane-bound proteins			
NP_037223	P28825	NP_037275	NP_032013
NP_032612	P50555	P25118	P18519
O19131	Q63199	O77736	P51867
NP_036742	NP_037315	NP_005916	NP_005579
NP_000034	NP_001056	NP_003781	NP_002498
O02703	Q13014	NP_031553	NP_031549
Q63690	Q07820	NP_001179	Q91828
Q91827	Q07812		
(3) 13 Mitochondrial inner and outer proteins			
XP_008738	O77737	Q00709	NP_033873
P10417	P53563	Q07816	P49950
Q07817	O95831	Q9OX1	Q9JM53
Q9VQ79			
(4) 12 Other proteins[a]			
Q63369	Q90660	Q00653	Q04861
P19838	NP_032715	P98150	Q15121
Q62048	NP_033872	NP_004040	NP_005736

†Derived from SWISS-PROT data bank.[40]
[a]Of the 12 other apoptosis proteins, five are located in nucleus, two in endoplasmic reticulum, one in microtubule, and one in lysosome.

leaving out a different component? According to Chou's invariance theorem (see Appendix A of Chou[22]), the value of the Mahalanobis distance as well as the value of the determinant of $\mathbf{S}_m$ will remain exactly the same regardless of which one of the 20 components is left out. Therefore, the value of the covariant discriminant function (Eq. 4) can be uniquely defined through such a dimension-reducing procedure.

In addition to the prediction algorithm, we also need to construct a training data set to complete the establishment of a statistical prediction method. To realize this, based on the SWISS-PROT data bank,[40] 98 apoptosis proteins were classified into the following four subcellular locations: (1) cytoplasmic, (2) plasma membrane-bound, (3) mitochondrial, and (4) other (Table I).

RESULTS AND DISCUSSION

By means of the covariant-discriminant algorithm described in the last section, a statistical prediction was performed for the 98 apoptosis proteins listed in Table I. The prediction was conducted by two different approaches, the re-substitution test and the jackknife test. The results are given in Table II.

Re-Substitution Test

The so-called re-substitution test is an examination for the self-consistency of a prediction method. When the re-substitution test was performed for the current study, the type of each apoptosis protein in a data set was in turn identified using the rule parameters derived from the same data set, the so-called training data set. As shown in Table II, the overall success rate thus obtained for the 98 apoptosis proteins in Table I was 90.8%, indicating an excellent self-consistency. However, during the process of the re-substitution test, the rule parameters derived from the training data set include the information of the query protein later plugged back in the test. This will certainly underestimate the error and enhance the success rate because the same proteins are used to derive the rule parameters and to test themselves. Accordingly, the success rate thus obtained represents some sort of optimistic estimation.[22,24,26,27] Nevertheless, the re-substitution test is absolutely necessary because it reflects the self-consistency of a prediction method, especially for its algorithm part. A prediction algorithm certainly cannot be deemed as a good one if its self-consistency is poor. In other words, the re-substitution test is necessary but not sufficient for evaluating a prediction method. As a complement, a cross-validation test for an independent testing data set is needed because it can reflect the effectiveness of a prediction method in practical application. This is important especially for checking the validity of a training data set—whether it contains sufficient information to reflect all the important features concerned so as to yield a high success rate in application.

Jackknife Test

As is well known, the independent data set test, subsampling test, and jackknife test are the three methods often used for cross-validation in statistical prediction. Among these three, however, the jackknife test is deemed as the most effective and objective one (see Chou and Zhang[21] for a comprehensive discussion about this, and Mardia et al.[41] for the mathematical principle). During jackknifing, each protein in the data set is in turn singled out as a tested protein and all the rule parameters are calculated based on the remaining proteins. In other words, the subcellular location of each apoptosis protein is identified by the rule parameters derived using all the other apoptosis proteins except the one that is being identified. During the process of jackknifing, both the training data set and testing data set are actually open, and a protein will in turn move from one to the other. As expected, the success prediction rates by jackknife test were decreased compared with those by the re-substitution test. Such a decrement is particularly more remarkable for small subsets. This is because the cluster-tolerant capacity[28] for small subsets is usually low. And hence the information loss resulting from jackknifing will have a

TABLE II. Tested Results for the 98 Apoptosis Proteins in Table I by Both Re-Substitution Test and Jackknife Test

	Success rate				
Test method	Cytoplasmic	Membrane-bound	Mitochondrial	Other	Overall
Re-substitution	43/43 = 100%	30/30 = 100%	9/13 = 60.2%	7/12 = 58.3%	89/98 = 90.8%
Jackknife	42/43 = 97.7%	22/30 = 73.3%	4/13 = 30.8%	3/12 = 25.0%	71/98 = 72.5%

greater impact on the small subsets than the large ones. Nevertheless, as shown in Table II, the overall jackknife rate for the data set of the 98 apoptosis proteins could still reach 72.5%. It is expected that the success rate for identifying the subcellular location of apoptosis proteins can be further enhanced by improving the training data of small subsets by adding into them more new proteins that have been found belonging to the subcellular location defined by these subsets.

CONCLUSIONS

If the samples of apoptosis proteins are completely randomly distributed among the four subsets, the rate of correct identification by random assignment would generally be 1/4 = 25%; if the distribution is weighted according to the sizes of subsets, then the rate of correct identification by the weighted random assignment would be $(43/98)^2 + (30/98)^2 + (13/98)^2 + (12/98)^2 \approx 30.4\%$. Therefore, the rates of correct identification obtained based on the amino acid composition in both the re-substitution and jackknife tests are much higher than the corresponding completely randomized rate and weighted randomized rate, implying that the subcellular location of apoptosis proteins is considerably correlated with the amino acid composition. This suggests that their subcellular locations are predictable to a considerably accurate extent if a complete or quasi-complete training data set can be established for that purpose. Establishment of such a fast identification tool will certainly speed up the pace of the studies in the apoptosis and related areas.

ACKNOWLEDGMENTS

We thank Dr. D. Kedra for converting the SWISS-PROT codes of the apoptosis proteins into their accession numbers.

REFERENCES

1. Vaux DL, Heacker G, Strasser A. An evolutionary perspective on apoptosis. Cell 1994;76:777–779.
2. Jacobson MD, Weil M, Raff MC. Programmed cell death in animal development. Cell 1997;88:347–354.
3. Chou JJ, Matsuo H, Duan H, Wagner G. Solution structure of the RAIDD CARD and model for CARD/CARD interaction in caspase-2 and caspase-9 recruitment. Cell 1998;94:171–180.
4. Chou JJ, Li H, Salvesen GS, Yuan J, Wagner G. Solution structure of BID, an intracellular amplifier of apoptotic signaling. Cell 1999;96:615–624.
5. Chou KC, Tomasselli AG, Heinrikson RL. Prediction of the tertiary structure of a caspase-9/inhibitor complex. FEBS Lett 2000;470:249–256.
6. Chou KC, Jones D, Heinrikson RL. Prediction of the tertiary structure and substrate binding site of caspase-8. FEBS Lett 1997;419:49–54.
7. Lugovskoy AA, Zhou P, Chou JJ, McCarty JS, Li P, Wagner G. Solution structure of the CIDE-N domain of CIDE-B and a model for CIDE-N/CIDE-N interactions in the DNA fragmentation pathway of apoptosis. Cell 1999;99:747–755.
8. Zhou P, Chou JJ, Olea RS, Yuan J, Wagner G. Solution structure of Apaf-1 CARD and its interaction with caspase-9 CARD: a structural basis for specific adaptor/caspase interaction. Proc Natl Acad Sci USA 1999;96:11265–11270.
9. Kerr JF, Wyllie AH, Currie AR. Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics. Br J Cancer 1972;26:239–257.
10. Evan G, Littlewood T. A matter of life and cell death. Science 1998;281:1317–1322.
11. Adams JM, Cory S. The Bcl-2 protein family: arbiters of cell survival. Science 1998;281:1322–1326.
12. Reed JC, Paternostro G. Postmitochondrial regulation of apoptosis during heart failure. Proc Natl Acad Sci USA 1999;96:7614–7616.
13. Schulz JB, Weller M, Moskowitz MA. Caspases as treatment targets in stroke and neurodegenerative diseases. Ann Neurol 1999;45:421–429.
14. Barinaga M. Stroke-damaged neurons may commit cellular suicide. Science 1998;281:1302–1303.
15. Suzuki M, Youle RJ, Tjandra N. Structure of Bax: coregulation of dimer formation and intracellular location. Cell 2000;103:645–654.
16. Wolter KG, Hsu YT, Smith CL, Nechushtan A, Xi XG, Youle RJ. Movement of Bax from the cytosol to mitochondria during apoptosis. J Cell Biol 1997;139:1281–1292.
17. Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 2001;43:246–255 (Erratum: Proteins 2001;44:60).
18. Andrade MA, O'Donoghue SI, Rost B. Adaptation of protein surfaces to subcellular location. J Mol Biol 1998;276:517–525.
19. Chou PY. Amino acid composition of four classes of proteins. Abstracts of papers, Part I. Second Chemical Congress of the North American Continent, Las Vegas; 1980.
20. Chou PY. Prediction of protein structural classes from amino acid composition. In: Fasman GD, editor. Prediction of protein structure and the principles of protein conformation. New York: Plenum Press; 1989. p 549–586.
21. Chou KC, Zhang CT. Prediction of protein structural classes. Crit Rev Biochem Mol Biol 1995;30:275–349.
22. Chou KC. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins 1995;21:319–344.
23. Bahar I, Atilgan AR, Jernigan RL, Erman B. Understanding the recognition of protein structural classes by amino acid composition. Proteins 1997;29:172–185.
24. Chou KC, Zhang CT. Predicting protein folding types by distance functions that make allowances for amino acid interactions. J Biol Chem 1994;269:22014–22020.
25. Zhou GP. An intriguing controversy over protein structural class prediction. J Protein Chem 1998;17:729–738.
26. Zhou GP, Assa-Munt N. Some insights into protein structural class prediction. Proteins 2001;44:57–59.
27. Cai YD. Is it a paradox or misinterpretation? Proteins 2001;43:336–338.
28. Chou KC. A key driving force in determination of protein structural classes. Biochem Biophys Res Commun 1999;264:216–224.
29. Chou KC, Elrod DW. Protein subcellular location prediction. Protein Eng 1999;12:107–118.
30. Chou KC. Review: Prediction of protein structural classes and subcellular locations. Curr Protein Peptide Sci 2000;1:171–208.
31. Nakashima H, Nishikawa K. Discrimination of intracellular and

extracellular proteins using amino acid composition and residue-pair frequencies. J Mol Biol 1994;238:54–61.
32. Cedano J, Aloy P, Perez-Pons JA, Querol E. Relation between amino acid composition and cellular location of proteins. J Mol Biol 1997;266:594–600.
33. Reinhardt A, Hubbard T. Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res 1998;26: 2230–2236.
34. Chou KC, Liu W, Maggiora GM, Zhang CT. Prediction and classification of domain structural classes. Proteins 1998;31:97–103.
35. Chou KC, Maggiora GM. Domain structural class prediction. Protein Eng 1998;11:523–538.
36. Liu W, Chou KC. Prediction of protein structural classes by modified Mahalanobis discriminant algorithm. J Protein Chem 1998;17:209–217.
37. Chou KC, Elrod DW. Prediction of membrane protein types and subcellular locations. Proteins 1999;34:137–153.
38. Mahalanobis PC. On the generalized distance in statistics. Proc Natl Inst Sci India 1936;2:49–55.
39. Pillai KCS. Mahalanobis D2. In: Kotz S, Johnson NL, editors. Encyclopedia of statistical sciences. Vol. 5. New York: John Wiley & Sons; 1985. p 176–181.
40. Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL. Nucleic Acids Res 2000;25:31–36.
41. Mardia KV, Kent JT, Bibby JM. Multivariate analysis. London: Academic Press; 1979. p 322 and 381.