Amino Acids (2007) 33: 669–675
DOI 10.1007/s00726-006-0465-0
Printed in The Netherlands

Amino Acids

Prediction of mitochondrial proteins based on genetic algorithm – partial least squares and support vector machine

F. Tan, X. Feng, Z. Fang, M. Li, Y. Guo, and **L. Jiang**

College of Chemistry, Sichuan University, Chengdu, China

Received September 21, 2006
Accepted October 15, 2006
Published online August 15, 2007;

Summary. Mitochondria are essential cell organelles of eukaryotes. Hence, it is vitally important to develop an automated and reliable method for timely identification of novel mitochondrial proteins. In this study, mitochondrial proteins were encoded by dipeptide composition technology; then, the genetic algorithm-partial least square (GA-PLS) method was used to evaluate the dipeptide composition elements which are more important in recognizing mitochondrial proteins; further, these selected dipeptide composition elements were applied to support vector machine (SVM)-based classifiers to predict the mitochondrial proteins. All the models were trained and validated by the jackknife cross-validation test. The prediction accuracy is 85%, suggesting that it performs reasonably well in predicting the mitochondrial proteins. Our results strongly imply that not all the dipeptide compositions are informative and indispensable for predicting proteins. The source code of MATLAB and the dataset are available on request under liml@scu.edu.cn.

Keywords: Mitochondrial proteins – Dipeptide composition – Genetic algorithm-partial least square – Support vector machine

1. Introduction

Mitochondria carry out a wide variety of biochemical processes within the eukaryotic cell. They fulfill most of the energy requirements of aerobic cells and are essential for the metabolism of a number of important biological compounds (Scharfe et al., 2000). Of 1331 identified human disease proteins, a proportion of ∼10% (129) of the known disease proteins have been known to be localized in mitochondria (Andreoli et al., 2004). Meanwhile, in recent years, there has been an unprecedented increase in the production of mitochondria sequences. Unfortunately, most of these sequences will remain of no avail to the understanding of their biological significance until properly analyzed.

The contradictions between their critical roles in a variety of complex biochemical processes and the deposition in the public data bank have made it a crucial issue to predict mitochondrial proteins. Most of existing prediction methods fall into two categories: one trends to combine several source of biological information for prediction, the other is based on the sequence composition incorporating some pattern recognition or machine learning methods. Some prevalent approaches in the first category include Target P (Emanuelsson et al., 2000), SignalP 3.0 (Bendtsen et al., 2004), WoLF PSORT (Horton et al., 2006), TargetLoc (Höglund et al., 2006), MitoProt II (Claros and Vincens, 1996), MITOPRED (Guda et al., 2004) and so on. Although these methods are popularly used, these algorithms rely strongly on the existence of leader sequences. Proteins localized in the same organelle have been reported to show a similar overall amino acid composition (Andrade et al., 1998). It means that a method based on the amino acid or dipeptide composition would be possible and more useful in practical applications. A number of different computational approaches based on amino acid or dipeptide compositions have been presented, including the covariant discriminant algorithm (Chou and Elrod, 1999), Markov chain models (Yuan, 1999), support vector machine (Hua and Sun, 2001; Chou and Cai, 2002; Kumar, 2006), fuzzy K-NN method (Huang and Li, 2004), etc. Because of the special properties of mitochondrial proteins, the prediction accuracy is much lower than proteins in other locations.

To approximately incorporate the sequence-order effects (Chou, 2000a), the concept of the pseudo amino acid composition (PseAA) was proposed (Chou, 2001, 2005a, b) and has been used via various approaches to enhance the prediction quality (Chou and Cai, 2003; Chou and Shen, 2006a; Gao et al., 2005; Xiao et al., 2005a, 2006a). Recently, two very powerful predictors were developed.

One is Hum-PLoc (Chou and Shen, 2006b, c), which can identify mitochondrion proteins among 12 human protein subcellular locations. And the other is Euk-OET-PLoc (Chou and Shen, 2006d), which can identify mitochondrion proteins among 16 eukaryotic protein subcellular locations.

The present study was initiated in order to develop an integrative method for recognizing mitochondria proteins on the base of dipeptide composition. First, the inquired sequence was transformed to numeric series by dipeptide composition technology; then, we applied the GA-PLS method to extract the more important dipeptide composition elements; finally, the selected dipeptide composition elements were introduced as input to construct SVM models. The results suggest GA-PLS is an effective tool for extracting feature dipeptide composition elements and not all the dipeptide compositions are informative in predicting proteins.

2. Materials and methods

2.1 Data sets

The dataset used in this paper was generated in our previous work (Jiang et al., 2006). All the sequences were extracted from Swiss-Prot release 46.6 (Boeckmann et al., 2003) by the keyword mitochondrial and 2833 entries were obtained. All sequences with ambiguous words, such as POTENTIAL, BY SIMILARITY, or PROBABLE and fragments were all excluded. The final dataset consisted of 499 entries that appeared as whole sequences and had reliable experimental annotations for localization. They consisted of a credible mitochondrial database and were used as positive samples. Meanwhile 681 entries were also extracted by selected one out of every 250 entries in Swiss-Prot and mitochondrial protein sequences or fragments were all deleted. They were used as negative samples. Some sequences with high identity of 90% were not removed in order to provide a wide range prediction, while most sequences were clustered lower than 20% identity using Clustal W program (Thompson et al., 1994).

2.2 Dipeptide composition

The dipeptide composition (Liu and Chou, 1999) and pair-coupled amino acid composition (Chou, 1999) have been successfully used to predict protein secondary structure contents. The dipeptide composition used as input can provide global information on protein features in the form of fixed-length vector. It is calculated as follows for each protein:

$$\mathrm{Fdip(i)} = \frac{\text{total number of dip(i)}}{\text{total number of all dipeptides}} \quad (1)$$

where Fdip(i) is the fraction of dip(i) that is the *i*th dipeptide out of 400 dipeptides.

Compared with amino acid composition, the advantage of dipeptide composition is that it incorporates some sequence-order information. With dipeptide composition coding scheme, each protein was represented as a fixed pattern length of 400 elements.

2.3 Performance evaluation

The jackknife test has been considered as one of the most objective test methods in examining the power of a prediction method, as illustrated in a comprehensive review article (Chou and Zhang, 1995). It has been adopted by more and more leading investigators to test the powers of various predictors (see, e.g., Cai and Chou, 2005; Chou, 1995, 2000b, 2005c; Chou and Cai, 2004a, b, 2006; Chou and Elrod, 1998, 2002, 2003; Chou et al., 1998; Chou and Maggiora, 1998; Gao et al., 2005; Guo et al., 2006; Liu et al., 2005; Shen and Chou, 2005a, b, 2006; Shen et al., 2005, 2006; Sun and Huang, 2006; Wen et al., 2006; Xiao et al., 2005b, 2006b, c; Zhang et al., 2006; Zhou, 1998; Zhou and Assa-Munt, 2001; Zhou and Doctor, 2003). In this paper, a jackknife procedure was carried out. All the protein sequences in the training dataset were in turn singled out as a 'tested protein' and all the rule parameters were determined from the remaining proteins. Meanwhile each non-mitochondrial protein sequence in test dataset was also compared with the 499 mitochondrial sequences. Four parameters were employed to measure the performance of models, including Acc, SE, SP, and MCC. Details of these indices are listed in Table 1 (Liu et al., 2006).

Table 1. Indices introduced to evaluate the dipeptide composition-based support vector machine method using jackknife test

Index	Formula
Acc	$(TP + TN)/(TP + TN + FP + FN)$
Sen	$TP/(TP + FN)$
Sp	$TP/(TP + FP)$
MCC	$(TP \cdot TN - FN \cdot FP) \Big/ \sqrt{(TP + FN) \cdot (TP + FP) \cdot (TN + FN) \cdot (TN + FP)}$

Acc Total accuracy; *Sen* sensitivity; *Sp* specificity; *MCC* Matthews's correlation coefficient. *TP (true positive)* The number of observed positive samples, predicted positive samples; *TN (true negative)* the number of observed negative samples, predicted negative samples; *FP (false positive)* the number of observed negative samples, predicted positive samples; *FN (false negative)* the number of observed positive samples, predicted negative samples

2.4 Support vector machine

SVM is a statistical learning theory based on machine learning algorithm presented by Vapnik (1998). A brief and clear description for how to use SVM to do classification also has been given by Chou and Cai (2002) and Cai et al. (2003). In this particular work, the mitochondrial proteins were defined as one class (labeled as $+1$) and the non-mitochondrial proteins were defined as the other one (labeled as -1) while radial basis function (RBF) was selected as the kernel function and quadratic programming (QP) method was introduced to solve the optimization problem. All the parameters were kept constant except for C (regulatory parameter) and σ (kernel width parameter). In the training process, C and σ were optimized (Guo et al., 2006). The fixed length feature vector was obtained using dipeptide composition.

2.5 Genetic algorithm and partial least square

GA-PLS is a sophisticated hybrid approach that combines GA as a powerful optimization method with PLS as a robust statistical method for variable selection. GA is inspired by the biological concept of natural selection and evolution. Just as the most fit organisms are most likely to survive and be reproduced by crossover together with random mutations of chromosomes in the surviving ones. In GA-PLS, the chromosome and its fitness in the species correspond to a set of variables and internal prediction of the derived PLS model, respectively (Hasegawa et al., 1999). In this study, the GA-PLS programs were implemented using the freely downloadable software package PLS_Genetic Algorithm Toolbox

Table 2. Parameters of the genetic algorithm (Leardi, 2000)

Population size: 30 chromosomes
On average, five variables per chromosome in the original population
Regression method: PLS
Response: cross-validated % explained variance (five deletion groups; the number of components is determined by cross-validation)
Maximum number of variables selected in the same chromosome: 30
Probability of mutation: 1%
Maximum number of components: the optimal number of components determined by cross-validation on the model containing all the variables (not higher than 15)
Number of runs: 100
Backward elimination after every 100th evaluation and at the end (if the number of evaluations is not a multiple of 100)

written by Leardi (Leardi and Lupiáñez, 1998). The values of empirical parameters affecting the performance of GA-PLS were defined as Table 2 (Leardi, 2000). The exploration of possible variable combinations was then done by PLS models. Selection of useful variables was based on their frequency of occurrence in the best models obtained for each program.

3. Results and discussion

3.1 Prediction procedure

Recently, Kumar et al. (2006) have developed dipeptide composition-based method to predict mitochondrial proteins and achieved sound accuracy. This work expects to predict mitochondrial proteins with higher accuracy while using as fewer dipeptide composition elements as possible. GA is a powerful optimization method in selecting the most relevant molecular descriptors for database mining in biochemistry (Ros et al., 2002). Compared with other optimization methods, the genetic algorithm selects features themselves other than their assemblage. This makes it possible to progress further study of protein structure and biological function on the base of these selected dipeptides.

Since the GA is mainly a stochastic algorithm, the results of different GA applications can therefore be slightly different. In order to get more consistent results, the GA process needs to repeat many times to give a more reliable model. In this particular work, GA process was finally repeated 500 times and the selection of useful variables was based on their frequency of occurrence in the models with the maximal C.V. % (Cross-validated explained variance) obtained for each operation. The frequency was calculated by Eq. (2). The frequency of each dipeptide composition occurrence in GA-PLS model is listed in Table 3. Dipeptide compositions with higher frequency were considered as more important in identifying mitochondrial proteins. Figure 1 shows the number of dipeptide compositions with a frequency above 70% in each 100 operations.

$$\text{frequency(i)} = \frac{\text{the total number of dip(i) selected by GA-PLS}}{\text{the times of operation using GA-PLS}} \quad (2)$$

where i is the *i*th dipeptide out of 400 dipeptides.

It was observed that the number of common dipeptide compositions with frequency above 90% and 80% stabilized around 84 and 112, respectively; and the number of common dipeptide compositions with frequency 100% trended to 26 gradually. The number with other frequencies also shows a similar phenomenon. So we utilized these common dipeptide compositions as input to construct the SVM models, respectively. The details are listed in Table 4.

It was obvious that SVM model using 84 dipeptide compositions (84-D) with frequency of 90% showed the best performance with MCC of 0.6913. In order to validate the importance of the 84 dipeptides, a comparison was completed on the base of other 316 dipeptides. Although the number of features (dipeptide compositions) was far more than 84, the result showed worse performance. The 26 dipeptides were considered as the most important dipeptides. However, due to the fact that none of them can represent significant information, more than 84 dipeptide compositions were needed for SVM-model construction. We believe that in the further study, more useful biological information on the base of these dipeptides will be found.

In order to explore the performance of the method in detail, we further analyzed the result on the base of species composed of the dataset, it observed that after feature selection by GA, the accuracies of most species were influenced a little, and the accuracy of fungus was even greatly improved (Table 5). It proved that the 84 dipeptide compositions were sufficient and efficient in identifying mitochondrial proteins.

3.2 Comparison with other prediction methods

In comparison, we made our method compare with other existing prediction methods using the same dataset with a jackknife test; all prediction results are listed in Table 6. The results depicted that the selected 84 dipeptide compositions (84-D) can identify mitochondrial proteins from other proteins with reasonable accuracy of 0.85 and MCC of 0.69, respectively. MITOPRED showed the

Table 3. Frequency (%) of each dipeptide composition occurrence in GA-PLS model

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	100	11	99	12	24	84	11	91	52	48	95	42	34	6.8	100	100	95	41	5.4	63
C	1.8	100	73	68	71	92	5.0	14	95	42	8.0	97	7.4	30	99	73	40	2.4	85	59
D	91	90	100	75	0.0	92	5.2	5.0	2.0	100	99	90	100	99	100	100	66	18	46	0.0
E	4.2	98	46	56	7.0	71	18	100	20	1.6	2.6	3.0	0.2	1.4	4.2	6.6	99	21	5.4	16
F	98	2.6	51	5.6	69	64	88	82	59	91	2.2	92	56	100	47	1.4	3.8	85	25	99
G	99	24	100	64	2.0	54	15	4.6	50	98	0.8	93	94	69	38	99	9.4	70	37	4.0
H	1.4	4.0	97	4.0	42	1.4	3.8	61	8.4	86	1.2	86	88	6.8	3.8	3.2	58	67	63	33
I	100	85	33	91	84	11	6.0	100	24	100	12	11	8.0	67	3.6	73	100	100	68	52
K	69	95	19	64	6.0	49	31	37	74	87	1.6	56	94	73	97	79	4.0	27	28	51
L	37	8.4	74	4.0	12	100	3.2	100	95	97	14	20	1.2	81	100	87	45	100	82	33
M	90	9.6	57	100	54	9.4	72	95	98	100	4.8	0.0	46	30	46	12	22	38	89	59
N	2.2	95	83	27	52	1.4	0.4	40	23	85	71	44	45	96	80	65	45	75	20	93
P	98	81	32	11	31	44	0.2	74	84	97	0.6	8.0	30	1.2	10	5.8	70	67	6.2	0.0
Q	1.8	1.0	42	31	6.0	0.0	91	12	99	0.0	28	0.0	56	50	12	0.2	20	64	6.8	2.0
R	99	0.6	2.4	0.4	1.0	42	1.2	32	91	21	1.0	6.0	89	0.0	1.6	54	47	0.0	0.0	32
S	95	6.4	100	5.6	2.0	38	2.4	6.0	72	44	35	86	0.8	0.6	5.0	17	93	93	26	11
T	21	99	0.2	17	32	19	99	84	29	95	0.6	49	34	98	89	96	27	65	89	0.0
V	99	61	95	7.8	1.0	17	0.8	95	2.8	97	64	26	100	46	24	0.0	1.2	71	0.0	34
W	49	0.0	0.2	65	70	27	3.2	4.6	100	1.8	67	40	0.2	86	34	68	2.4	6.8	8.4	83
Y	42	67	7.4	30	12	1.6	3.8	95	52	73	5.0	1.0	3.0	19	79	33	99	0.8	0.0	83

The shadow represented the most informative dipeptide compositions in identifying the mitochondrial proteins

best performance with accuracy of 0.9568; MitoProt performed marginally better than dipeptide composition-based method. But our work remained high accuracy without any biological relevant information. In practice, not all the biological information can be easily attained. Once such information absent, our method will be influenced little. As the same as MITOPRED, MitoProt also has some limitations: it can only predict the sequences started by a methionine and the mature proteins which cleaved the precursor or the long sequences can't be predicted, either. Moreover, the datasets of MITOPRED and MitoProt may include part of our sequences, which resulted in a better performance than this work. Considering this, the differences between MITOPRED, MitoProt and this work will decrease. The discrete wavelet transform method (Jiang et al., 2006), based on the sequence–scale similarity measurement, does not rely on subcellular locations information and can directly predict protein sequences with different length. Although the performance of specificity is relative higher, the accuracy is poor. It is usually due to specific properties of mitochondrial protein that make it difficult to discriminate from other proteins by just one method, or simply because the number of proteins present in the mitochondrion is

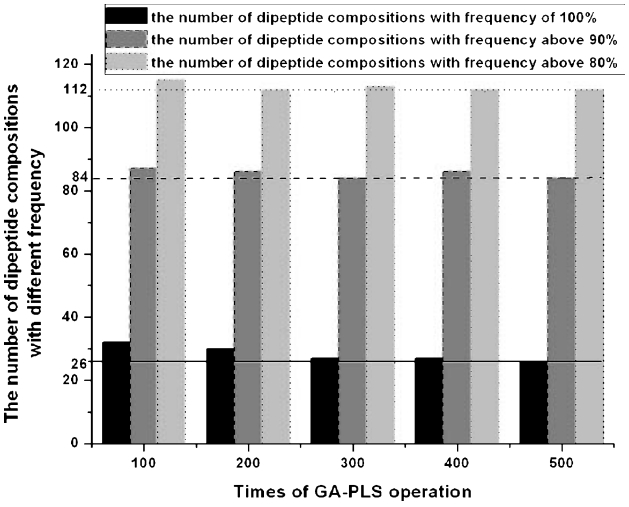

Fig. 1. The number of common dipeptide composition elements at different frequency. The abscissa represents the times of operation using GA, and the ordinate represents the number of common dipeptide composition elements at different frequency corresponding to the times of GA run

Table 4. Performance of the method in identifying mitochondrial proteins using jackknife test at different frequency thresholds

Frequency threshold	No. of dipeptides	Sen	Sp	Acc	MCC
100	26	0.7054	0.7680	0.7415	0.4721
95	66	0.7756	0.7944	0.7864	0.5664
90	84	0.7916	0.8928	0.8500	0.6913
85	100	0.7976	0.8223	0.8119	0.6170
80	112	0.7916	0.8076	0.8008	0.5956
70	133	0.8076	0.8194	0.8144	0.6232
60	157	0.7876	0.8179	0.8051	0.6029
0	400	0.8076	0.8458	0.8297	0.6521
<90	316	0.7010	0.8370	0.7800	0.5452

Table 5. Performance of our method in identifying mitochondrial proteins of each species using jackknife test

Species	No. of sequences	400-D dipeptides		84-D dipeptides	
		Correct hit	Accuracy (%)	Correct hit	Accuracy (%)
Human	88	81	92.05	77	87.50
Animal	207	177	85.51	169	81.64
Plant	52	43	82.70	45	86.54
Fungus	152	83	54.61	109	71.71
Total	499	384	76.95	400	80.16

400-D 400 Dipeptide composition technology; *84-D* 84 dipeptide composition technology

unmanageable (Cameron, et al., 2005). With the increase of the exact experimental mitochondrial proteins, the performance should also be improved significantly.

Table 6. Comparison of performances of commonly-used mitochondrial protein prediction programs and our method using the same datasets

Method	Sen	Sp	Acc	MCC
SVM(400-D)	0.8076	0.8458	0.8297	0.65
SVM(84-D)	0.7916	0.8928	0.8500	0.69
DWT method	0.5030	0.9574	0.7653	0.54
MITOPRED[a]	0.9279	0.9780	0.9568	0.89
MitoProt[b]	0.8617	0.8414	0.8508	0.70

400-D 400 Dipeptide composition technology; *84-D* 84 dipeptide composition technology

[a] Prediction performances of MITOPRED were calculated at a confidence cutoff of 0.85

[b] Prediction performances of MitoProt were calculated at a threshold of 0.70

4. Conclusions

In this paper, GA-PLS and SVM methods based on dipeptide composition of proteins were developed to predict the mitochondrial proteins. Traditionally, all the 400 dipeptides are used as features for proteins prediction. This report proclaims that not all the dipeptides are informative in identifying proteins. The GA-PLS selects features themselves other than their assemblage, which makes it possible to progress further study of protein structure and biological function on the base of these selected dipeptides.

This work describes a statistical prediction method to distinguish mitochondrial proteins only using raw sequence data. Therefore, it is helpful in annotating the mitochondrial proteins in the absence of experiment data. It is anticipated that it can play a supplementary role to biochemical experiments and help to provide insights in selecting the peptides for drug discovery in further study.

Acknowledgement

This work was supported by the foundation of the State Key Laboratory of Chemo/Biosensing and Chemometrics.

References

Andrade MA, O'Donoghue SI, Rost B (1998) Adaption of protein surfaces to subcellular location. J Mol Biol 276: 517–525

Andreoli C, Prokisch H, Hortnagel K, Mueller JC, Munsterkotter M, Scharfe C, Meitinger T (2004) MitoP2, an integrated database on mitochondrial proteins in yeast and man. Nucleic Acids Res 32: 459–462

Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340: 783–795

Boeckmann B, Bairoch A, Apweiler R, Blatter M, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schn M (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 31: 365–370

Cai YD, Chou KC (2005) Predicting enzyme subclass by functional domain composition and pseudo amino acid composition. J Proteome Res 4: 967–971

Cai YD, Zhou GP, Chou KC (2003) Support vector machines for predicting membrane protein types by using functional domain composition. Biophys J 84: 3257–3263

Cameron JM, Hurd T, Robinson BH (2005) Computational identification of human mitochondrial proteins based on homology to yeast mitochondrially targeted proteins. Bioinformatics 21: 1825–1830

Chou KC (1995) A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins Struct Funct Genet 21: 319–344

Chou KC (1999) Using pair-coupled amino acid composition to predict protein secondary structure content. J Protein Chem 18: 473–480

Chou KC (2000a) Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochem Biophys Res Commun 278: 477–483

Chou KC (2000b) Review: prediction of protein structural classes and subcellular locations. Curr Protein Pept Sci 1: 171–208

Chou KC (2001) Prediction of protein cellular attributes using pseudo amino acid composition. Proteins Struct Funct Genet 43: 246–255 [Erratum ibid. (2001) 44: 60]

Chou KC (2005a) Review: progress in protein structural class prediction and its impact to bioinformatics and proteomics. Curr Protein Pept Sci 6: 423–436

Chou KC (2005b) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21: 10–19

Chou KC (2005c) Prediction of G-protein-coupled receptor classes. J Proteome Res 4: 1413–1418

Chou KC, Cai YD (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. J Biol Chem 277: 45765–45769

Chou KC, Cai YD (2003) Prediction and classification of protein subcellular location: sequence-order effect and pseudo amino acid composition. J Cell Biochem 90: 1250–1260 (Addendum, ibid. 2004, 91, 1085)

Chou KC, Cai YD (2004a) Predicting enzyme family class in a hybridization space. Protein Sci 13: 2857–2863

Chou KC, Cai YD (2004b) Predicting protein structural class by functional domain composition. Biochem Biophys Res Commun 321: 1007–1009 (Corrigendum: ibid., 2005, Vol. 329, 1362)

Chou KC, Cai YD (2006) Predicting protein–protein interactions from sequences in a hybridization space. J Proteome Res 5: 316–322

Chou KC, Elrod DW (1998) Using discriminant function for prediction of subcellular location of prokaryotic proteins. Biochem Biophys Res Commun 252: 63–68

Chou KC, Elrod DW (1999) Protein subcellular location prediction. Protein Eng 12: 107–118

Chou KC, Elrod DW (2002) Bioinformatical analysis of G-protein-coupled receptors. J Proteome Res 1: 429–433

Chou KC, Elrod DW (2003) Prediction of enzyme family classes. J Proteome Res 2: 183–190

Chou KC, Liu W, Maggiora GM, Zhang CT (1998) Prediction and classification of domain structural classes. Proteins Struct Funct Genet 31: 97–103

Chou KC, Maggiora GM (1998) Domain structural class prediction. Protein Eng 11: 523–538

Chou KC, Shen HB (2006a) Predicting protein subcellular location by fusing multiple classifiers. J Cell Biochem 99: 517–527

Chou KC, Shen HB (2006b) Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. Biochem Biophys Res Commun 347: 150–157

Chou KC, Shen HB (2006c) Addendum to ''Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization''. Biochem Biophys Res Commun 348: 1479

Chou KC, Shen HB (2006d) Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. J Proteome Res 5: 1888–1897

Chou KC, Zhang CT (1995) Review: prediction of protein structural classes. Crit Rev Biochem Mol Biol 30: 275–349

Claros MG, Vincens P (1996) Computational method to predict mitochondrial proteins and their targeting sequences. Eur J Biochem 241: 779–786

Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300: 1005–1016

Gao Y, Shao SH, Xiao X, Ding YS, Huang YS, Huang ZD, Chou KC (2005) Using pseudo amino acid composition to predict protein subcellular location: approached with Lyapunov index, Bessel function, and Chebyshev filter. Amino Acids 28: 373–376

Guda C, Fahy E, Subramaniam S (2004) MITOPRED: a genome-scale method for prediction of nucleus-encoded mitochondrial proteins. Bioinformatics 20: 1785–1794

Guo YZ, Li M, Lu M, Wen Z, Wang K, Li G, Wu J (2006) Classifying GPCRs and NRs based on protein power spectrum from fast Fourier transform. Amino Acids 30: 397–402

Hasegawa K, Kimura T, Funatsu K (1999) GA strategy for variable selection in QSAR studies: enhancement of comparative molecular binding energy analysis by GA-based PLS method. Quant Struct Acta Relat 18: 262–272

Höglund A, Dönnes P, Blum T, Adolph HW, Kohlbacher O (2006) MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. Bioinformatics 22: 1158–1165

Horton P, Park KJ, Obayashi T, Nakai K (2006) Protein subcellular localization prediction with WoLF PSORT. Proceedings of the 4th Annual Asia Pacific Bioinformatics Conference APBC06, Taipei, Taiwan, pp 39–48

Hua SJ, Sun ZR (2001) Support vector machine approach for protein subcellular localization prediction. Bioinformatics 17: 721–728

Huang Y, Li Y (2004) Prediction of protein subcellular locations using fuzzy k-NN method. Bioinformatics 20: 21–28

Jiang L, Li ML, Wen ZN, Wang KL, Diao YB, Guo YZ, Liu LX (2006) Prediction of mitochondrial proteins using discrete wavelet transform. Protein J 25: 241–249

Kumar M, Verma R, Raghava GPS (2006) Prediction of mitochondrial proteins using support vector machine and hidden Markov model. J Biol Chem 281: 5357–5363

Leardi R (2000) Application of genetic algorithm-PLS for feature election in spectral data sets. J Chemometrics 14: 643–655

Leardi R, Lupiáñez A (1998) Genetic algorithms applied to feature selection in PLS regression: how and when to use them. Chemom Intell Lab Syst 41: 195–207

Liu H, Yang J, Wang M, Xue L, Chou KC (2005) Using Fourier spectrum analysis and pseudo amino acid composition for prediction of membrane protein types. Protein J 24: 385–389

Liu LX, Li ML, Tan FY, Lu MC, Wang KL, Guo YZ, Wen ZN, Jiang L (2006) Local sequence information-based support vector machine to classify voltage-gated potassium channels. Acta Biochim Biophys Sin 38: 363–371

Liu W, Chou KC (1999) Protein secondary structural content prediction. Protein Eng 12: 1041–1050

Matthews BW (1975) Comparison of predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405: 442–451

Ros F, Pintore M, Chrétien JR (2002) Molecular descriptor selection combining genetic algorithms and fuzzy logic: application to database mining procedures. Chemom Intell Lab Syst 63: 15–26

Scharfe C, Zaccaria P, Hoertnagel K, Jaksch M, Klopstock T, Lill R, Prokisch H, Gerbitz KD, Mewes HW, Meitinger T (2000) MITOP, the mitochondrial proteome database: 2000 update. Nucleic Acids Res 28: 155–158

Shen HB, Chou KC (2005a) Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo amino acid composition to predict membrane protein types. Biochem Biophys Res Commun 334: 288–292

Shen HB, Chou KC (2005b) Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition. Biochem Biophys Res Commun 337: 752–756

Shen HB, Chou KC (2006) Ensemble classifier for protein fold pattern recognition. Bioinformatics 22: 1717–1722

Shen HB, Yang J, Chou KC (2006) Fuzzy KNN for predicting membrane protein types from pseudo amino acid composition. J Theor Biol 240: 9–13

Shen HB, Yang J, Liu XJ, Chou KC (2005) Using supervised fuzzy clustering to predict protein structural classes. Biochem Biophys Res Commun 334: 577–581

Sun XD, Huang RB (2006) Prediction of protein structural classes using support vector machines. Amino Acids 30: 469–475

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680

Vapnik VN (1998) Statistical learning theory. J. Wiley, New York

Wen Z, Li M, Li Y, GuoY, Wang K (2006) Delaunay triangulation with partial least squares projection to latent structures: a model for G-protein coupled receptors classification and fast structure recognition. Amino Acids 32: 277–283

Xiao X, Shao S, Ding Y, Huang Z, Chen X, Chou KC (2005a) An application of gene comparative image for predicting the effect on replication ratio by HBV virus gene missense mutation. J Theor Biol 235: 555–565

Xiao X, Shao S, Ding Y, Huang Z, Huang Y, Chou KC (2005b) Using complexity measure factor to predict protein subcellular location. Amino Acids 28: 57–61

Xiao X, Shao SH, Chou KC (2006a) A probability cellular automaton model for hepatitis B viral infections. Biochem Biophys Res Commun 342: 605–610

Xiao X, Shao S, Ding Y, Huang Z, Chou KC (2006b) Using cellular automata images and pseudo amino acid composition to predict protein sub-cellular location. Amino Acids 30: 49–54

Xiao X, Shao SH, Huang ZD, Chou KC (2006c) Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. J Comput Chem 27: 478–482

Yuan Z (1999) Prediction of protein subcellular locations using Markov chain models. FEBS Lett 451: 23–26

Zhang SW, Pan Q, Zhang HC, Shao ZC, Shi JY (2006) Prediction protein homo-oligomer types by pseudo amino acid composition: approached with an improved feature extraction and naive Bayes feature fusion. Amino Acids 30: 461–468

Zhou GP (1998) An intriguing controversy over protein structural class prediction. J Protein Chem 17: 729–738

Zhou GP, Assa-Munt N (2001) Some insights into protein structural class prediction. Proteins 44: 57–59

Zhou GP, Doctor K (2003) Subcellular location prediction of apoptosis proteins. Proteins 50: 44–48

Authors' address: Prof. Menglong Li, College of Chemistry, Sichuan University, Chengdu 610064, P.R. China,
Fax: +86-28-85412356, E-mail: liml@scu.edu.cn