Amino Acids (2008) 34: 119–125
DOI 10.1007/s00726-007-0545-9
Printed in The Netherlands

Amino Acids

Using pseudo amino acid composition to predict protein subnuclear location with improved hybrid approach

F.-M. Li[1,2] and **Q.-Z. Li**[1]

[1] Laboratory of Theoretical Biophysics, Department of Physics, College of Sciences and Technology, Inner Mongolia University, Hohhot, China
[2] College of Sciences, Inner Mongolia Agricultural University, Hohhot, China

Received February 21, 2007
Accepted March 7, 2007
Published online May 21, 2007; © Springer-Verlag 2007

Summary. The subnuclear localization of nuclear protein is very important for in-depth understanding of the construction and function of the nucleus. Based on the amino acid and pseudo amino acid composition (PseAA) as originally introduced by K. C. Chou can incorporate much more information of a protein sequence than the classical amino acid composition so as to significantly enhance the power of using a discrete model to predict various attributes of a protein, an algorithm of increment of diversity combined with the improved quadratic discriminant analysis is proposed to predict the protein subnuclear location. The overall predictive success rates and correlation coefficient are 75.4% and 0.629 for 504 single localization proteins in jackknife test, and 80.4% for an independent set of 92 multi-localization proteins, respectively. For 406 single localization nuclear proteins with ≤25% sequence identity, the results of jackknife test show that the overall accuracy of prediction is 77.1%.

Keywords: Subnuclear localization – Increment of diversity – Quadratic discriminant – Covariant discriminant function – Chou's pseudo amino acid composition

Introduction

The cell nucleus is a highly complex organelle that organizes the comprehensive assembly of genes and their corresponding regulatory factors. It reflects the intricate regulation of various biological activities. The nucleus contains many proteins, and its biological functions are closely relevant to the proteins therein. Although protein complexes disperse throughout the entire organelle, it is known that many nuclear proteins participating in life processes tend to concentrate on subnuclear compartments (Heidi et al., 2001; Joanna and Wendy, 1998). The concentration within specific subnuclear compartments is important for the construction and function of the nucleus. The importance of this organization is revealed by the mis-localization of nuclear proteins in human genetic disease (Marsh et al., 1998; Wilson et al., 2001), in cancers (Koken et al., 1997; Phair and Misteli, 2000) and in virally infected cells (Bell et al., 2000). The knowledge of protein subcellular or subnuclear localization can provide valuable clues about its molecular function, as well as the biological pathway in which it participates (Chou, 2000b, 2002; Chou et al., 2006; Cocco et al., 2004; Itoh et al., 2005). Accordingly, the knowledge of protein subnuclear localization is very important for in-depth understanding of the biochemical process of the nucleus.

Advances in experimental technology have enabled the large-scale identification of nuclear proteins. However, at the same time, the sequencing of both the human and mouse genomes has generated an enormous inventory of primary sequences with unknown functions. The avalanche of these protein sequences has called for development of automated methods for fast identifying the localization of uncharacterized proteins in cell. Therefore, accurately predicting protein subnuclear localization is crucial for understanding genome regulation and functions. Many of the existing methods were focused on the prediction of protein subcellular localizations from primary protein sequences (Bulashevska and Eils, 2006; Cai and Chou, 2003; Cai et al., 2002; Chou and Cai, 2002; Chou and Elrod, 1999; Chou and Shen, 2006a, b, c, d, 2007; Du and Li, 2006; Feng, 2001, 2002; Gao et al., 2005a, b; Guo et al., 2006a; Pan et al., 2003; Shen and Chou, 2007; Shen et al., 2007; Xiao et al., 2006a; Zhou and Doctor, 2003). However, the prediction of protein localization from primary protein sequences at subnuclear level is challenging compared with that at the subcellular level (Lei and

Dai, 2005; Shen and Chou, 2005). In order to extend the prediction of protein subcellular location into a deeper level, i.e., the subnuclear level, a novel approach is proposed for predicting the subnuclear location of 370 nuclear proteins (Shen and Chou, 2005).

Recently, Lei and Dai assessed protein similarity with gene ontology (GO) and then used new kernel functions in a support vector machine (SVM) learning model for classifying the 504 single-localization and the 92 multi-localization nuclear proteins into their respective subnuclear location based on their primary sequence. The overall accuracy is elevated from 50.0 to 66.5% for single-localization proteins in jackknife test; and from 65 to 65.2% for an independent set of multi-localization proteins (Lei and Dai, 2005, 2006). In addition, an evolutionary support vector machine (ESVM) is proposed to predict subnuclear localization (Huang, 2007). The overall accuracy of prediction is 56.37% for the 504 proteins in the same dataset.

In this article, an algorithm of increment of diversity (ID) combined with improved quadratic discriminant analysis (IDQD) is introduced to predict the subnuclear location of nuclear proteins by using of amino acid compositions (AA) and pseudo amino acid compositions (PseAA). Compared with AA, using PseAA can avoid completely lose the sequence-order information, as elaborated in the original paper by Chou (2001). The concept of Chou's pseudo amino acid composition has stimulated a series of studies to use such an approach to improve the prediction quality in various areas (see, e.g., Chen et al., 2006a, b; Chou and Cai, 2003; Du and Li, 2006; Guo et al., 2006b; Lin and Li, 2007a, b; Mondal et al., 2006; Shen and Chou, 2005; Shi et al., 2007; Wang et al., 2005; Xiao et al., 2005b; Zhang et al., 2006). The algorithm of increment of diversity (ID) which was first introduced and employed in biogeography is a kind of information description on state space and a measure of whole uncertainly and total information of a system (Laxton, 1978). In order to compare the distribution of two species, one defines the increment of diversity (ID) by the difference of the total diversity measure of two systems and the diversity measure of the mixed system. It can be proved that the higher the similarity of two sources, the smaller the ID. So, the increment of diversity and diversity coefficient of two sources are essentially a measure of their similarity level. Therefore, the ID algorithm and the IDQD model had, respectively, been applied in the recognition of protein structural class (Li and Lu, 2001; Lin and Li, 2007a), the exon-intron splice site prediction (Zhang and Luo, 2003), the subcellular location of an apoptosis protein (Chen and Li, 2007) and conotoxin superfamily (Lin and Li, 2007b). By generalizing the IDQD model from two-classes predictive problem to multi-classes prediction problem, an improved IDQD is applied in the prediction of subnuclear location. The performance of the new system proposed here was compared with recent predicting method using a set of proteins resided within 6 localizations collected from the nuclear protein database (NPD) (Dellaire et al., 2003; Lei and Dai, 2006). The predictive results of the jackknife test show significant improvement compared with other methods.

Materials and methods

Datasets

In order to have sufficient number of proteins for training and testing, 504 single localization nuclear proteins with resided within 6 localizations were constructed by Lei and Dai (2005, 2006). These nuclear proteins derived from the nuclear protein database (NPD) (Dellaire et al., 2003) can be classified into 6 localizations: PML BODY (38), nuclear lamina (55), nuclear splicing speckles (56), chromatin (61), nucleoplasm (75) and nucleolus (219). The sequence identity is analyzed by a culling program (Wang and Dunbrack, 2003). Their sequence identity is less than 65%. And the 92 multi-localization proteins are selected as an independent testing set from same dataset (Lei and Dai, 2005, 2006).

In order to estimate the effectiveness of the new prediction method and the effect of the sequence identity on predicting results, the 406 proteins with sequence identity $\leq 25\%$ are chosen by a culling program (Wang and Dunbrack, 2003) from 504 single-localization proteins. These are 35 PML BODY, 48 nuclear lamina, 44 nuclear splicing speckles, 44 chromatin, 62 nucleoplasm and 173 nucleolus.

Increment of diversity

In a state space of d dimension, the standard diversity measure for diversity source X: $\{n_1, n_2, \ldots, n_i, \ldots, n_d\}$ is defined as (Li and Lu, 2001):

$$D(X) = D(n_1, n_2, \ldots, n_d) = N \log N - \sum_{i=1}^{d} n_i \log n_i \tag{1}$$

here $N = \Sigma_{i=1}^{d} n_i$, n_i indicates the absolute frequency of the i-th state. If n_i equals zero, then $n_i \log n_i = 0$.

In general, for two sources of diversity in the same space of d dimension, X: $\{n_1, n_2, \ldots, n_i, \cdots, n_d\}$ and S: $\{m_1, m_2, \ldots, m_i, \ldots, m_d\}$, the increment of diversity is defined by

$$ID(S, X) = D(S + X) - D(S) - D(X) \tag{2}$$

where $D(S + X)$ is the measure of diversity of the mixed source $X + S$: $\{n_1 + m_1, n_2 + m_2, \ldots, n_i + m_i, \ldots, n_d + m_d\}$.

It is easily proved that the increment of diversity (Eq. (2)) is nonnegative, and can be written as

$$ID(S, X) = D(M, N) - \sum_{i=1}^{d} D(m_i, n_i) \tag{3}$$

where $M = \Sigma_{i=1}^{d} m_i$, $N = \Sigma_{i=1}^{d} n_i$.

$$D(M, N) = (M + N) \log (M + N) - M \log M - N \log N$$

$$D(m_i, n_i) = (m_i + n_i) \log (m_i + n_i) - m_i \log m_i - n_i \log n_i$$

According to the definition of increment of diversity (Eq. (3)), a diversity coefficient $DC(S,X)$ for measuring the similarity level of two sources can be defined as

$$DC(S,X)=\frac{ID(S,X)}{D(M,N)}=1-\sum_{i=1}^{d}\frac{D(m_i,n_i)}{D(M,N)} \tag{4}$$

So, the diversity coefficient of two sources is essentially a measure of their similarity level.

In the same of state space, for an arbitrary protein sequence S to be predicted, six increments of diversity $ID(S,X^{\xi})$ ($\xi=B$, L, S, C, P or N) between the sequence S and the six standard measure of diversities in training sets corresponding, respectively, to PML BODY (B), nuclear lamina (L), nuclear splicing speckles (S), chromatin (C), nucleoplasm (P) and nucleolus (N) protein sequences may be calculated by the following formula:

$$ID(S,X^{\xi})=D(S+X^{\xi})-D(S)-D(X^{\xi}) \quad (\xi=B,L,S,C,P \text{ or } N) \tag{5}$$

here $D(S+X^{\xi})$, $D(S)$, $D(X^{\xi})$ denote standard diversity measure of source: $S+X^{\xi}$, S and X^{ξ} calculated by Eq. (1), respectively.

The six diversity coefficients $DC(S,X^{\xi})$ can be calculated by using Eq. (4). Then the protein S can be predicted to be the subnuclear location for which the corresponding diversity coefficient has the minimum value, and can be formulated as follows:

$$DC(S,X^{\xi})=\mathbf{Min}\{DC(S,X^{B}),DC(S,X^{L}),DC(S,X^{S}),DC(S,X^{C}),DC(S,X^{P}),DC(S,X^{N})\} \tag{6}$$

where ξ can be PML BODY, nuclear lamina, nuclear splicing speckles, chromatin, nucleoplasm or nucleolus and the operator **Min** means taking the minimum value among those in the parentheses as defined by Chou (1995) and Chou and Zhang (1994), then the ξ in Eq. (6) will give the protein location to which the predicted protein sequence S should belong.

Quadratic discriminant and covariant discriminant function

The covariant discriminant function can be used as quadratic discriminant (QD) that was given by Chou (2000a, 2005), Chou and Elrod (1998), Chou and Maggiora (1998), and Liu and Chou (1998):

$$QD_{\xi}=(x-\bar{x}_{\xi})^{T}\Sigma_{\xi}^{-1}(x-\bar{x}_{\xi})+\log|\Sigma_{\xi}| \tag{7}$$

here $\bar{x}_{\xi}$ and Σ_{ξ} are the group mean and covariance matrix, respectively (computed from the ξ training set), the symbol ξ is same to Eq. (5).

The recognition rule should be given by:

$$QD_{\xi}=\mathbf{Min}\{QD_B,QD_L,QD_S,QD_C,QD_P,QD_N\} \tag{8}$$

The meaning of symbol ξ and the operator **Min** is the same as Eq. (6).

Schemes of information parameters

The amino acid hydropathy compositions

The description of a protein sequence can be based on the n-peptide composition coding, denoted by A_n. In the case of $n=1$, the coding reduces to the usual amino acid composition, which can be considered as the first-order approximation to the complete protein sequence. For $n=2$, the coding gives the dipeptide composition. As n increases, the coding provides progressively more detailed sequential information. But at the same time, such a coding scheme becomes not only impractical from a computational viewpoint but also undoable from a learning viewpoint. However, it was demonstrated that in the definition of global protein structure, the patterns of hydrophobic and hydrophilic residues have major significance. To obtain the hydropathy characteristics, the amino acids are divided into groups using their individual hydropathies according to the ranges of the hydropathy scale. The three classifications of the amino acids were derived from their individual hydropathies. In addition, proline, glycine and cysteine are classified into single three groups because of their unique backbone properties. The six groups of 20 amino acids are shown in Table 1. So a protein sequence with 20 amino acids can be represented by a sequence with 6 characters (L (strongly hydrophilic or polar), B (strongly hydrophobic), W (weakly hydrophilic or weakly hydrophobic), P (proline), G (glycine) and C (cysteine)) (Chen and Li, 2007). The n-peptide composition of the six characters along the protein sequence can be selected as the information parameters of a protein, denoted by H_n.

Table 1. Classification of amino acids

Classification	Abbreviation	Amino acids
Strongly hydrophilic or polar	L	R, D, E, N, Q, K, H
Strongly hydrophobic	B	L, I, V, A, M, F
Weakly hydrophilic or weakly hydrophobic	W	S, T, Y, W
Proline	P	P
Glycine	G	G
Cysteine	C	C

The local g-gap dipeptide composition in segmental fragments

Another generalized sequence composition is the g-gap dipeptide compositions, denoted by D_g, in which we compute the composition of the sequence of the form $a(x)_g b$, where a and b denote two specific amino acid types, and $(x)_g$ denotes g intervening amino acids of arbitrary type x. Note that in the special case of $g=0$, D_0 is equivalent to the dipeptide composition A_2. So the local g-gap dipeptide compositions on the N-terminal region with m residues in segmental fragments of protein sequence are chosen as inputting parameters of IDQD.

Results and discussion

Performance assessment and jackknife test

In order to evaluate the predictive capability and reliability of the algorithm, the sensitivity (S_n), specificity (S_p) and correlation coefficient (CC) are defined by

$$S_n=TP/(TP+FN),$$

$$S_p=TP/(TP+FP),$$

$$CC=\frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)\times(TN+FN)\times(TP+FN)\times(TN+FP)}}$$

where TP denotes the numbers of the correctly recognized positives, FN denotes the numbers of the positives recognized as negatives, FP denotes the numbers of the negatives recognized as positives, TN denotes the numbers of correctly recognized negatives.

In statistical prediction, the following three cross-validation tests are often used to examine the power of a predictor: independent dataset test, sub-sampling test and jackknife test. Of these three, the jackknife test is thought the most rigorous and objective one (see Chou and Zhang (1995) for a comprehensive review in this regard), and hence has been used by more and more investigators (Chou, 1995; Feng, 2001, 2002; Guo et al., 2006b; Liu et al., 2005a, b; Luo et al., 2002; Sun and Huang, 2006; Wang et al., 2005; Wen et al., 2007; Xiao et al., 2005a, b, 2006a, b; Zhang et al., 2006; Zhou, 1998; Zhou and Assa-Munt, 2001) in examining the power of various prediction methods. During jackknifing, the subnuclear location of each nuclear protein is identified by the rule parameters derived using all the other nuclear proteins except the one that is being identified.

Diversity coefficient (DC) prediction

In order to predict the subnuclear location of a protein, it is very important to choose classifier and a set of reasonable information parameters from protein sequence. According to the concept of the Chou's PseAA (Chou, 2001, 2005), the 1-peptide composition A_1 of the 20 amino acid compositions and the 2-peptide composition H_2 of the six characters along the protein sequence are first selected as inputting parameters, which are defined in a 56-D space, formulated as:

$$X_{56} = [x_1 \cdots x_i \cdots x_{56}]^T \tag{9}$$

where x_i ($i = 1, 2, \ldots, 20$) and x_i ($i = 21, 22, \ldots, 56$) are, respectively, the absolute occurrence frequencies of the 20 native amino acids and 36 hydropathy dipeptides.

For an arbitrary protein sequence S to be predicted, based on above two kinds of parameters (20 AA and 36 PseAA), the six diversity coefficients DC values between sequence S and six training sets (B, L, S, C, P and N) corresponding to six subnuclear locations (PML BODY, nuclear lamina, nuclear splicing speckles, chromatin, nucleoplasm and nucleolus) can be calculated by using Eq. (4). Then the subnuclear location of the protein S can be predicted by Eq. (6). If two diversity coefficients (DC) have same minimum value, then the test protein will be assigned as "unpredicted". The unpredicted proteins will be passed on the quadratic discriminant (QD) module.

Quadratic discriminant (QD) prediction

Based on the concept of the local g-gap dipeptide compositions D_g, the 1-gap dipeptide compositions D_1 and the 2-gap dipeptide compositions D_2 on the N-terminal region with m residues in protein sequence are selected as the information parameters of a diversity source, which are defined in 400-D space, the predictive results indicate that when $m = 25$, sensitivity (S_n), specificity (S_p) and correlation coefficient (CC) are higher.

$$X_{400} = [x_1 \cdots x_i \cdots x_{400}]^T \quad \text{and} \quad Y_{400} = [y_1 \cdots y_i \cdots y_{400}]^T \tag{10}$$

where x_i ($i = 1, 2, \ldots, 400$) and y_i ($i = 1, 2, \ldots, 400$) are, respectively, the absolute occurrence frequencies of the D_1 and D_2.

For an arbitrary protein sequence S to be predicted, based on above two kinds of parameters(400 PseAA), twelve ($2 \times 6 = 12$) ID values between sequence S and six training sets (B, L, S, C, P or N) are calculated and selected as inputting parameters of QD module.

Results and discussion

For 504 single localization proteins, 308 (238 true predictions and 70 false predictions) out of 504 single localization proteins were predicted by DC module in the jackknife test, and the remaining 196 were passed on to the QD module. The 196 proteins (142 true predictions and 54 false predictions) are predicted by QD module in the jackknife test. For the independent test set of proteins with multi-localizations, 82 (67 true predictions and 15 false predictions) out of 92 proteins were predicted by the DC module, and the remaining 10 proteins were passed on to the QD module. The 10 proteins (7 true predictions and 3 false predictions) are predicted by the QD module.

The overall accuracies (S_n) of prediction are 75.4% for 504 single localization proteins in the jackknife test; and 80.4% for an independent set of multi-localization proteins (Table 2). The overall CC value is 0.629 for single-localization proteins in jackknife test. For the purpose of comparing the predictive capability of IDQD algorithm, the predicted results of other predictive methods are enumerated in Table 2 for the same dataset.

For 406 single localization proteins with sequence identity $\leq$25%, 255 (194 true predictions and 61 false predictions) proteins were predicted by DC module in the jackknife test, and the remaining 151 were passed on to the QD module. The 151 proteins (119 true predictions and 32 false predictions) are predicted by QD module in jackknife test. In addition, for an independent set of 92 multi-localization proteins, the 89 (70 true predictions and 19 false predictions) proteins were predicted by the DC module, and the remaining 3 were passed on to the QD module. The 3 proteins (2 true predictions and 1 false

Table 2. Predictive results of IDQD algorithm compared with other predictive methods

Compartment	Lei-SVM[a]		ESVM[b]		IDQD method		
	S_n (%)	CC	S_n (%)	CC	S_n (%)	S_p (%)	CC
PML BODY	34.2	0.253	18.42	–	44.7	65.4	0.511
Nuclear lamina	63.6	0.578	36.37	–	61.9	69.4	0.615
Nuclear splicing speckles	62.5	0.607	26.79	–	64.3	72.0	0.643
Chromatin	60.7	0.518	21.31	–	60.6	69.8	0.607
Nucleoplasm	56.0	0.504	42.67	–	68.0	70.8	0.642
Nucleolus	79.0	0.656	90.32	–	93.6	80.7	0.758
Overall for single-localization	66.5	0.519	56.37	–	75.4	71.4	0.629
Multi-localization	65.2	–	–	–	80.4	–	–

[a] Source: Lei and Dai (2006)
[b] Source: Huang et al. (2007)

Table 3. Predictive results of the IDQD model by the jackknife test for 498 (406 + 92) proteins

Compartment	S_n (%)	S_p (%)	CC
PML BODY	51.4	81.8	0.624
Nuclear lamina	70.8	69.4	0.661
Nuclear splicing speckles	68.2	85.7	0.740
Chromatin	56.8	83.3	0.659
Nucleoplasm	69.4	63.2	0.599
Nucleolus	94.2	80.7	0.771
Overall for single-localization	77.1	77.4	0.676
Multi-localization	78.3	–	–

prediction) are predicted by the *QD* module. The overall accuracy (S_n) of prediction is 77.1% for 406 single localization proteins in the jackknife test; and 78.3% for an independent set of multi-localization proteins (Table 3).

The results in Table 2 show that the overall jackknife success rate obtained by the IDQD is about 8.9% higher than two other algorithms for 504 single localization proteins; and 15.2% higher than Lei's SVM methods for an independent set of 92 multi-localization proteins, respectively. The result of prediction for 406 single localization proteins with sequence identity $\leq$25% indicates that the IDQD model is helpful for subnuclear location prediction of proteins.

Using the increment of diversity as quadratic discriminant function parameters can reduce dimension of inputting vector, improve calculating efficiency and extract important classify information. It is also evidence that the primary sequence contain important information determined protein advance structure. The local *g*-gap dipeptide compositions can reflect correlation of proximate peptide and successfully enhance the prediction quality for subnuclear location of protein.

Acknowledgments

We are grateful to editor and referees for their careful review and valuable comments on our manuscript. This work was supported by National Natural Science Foundation of China (30560039).

References

Bell P, Lieberman PM, Maul GG (2000) Lytic but not latent replication of Epstein-barr virus is associated with PML and induces sequential release of nuclear domain proteins. J Virol 74: 11800–11810

Bulashevska A, Eils R (2006) Predicting protein subcellular locations using hierarchical ensemble of Bayesian classifiers based on Markov chains. BMC Bioinformatics 7: 298

Cai YD, Chou KC (2003) Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition. Biochem Biophys Res Commun 305: 407–411

Cai YD, Liu XJ, Xu XB, Chou KC (2002) Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect. J Cell Biochem 84: 343–348

Chen C, Tian YX, Zou XY, Cai PX, Mo JY (2006a) Using pseudo-amino acid composition and support vector machine to predict protein structural class. J Theor Biol 243: 444–448

Chen C, Zhou X, Tian Y, Zou X, Cai P (2006b) Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. Anal Biochem 357: 116–121

Chen YL, Li QZ (2007) Prediction of the subcellular location of apoptosis proteins. J Theor Biol 245: 775–783

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 (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 Peptide Sci 1: 171–208

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

Chou KC (2002) A new branch of proteomics: prediction of protein cellular attributes. In: Weinrer PW, Lu Q (eds) Gene cloning & expression technologies, Chapter 4. Eaton Publishing, Westborough, MA, pp 57–70

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

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 (Addendum, ibid. 2004, 91, 1085) 90: 1250–1260

Chou KC, Cai YD, Zhong WZ (2006) Predicting networking couples for metabolic pathways of Arabidopsis. EXCLI J 5: 55–65

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 D (1999) Protein subcellular location prediction. Protein Eng 12: 107–118

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

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

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

Chou KC, Shen HB (2006c) Large-scale predictions of Gram-negative bacterial protein subcellular locations. J Proteome Res 5: 3420–3428

Chou KC, Shen HB (2006d) Predicting protein subcellular location by fusing multiple classifiers. J Proteome Res 99: 517–527

Chou KC, Shen HB (2007) Large-scale plant protein subcellular location prediction. J Proteome Res 100: 665–678

Chou KC, Zhang CT (1994) Predicting protein folding types by distance functions that make allowances for amino acid interactions. J Biol Chem 269: 22014–22020

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

Cocco L, Manzoli L, Barnabei O, Martelli AM (2004) Significance of subnuclear localization of key players of inositol lipid cycle. Adv Enzyme Regul 44: 51–60

Dellaire G, Farrall R, Bickmore WA (2003) The nuclear protein database (NPD): subnuclear localisation and functional annotation of the nuclear proteome. Nucleic Acids Res 31: 328–330

Du P, Li Y (2006) Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physico-chemical features of segmented sequence. BMC Bioinformatics 7: 518

Feng ZP (2001) Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition. Biopolymers 58: 491–499

Feng ZP (2002) An overview on predicting the subcellular location of a protein. In Silico Biol 2: 291–303

Gao QB, Wang ZZ, Yan C, Du YH (2005a) Prediction of protein subcellular location using a combined feature of sequence. FEBS Lett 579: 3444–3448

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

Guo J, Lin Y, Liu X (2006a) GNBSL: a new integrative system to predict the subcellular location for Gram-negative bacteria proteins. Proteomics 6: 5099–5105

Guo YZ, Li M, Lu M, Wen Z, Wang K, Li G, Wu J (2006b) Classifying G protein-coupled receptors and nuclear receptors on the basis of protein power spectrum from fast Fourier transform. Amino Acids 30: 397–402

Heidi GES, Gail KM, Kathryn N, Lisa VF, Rachel F, Graham D, Javier FC, Wendy AB (2001) Large-scale identification of mammalian proteins localized to nuclear sub-compartments. Hum Mol Genet 10: 1995–2011

Huang WL, Tung CW, Huang HL, Hwang SF, Ho SY (2007) ProLoc: prediction of protein subnuclear localization using SVM with automatic selection from physico-chemical composition features. Biosystems (DOI: 10.1016/j.biosystems.2007.01.001)

Itoh K, Brott BK, Bae GU, Ratcliffe MJ, Sokol SY (2005) Nuclear localization is required for Dishevelled function in Wnt/beta-catenin signaling. J Biol 4: 3

Joanna MB, Wendy AB (1998) Putting the genome on the map. Trends Genet 14: 403–409

Koken MHM, Reid A, Quignon F, Chelbi-Alix MK, Davies JM, Kabarowski JHS, Zhu J, Dong S, Chen SJ, Chen Z, Tan CC, Licht J, Waxman S, de Thé H, Zelent A (1997) Leukemia-associated retinoic acid receptor alpha fusion partners, PML and PLZF, heterodimerize and colocalize to nuclear bodies. Proc Natl Acad Sci USA 94: 10255–10260

Laxton RR (1978) The measure of diversity. J Theor Biol 71: 51–67

Lei Z, Dai Y (2005) An SVM-based system for predicting protein subnuclear localizations. BMC Bioinformatics 6: 291

Lei Z, Dai Y (2006) Assessing protein similarity with Gene Ontology and its use in subnuclear localization prediction. BMC Bioinformatics 7: 491

Li QZ, Lu ZQ (2001) The prediction of the structural class of protein: application of the measure of diversity. J Theor Biol 213: 493–502

Lin H, Li QZ (2007a) Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components. J Comput Chem 28(9): 1463–1466

Lin H, Li QZ (2007b) Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant. Biochem Biophys Res Commun 354: 548–551

Liu H, Wang M, Chou KC (2005a) Low-frequency Fourier spectrum for predicting membrane protein types. Biochem Biophys Res Commun 336: 737–739

Liu H, Yang J, Ling JG, Chou KC (2005b) Prediction of protein signal sequences and their cleavage sites by statistical rulers. Biochem Biophys Res Commun 338: 1005–1011

Liu W, Chou KC (1998) Prediction of protein structural classes by modified Mahalanobis discriminant algorithm. J Protein Chem 17: 209–217

Luo RY, Feng ZP, Liu JK (2002) Prediction of protein structural class by amino acid and polypeptide composition. Eur J Biochem 269: 4219–4225

Marsh KL, Dixon J, Dixon MJ (1998) Mutations in the Treacher Collins syndrome gene lead to mislocalization of the nucleolar protein treacle. Hum Mol Genet 7: 1795–1800

Mondal S, Bhavna R, Mohan Babu R, Ramakumar S (2006) Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification. J Theor Biol 243: 252–260

Pan YX, Zhang ZZ, Guo ZM, Feng GY, Huang ZD, He L (2003) Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach. J Protein Chem 22: 395–402

Phair RD, Misteli T (2000) High mobility of proteins in the mammalian cell nucleus. Nature 404: 604–609

Shen HB, Chou KC (2005) 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 (2007) Virus-PLoc: a fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. Biopolymers 85: 233–240

Shen HB, Yang J, Chou KC (2007) Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction. Amino Acids (in press) (DOI: 10.1007/s00726-006-0478-8)

Shi JY, Zhang SW, Pan Q, Cheng YM, Xie J (2007) Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition. Amino Acids (in press) (DOI: 10.1007/s00726-006-0475-y)

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

Wang G, Dunbrack RL Jr (2003) PISCES: a protein sequence culling server. Bioinformatics 19: 1589–1591

Wang M, Yang J, Xu ZJ, Chou KC (2005) SLLE for predicting membrane protein types. J Theor Biol 232: 7–15

Wen Z, Li M, Li Y, Guo Y, Wang K (2007) 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

Wilson KL, Zastrow MS, Lee KK (2001) Lamins and disease: insights into nuclear infrastructure. Cell 104: 647–650

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, Ding YS, Huang ZD, Chou KC (2006a) Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. Amino Acids 30: 49–54

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

Zhang LR, Luo LF (2003) Splice site prediction with quadratic discriminant analysis using diversity measure. Nucleic Acids Res 31: 6214–6220

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 Struct Funct Genet 44: 57–59

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

Authors' address: Qian-Zhong Li, Laboratory of Theoretical Biophysics, Department of Physics, College of Sciences and Technology, Inner Mongolia University, Hohhot 010021, China,
Fax: +86-471-4993124, E-mail: qzli@imu.edu.cn